Using R-INLA in Bayesian Adaptive Designs

Lan Tran
Senior Specialist Statistics and Data Science, PharmaLex
lan.tran@pharmalex.com

with special thanks to my PharmaLex co-authors
Arnaud Monseur, MS, Marco Munda, PhD,
Bruno Boulanger, PhD, and Bradley P. Carlin, PhD

October 2023

Disclaimers

- The information provided during this presentation does not constitute legal advice. PharmaLex, and its parent Cencora, strongly encourage the audience to review available information related to the topics discussed during the presentation and to rely on their own experience and expertise in making decisions related thereto. Further, the contents of this presentation are owned by PharmaLex and reproduction of the slides used in today's presentation is not permitted without consent of PharmaLex.

- Disclaimer: all examples deal with simulated clinical trial settings.

- No conflict of interest by the presenter.

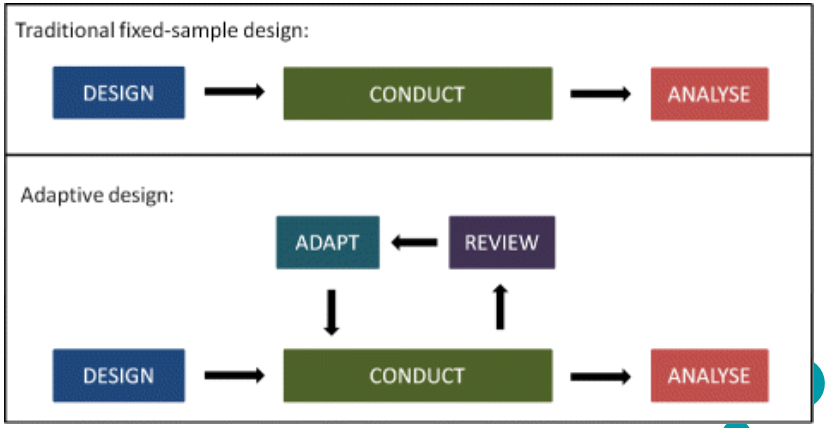
Adaptive designs in clinical trials

Pre-planned changes can include (but are not limited to)[7]

- Refine sample-size
- Drop doses that emerge as less promising
- Stop trial at an early stage for success or lack of efficacy
- Identify patients most likely to benefit from particular doses

Possible **advantages**

- More efficient, informative and ethical
- Save resources, time and money
- Fewer patients required



Source: Pallmann[10]

Key points of this presentation

- **Simulate what-if scenarios more efficiently using INLA**
  Some simulation algorithms take several hours (or days).
  Rerunning them to explore what-if scenarios can be time-consuming.
  INLA is a fast and accurate alternative to MCMC.

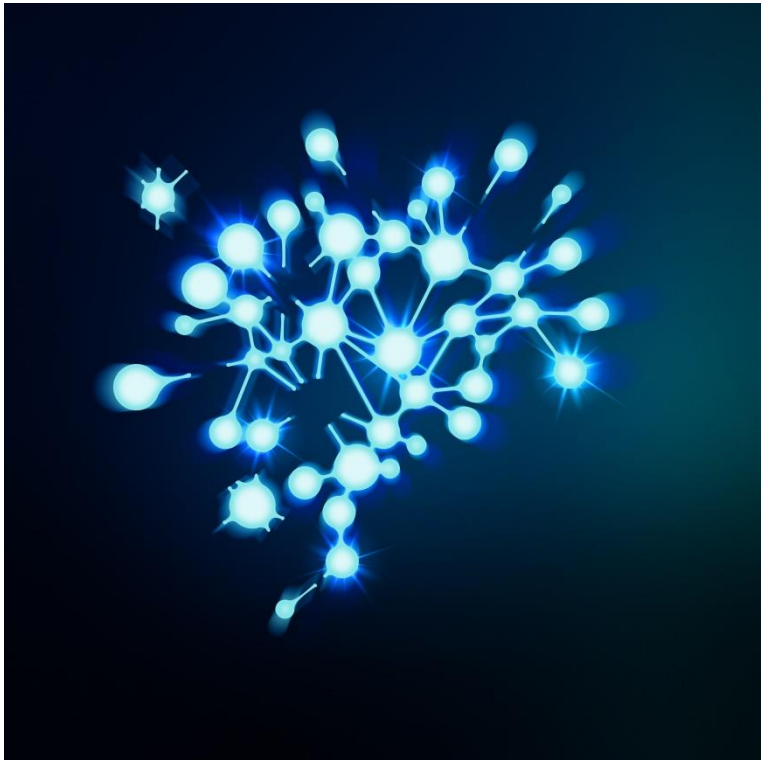
- **Comparison of INLA and STAN within Bayesian adaptive designs**
  Showcase results and computational time within three endpoints.

Keystones of INLA

R-INLA vs R-STAN in clinical trials

- Binary
- Time-to-event
- Continuous longitudinal

Takeaways



Keystones of INLA

Integrated nested Laplace approximation approach

- Suitable for **Bayesian inference**
- Relies on a generalization of the **Laplace approximation**: estimate a Gaussian distribution, using the first 3 terms of a Taylor series
- Applicable to **latent Gaussian models**: prior models that use normally distributed random effects to explicitly model dependence among samples[8]

Why INLA if MCMC exists?

- 1952: first MCMC algorithm designed by Metropolis et al. for use in statistical physics
- 1990 – recently: popularization MCMC for Bayesian analysis due to increased computational power
- MCMC is computationally intensive: Markov chain required, whose convergence must be diagnosed. INLA methods have now been generalized to handle models with Gaussian random effects

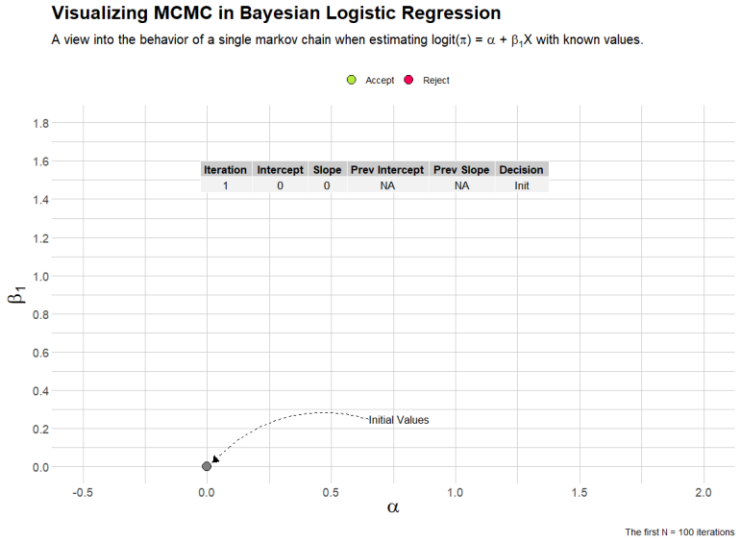
INLA vs MCMC

Advantages of INLA

- INLA is a deterministic algorithm
- Computationally faster than MCMC(2)
  - No burn-in or multiple chains needed
  - No inherent autocorrelations
  - Does not suffer from slow convergence and poor mixing

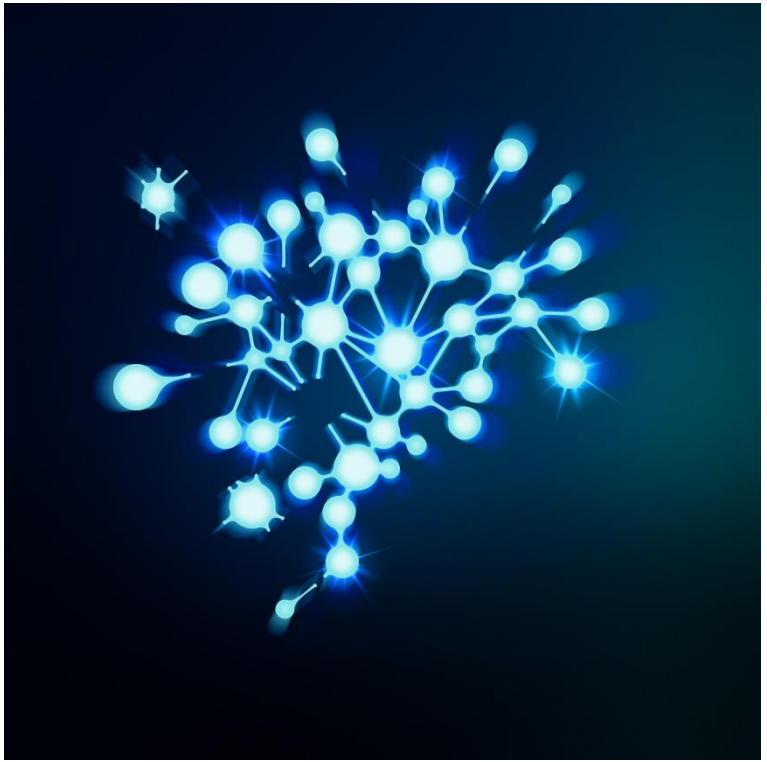
Possible Disadvantages of INLA

- Unlike MCMC, cannot be made arbitrarily accurate simply by running the algorithm longer
- Accuracy in any given problem is hard to judge since its justification relies on asymptotic arguments
  - Accuracy gets better with bigger sample sizes, however, additional data collection not always possible



First 100 iterations of a single chain from a random walk Metropolis Hastings algorithm, generated by Matt Kumar. Animation is slowed down for better visuals.

- Keystones of INLA
- **R-INLA vs R-STAN in clinical trials**
  - **Binary**
  - Time-to-event
  - Continuous longitudinal



Binary Endpoint

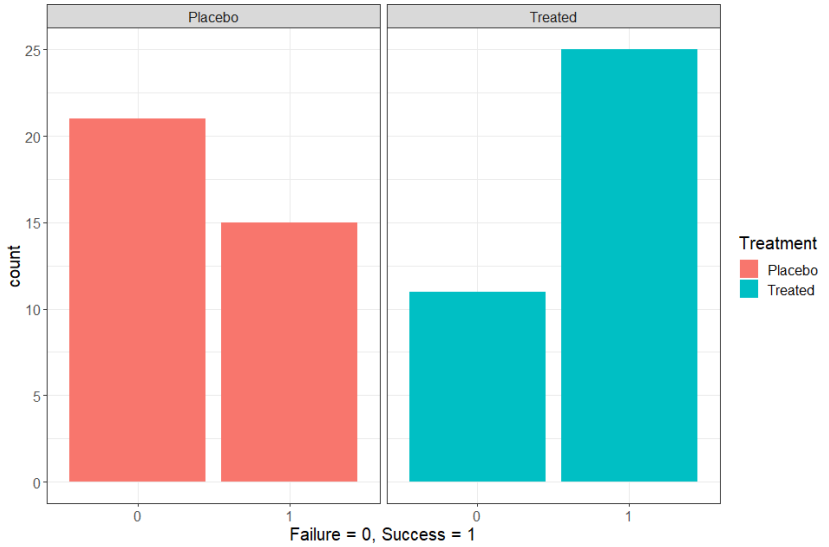
Background

- Logistic regression is used for binary endpoints

  Examples include
  - Successful replacement of a hip
  - Achieving a preset level of change, like increase in hemoglobin by 2g/dL

- Patients received either placebo or a treatment
- Success (1) or failure (0) rate of treatment recorded

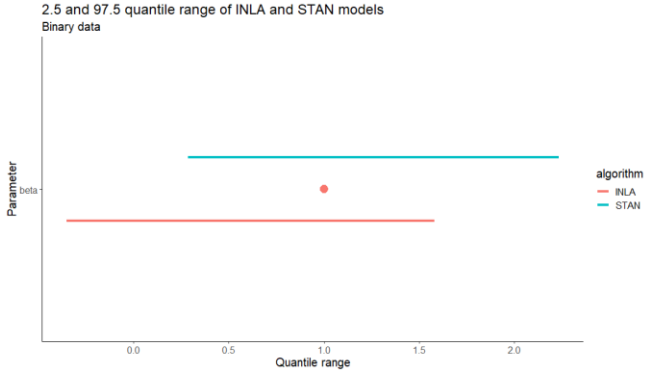
SubjID	Treatment	Result
1	0	0
⋮	⋮	⋮
68	1	1



Binary Endpoint

Comparison STAN and INLA

- Simulate 1000 times a dataset and model the success rates through the logistics function $p(x) = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$ with
  - α the intercept: base rate of placebo
  - β difference in rate of treatment group w.r.t. placebo. Parameter of interest is β.
  - x indicator for $0 =$ placebo and $1 =$ treatment
- Plot is one iteration of quantiles obtained by STAN and INLA models
  - Red dot: true value of β
  - Blue and red lines: 2.5% to 97.5% quantile estimates for β



Binary Endpoint

Results

- Here, β is of interest, as it shows the difference is rates between the two strata
- **Coverage** calculated as $\frac{\#\ 95\%\ \text{quantiles containing}\ \beta}{\#\ \text{simulations}}$
- **Bias** calculated as the mean $(\hat{\beta} - \beta)$, where $\hat{\beta}$ is the true value and β the estimated change

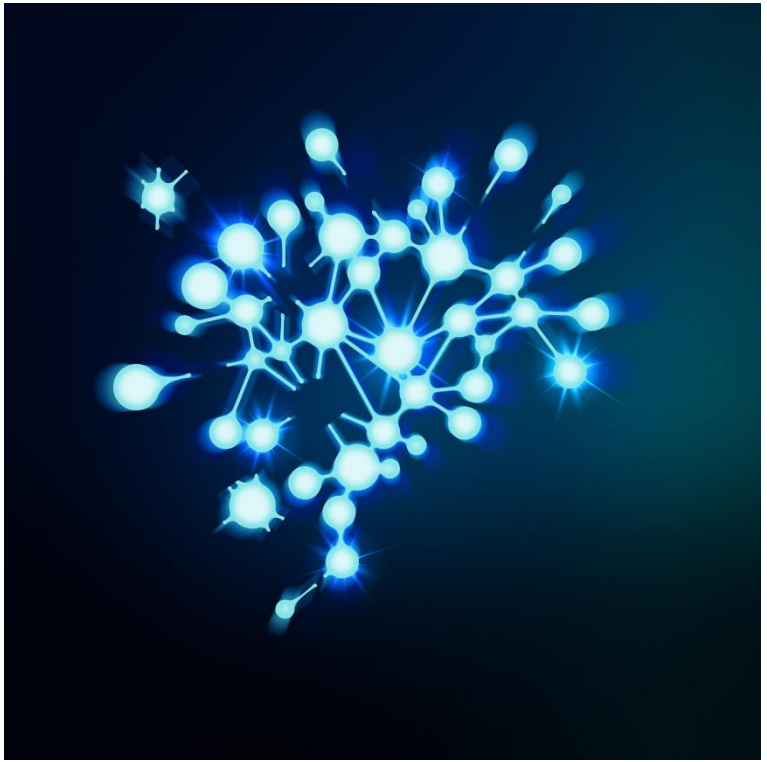
Parameter	Algorithm	Coverage (%)	Mean bias
beta	INLA	95	0.041
beta	STAN	94	0.043

nsim	Algorithm	Sys.time (sec)*
1000	INLA	726 (~12 min)
1000	STAN	978 (~16 min)

Keystones of INLA

R-INLA vs R-STAN in clinical trials

- Binary
- Time-to-event
- Continuous longitudinal



Time-to-event Endpoint

Background

- 1000 patients with a disease were followed over time
- Time = survival or censoring time
  Status = censoring status
  Complication occurred = 0 none, 1 yes

Subject	Time	Status	Complication
1	0.027	1	1
2	0.430	0	0
3	0.712	1	0
⋮	⋮	⋮	⋮
1000	0.270	1	1

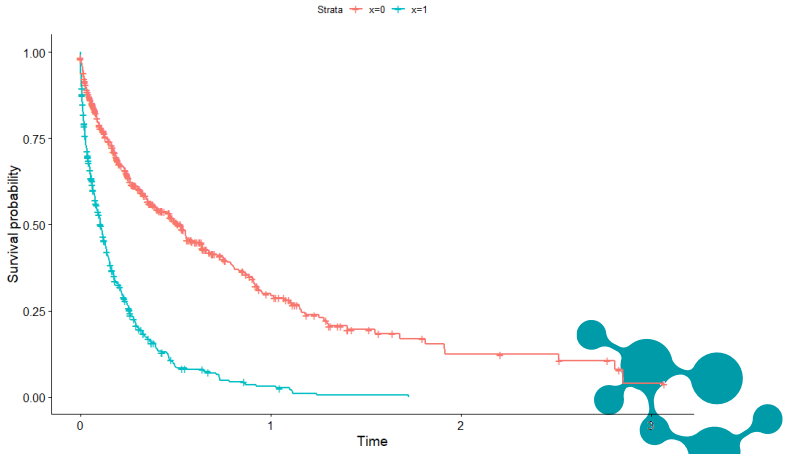
- Note: INLA estimates may struggle with numerical overflow when observed times are large
  Solution: re-scale time before fitting any model, e.g. time=time/max(time)

Time-to-event Endpoint

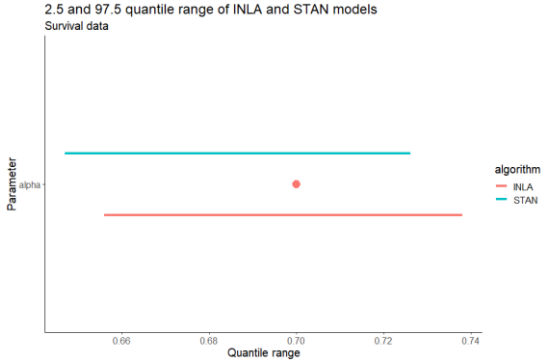
Model

- Data modelled through Weibull probability density function $f(x) = \frac{\alpha}{\sigma}\left(\frac{x}{\sigma}\right)^{\alpha-1} \exp\left(-\left(\frac{x}{\sigma}\right)^{\alpha}\right)$
  where x is time, α the shape parameter and σ the scale parameter
- Shape (or slope) parameter α is of most importance: indicator whether failure rate is increasing, constant or decreasing
  True $\alpha = 0.7$

- Survival probability of patients with a disease
  Stratum = 0: No complication during follow-up phase
  Stratum = 1: Complication occurred



Time-to-event Endpoint



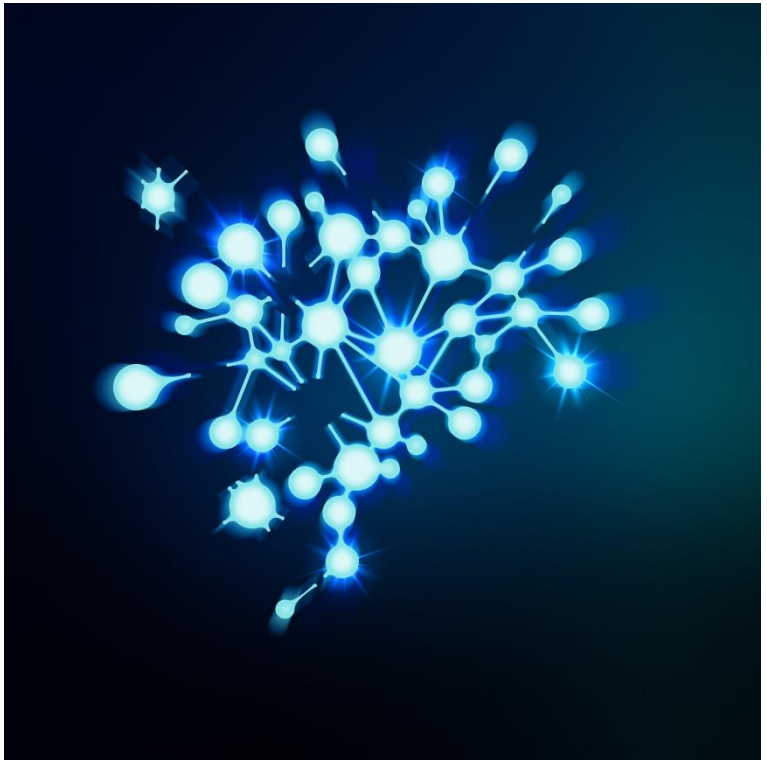
Results

- Simulate 1000 times a dataset and model the estimates
- **Coverage** calculated as $\frac{\#\ 95\%\ \text{quantiles containing}\ \alpha}{\#\ \text{simulations}}$
- **Bias** calculated as the mean$(\hat{\alpha} - \alpha)$, where $\hat{\alpha}$ is the true parameter and α is estimated value

Parameter	Algorithm	Coverage (%)	Mean bias
alpha	INLA	95	0.007
alpha	STAN	98	0.0003

nsim	Algorithm	Sys.time (sec)
1000	INLA	984 (~16 min)
1000	STAN	23034 (~6 h)

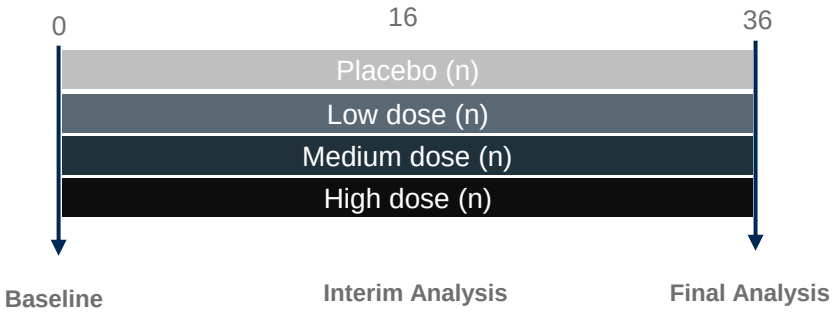
- Keystones of INLA
- **R-INLA vs R-STAN in clinical trials**
  - Binary
  - Time-to-event
  - **Continuous longitudinal**
- Takeaways



Continuous longitudinal Endpoint

Study setup

- Simulate data according to a model
- 4 treatment groups: placebo and low, medium, and high doses
- Measure disease severity (1 – 100%) over several weeks



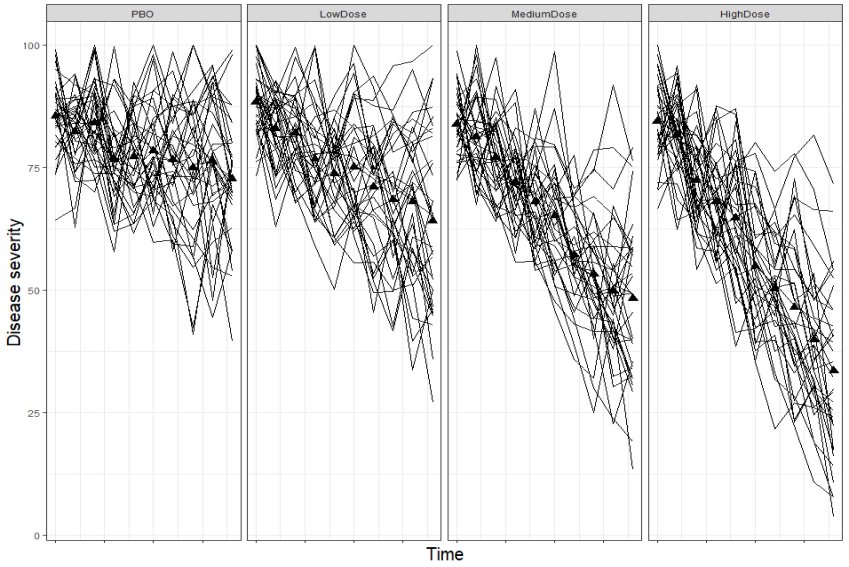
Continuous longitudinal Endpoint

- **Model** the response by

$$Y_{ijk} = \beta_0 + (\beta_1 + \theta_k + b_{1i}) \times t_j + b_{0i} + \epsilon_{ijk},$$

where

- Y_{ijk} is disease severity for subject i, at time j, taking treatment k
- β_0 is the intercept for the placebo group
- β_1 is the slope for the placebo group
- θ_k change in slope w.r.t. placebo group in treatment group $k =$ low, med, high. Define $\theta_{\text{PBO}} \equiv 0$
- t_j time in weeks for measurement point j
- $b_{0i} \sim N(0, \tau_0^2)$ is the random intercept for patient i
- $b_{1i} \sim N(0, \tau_1^2)$ is the random slope for patient i
- $\epsilon_{ijk} \sim N(0, \sigma^2)$ is a random noise term



- Here, θ_k is the parameter of interest

Continuous longitudinal Endpoint

Results

- **Coverage** calculated as $\frac{\#\ 95\%\ \text{quantiles containing}\ \theta}{\#\ \text{simulations}}$
- **Bias** calculated as the mean $\left(\hat{\theta}_k - \theta_k\right)$, where $\hat{\theta}$ is estimated change in slope w.r.t. PBO minus true θ for dose k

Dose	Algorithm	Coverage (%)	Mean bias
Low	INLA	98	0.02
Low	STAN	93	0.02
Medium	INLA	97	0.01
Medium	STAN	96	0.03
High	INLA	93	0.03
High	STAN	98	0.04

Discussion

- 60 simulations is limited but run times of STAN is more than 11 hours!
  Not feasible to explore many scenarios for adaptive designs
- INLA's run time is 442 times faster than STAN, while yielding similar results
- Inclusion of random effects affects model complexity and therefore computing time

nsim	Algorithm	Sys.time (sec)
1000	INLA	1705 (~28 min)
60	STAN	41520 (~11 h)

Continuous longitudinal Endpoint

Discussion

- 100 simulations is limited; not enough to distinguish between INLA's and STAN's coverage.

  If true coverage probability is 95%, then error is $\sqrt{\frac{0.95\times0.05}{100}} \approx 0.022$.

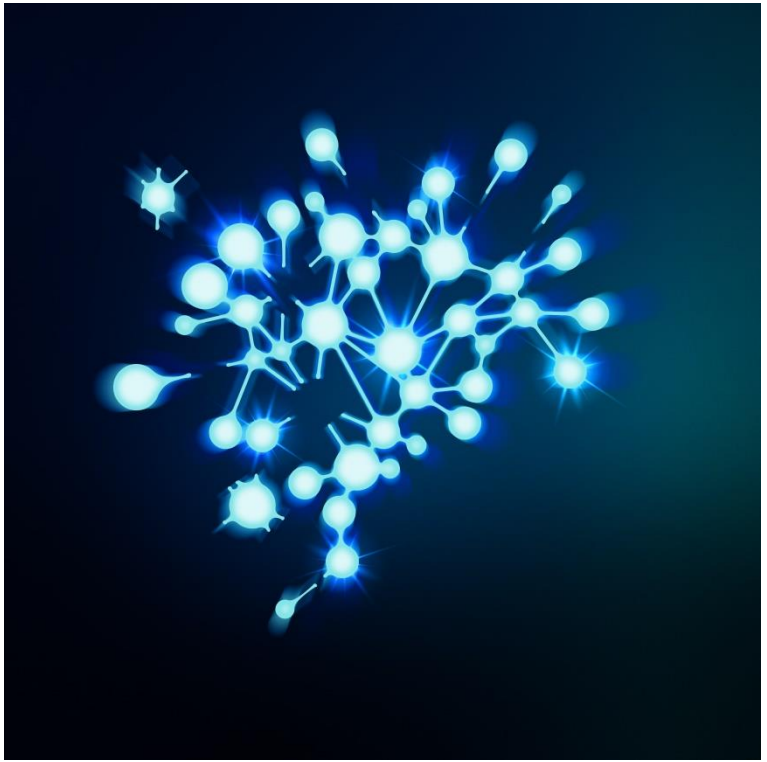
  Thus, its approximate 95% confidence interval is $\mu \pm 2\sigma \rightarrow 0.087$.
- Coverage probabilities must be between 0 and 1

- In practice, 1000 simulations are preferable: error would be $\sqrt{\frac{0.95\times0.05}{1000}} \approx 0.0068$

  95% CI width would be 0.0275, enough to distinguish between coverage rates

- STAN simulations were also employed on a 32-core server.
  Coverage and bias showed marginal differences (.01). Computing time was ~6 hours.

- Keystones of INLA
- R-INLA vs R-STAN in clinical trials
  - Binary
  - Time-to-event
  - Continuous longitudinal
- **Takeaways**



INLA vs MCMC

Summary

- Computation time and parameter estimates compared for INLA and STAN models
- Three common clinical trial settings explored: continuous longitudinal, time-to-event and binary

Performances

- Computation time of INLA is faster than STAN – sometimes marginally, sometimes substantially
- Parameter estimates of both algorithms are similar, as shown by the 95% quantile ranges. However, the limited number of simulations play a role in this
- Coverage is about ~95% for both algorithms
- Bias differs depending on the modeling scenario

Discussion

- Limited number of iterations of the simulation affects the precision
- Factors contributing to the speed of convergence may include, but are not limited to, initial values, model complexity (especially random effects), and data type
- Several studies have made a comparison between INLA and STAN[1, 2, 3, 6], showing similar results

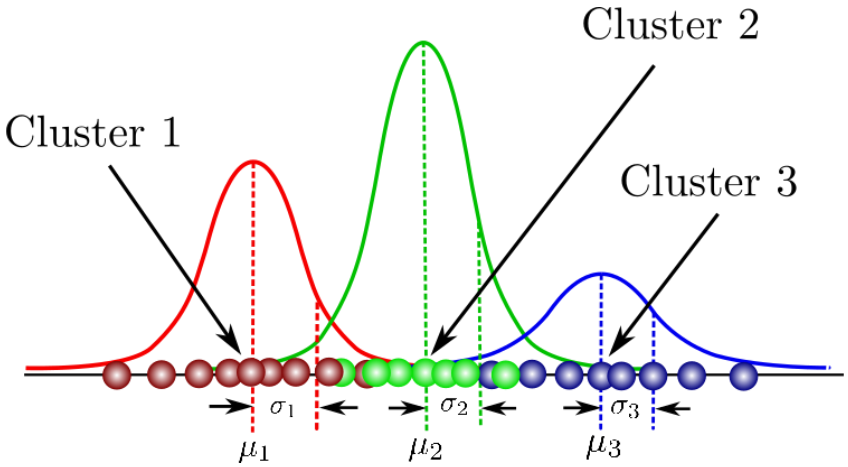
INLA vs MCMC

Limitations INLA

- MCMC can fit *any* hierarchical model

- INLA focusses on models which latent effects arise from a Gaussian Markov random field
- Consequently, INLA cannot fit the following:
  - Mixture models
  - Double hierarchical models
  - Any model where the random effects are not Gaussian (Student's T, Gamma, …)

- Solve the limitation by combining INLA and MCMC[6, 10]



Source: Kumar[9]

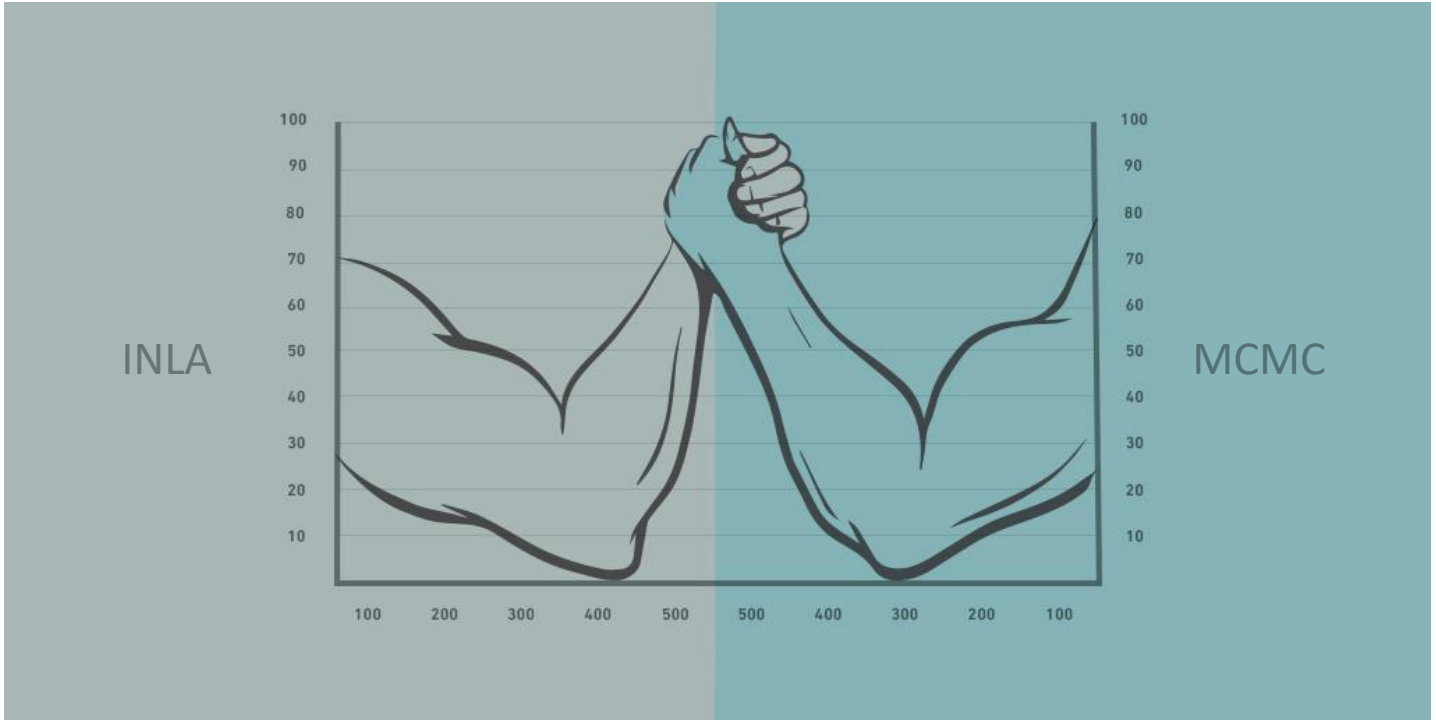


Image Adapted from AB Tasty[4, 5]

Contacts

- **Lan Tran**
  Senior Specialist, Statistics & Data Science
  Development Consulting & Scientific Affairs
  lan.tran@pharmalex.com

Slides and analyses developed with the help of

- **Bruno Boulanger**
  Senior Director, Global Head Data Science
  Development Consulting & Scientific Affairs
  bruno.boulanger@pharmalex.com
- **Arnaud Monseur**
  Associate Director, Statistics & Data Science
  Development Consulting & Scientific Affairs
  arnaud.monseur@pharmalex.com
- **Bradley P. Carlin**
  Senior Advisor, Data Science & Statistics
  Development Consulting & Scientific Affairs
  bradley.p.carlin@pharmalex.com
- **Marco Munda**
  Associate Director, Statistics & Data Science
  Development Consulting & Scientific Affairs
  marco.munda@pharmalex.com

Follow us on social media

/pharmalex-gmbh

@PharmalexGlobal

References

1. Alvares, D., Rustand, D., Krainski, E. T., van Niekerk, J., & Rue, H. (2022). Bayesian survival analysis with INLA. *arXiv preprint arXiv:2212.01900*. Darkwah, J. (2022). Bayesian inference for simple and generalized linear models: Comparing INLA and McMC. In *Advances in Phytochemistry, Textile and Renewable Energy Research for Industrial Growth* (pp. 62-67). CRC Press.
2. *Bayesian Adaptive Clinical Trial Designs: INLA vs. MCMC*. (n.d.). Www.cytel.com. Retrieved October 16, 2023, from https://www.cytel.com/blog/inla-vs-mcmc?utm_medium=social&utm_source=linkedin
3. De Smedt, T., Simons, K., Van Nieuwenhuyse, A., & Molenberghs, G. (2015). Comparing MCMC and INLA for disease mapping with Bayesian hierarchical models. *Archives of Public Health*, *73*, 1.
4. *Frequentist vs Bayesian Methods in A/B Testing - Which is Better?* (2021, January 7). AB Tasty. https://www.abtasty.com/blog/bayesian-ab-testing/
5. Georgiev. G. (2020, February 28). *Frequentist vs Bayesian Inference*. Blog for Web Analytics, Statistics and Data-Driven Internet Marketing | Analytics-Toolkit.com. https://blog.analytics-toolkit.com/2020/frequentist-vs-bayesian-inference/
6. Gómez-Rubio, V., & Rue, H. (2018). Markov chain Monte Carlo with the integrated nested Laplace approximation. *Statistics and Computing*, *28*, 1033-1051.
7. Held, L., Schrödle, B., & Rue, H. (2010). Posterior and cross-validatory predictive checks: a comparison of MCMC and INLA. *Statistical modelling and regression structures: Festschrift in honour of ludwig fahrmeir*, 91-110.
8. *In the Race to Develop a Vaccine For COVID-19, Is a Pull for R&D Essential or Optional?* (n.d.). Center for Global Development. https://www.cgdev.org/blog/race-develop-vaccine-covid-19-pull-rd-essential-or-optional
9. Kumar, A. (2022). Gaussian mixture models: What are they & when to use. *Online, Apr*.
10. Martino, S., & Riebler, A. (2019). Integrated nested Laplace approximations (INLA). *arXiv preprint arXiv:1907.01248*.
11. Pallmann, P., Bedding, A. W., Choodari-Oskooei, B., Dimairo, M., Flight, L., Hampson, L. V., ... & Jaki, T. (2018). Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC medicine*, *16*(1), 1-15.
12. Sigrist, F. (2022). Latent Gaussian model boosting. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(2), 1894-1905.

Appendix

INLA vs MCMC

Computation time INLA vs MCMC for different Bayesian clinical trial designs on a standard PC (Intel Core i7, 16GB RAM)

Comparison of Compute Times: JAGS vs. INLA			
Trial Primary Endpoint Type	**MCMC (sec.)**	**INLA (sec.)**	**MCMC:** 50K iterations, 3 chains **INLA:** Standard INLA Simplified Laplace Approximation
Survival (Oncology)	187	1.1	
Binary (Infectious Disease)	238	0.9	
Repeated Measures (Nephrology)	153	2.7	
Continuous (Rare Disease Biomarker)	36.2	1.0	
Survival (CV) (N=3000+)	>49K (13.7 hours)	27.35	
Repeated Measures (Lipids) (N=7000+)	>250K (~3 days)	396.2	

Source: Cytel[2]

Laplace approximations

- Laplace approximation is an old technique for the approximation of integrals

$$I_n = \int_x \exp\big(n(f)\big)dx$$

1. Approximate the target with a Gaussian
2. Match the mode and the curvature at the mode.
3. By interpreting f(x) as the sum of log-likelihoods and x as the unknown parameter, the Gaussian approximation will be exactas n→∞,if the central limit theorem holds.

- Let x_0 be the point in which $f(x)$ has its maximum. Then

$$I_n \approx \int_x \exp\left(n(f\left((x_0) + \frac{1}{2}(x - x_0)^2 f''(x_0)\right)\right)dx$$

$$= \exp\big(nf(x_0)\big)\sqrt{\frac{2\pi}{-nf''(x_0)}} = \tilde{I}_n$$

Laplace approximations

- **A Laplace approximation is used to estimate a Gaussian distribution, using the first 3 terms of a Taylor series**

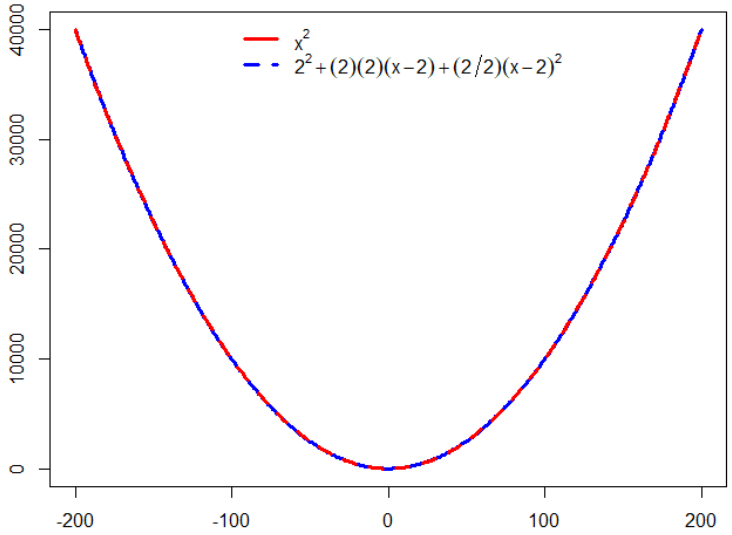
$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \cdots$$

- E.g. For a basic parabola with $y = x^2$, expanding around $a = 2$:

$$f(x) = x^2$$
$$f'(x) = 2x$$
$$f''(x) = 2$$
$$f'''(x) = 0$$

Therefore: $f(x) = x^2 = 2^2 + 2(2)(x-2) + \frac{2}{2}(x-2)^2$

- Thus, a function at point a can be expanded into a sum of terms. Using the first few terms serves as an approximation.



Latent Gaussian models

LGMs are flexible prior models which explicitly model dependence among samples and which allow for efficient learning of predictor functions and for making probabilistic predictions

- In these models, the response is assumed to belong to an exponential family, where the mean μ_i is linked to a structured additive predictor η_i through a link function $g(\cdot)$, so that $g(\mu_i) = \eta_i$

- η_i accounts for effects of various covariates in an additive way:

$$\eta_i = \alpha + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}) + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \epsilon_i$$

- $\{f^{(j)}(\cdot)\}$s are unknown functions of the covariates $\mathbf{u}$, the $\{\beta_l\}$s represent linear effect of covariates $\mathbf{z}$ and ϵ_is are unstructured terms.

Source: Alvares, D., Rustand, D., Krainski, E. T., van Niekerk, J., & Rue, H. (2022). Bayesian survival analysis with INLA. *arXiv preprint arXiv:2212.01900*.

Combining MCMC and INLA

Gómez-Rubio, V., & Rue, H. (2018). Markov chain Monte Carlo with the integrated nested Laplace approximation. *Statistics and Computing*, *28*, 1033-1051.

4 INLA within MCMC

In this Section, we will describe how INLA and MCMC can be combined to fit complex Bayesian hierarchical models. In principle, we will assume that the model cannot be fitted with **R-INLA** unless some of the parameters or hyperparameters in the model are fixed. This set of parameters is denoted by $\boldsymbol{z}_c$ so that the full ensemble of parameters and hyperparameters is $\boldsymbol{z} = (\boldsymbol{z}_c, \boldsymbol{z}_{-c})$. Here $\boldsymbol{z}_{-c}$ is used to denote all the parameters in $\boldsymbol{z}$ that are not in $\boldsymbol{z}_c$. Our assumptions are that the posterior distribution of $\boldsymbol{z}$ can be split as

$$\pi(\boldsymbol{z}|\boldsymbol{y}) \propto \pi(\boldsymbol{y}|\boldsymbol{z}_{-c})\pi(\boldsymbol{z}_{-c}|\boldsymbol{z}_c)\pi(\boldsymbol{z}_c) \tag{12}$$

and that $\pi(\boldsymbol{y}|\boldsymbol{z}_{-c})\pi(\boldsymbol{z}_{-c}|\boldsymbol{z}_c)$ is a latent Gaussian model suitable for INLA. This means that conditional models (on $\boldsymbol{z}_c$) can still be fitted with **R-INLA**, i.e., we can obtain marginals of the parameters in $\boldsymbol{z}_{-c}$ given $\boldsymbol{z}_c$. The conditional posterior marginals for the k-th element in vector $\boldsymbol{z}_{-c}$ will be denoted by $\pi(z_{-c,k}|\boldsymbol{z}_c, \boldsymbol{y})$. Also, the conditional marginal likelihood $\pi(\boldsymbol{y}|\boldsymbol{z}_c)$ can be easily computed with **R-INLA**.

4.1 Metropolis-Hastings with INLA

We will now discuss how to implement the Metropolis-Hastings algorithm to estimate the posterior marginal of $\boldsymbol{z}_c$. Note that this is a multivariate