Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations

Håvard Rue, Sara Martino, and Nicolas Chopin

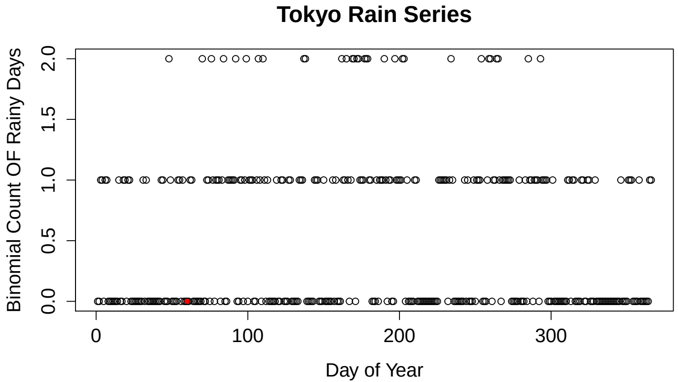
Aaron Zimmerman

Department of Statistics
University of Washington - Seattle

May 23, 2013

Tokyo Rain Dataset

- Binomial time series, taken from Kitagawa(1987)
- Daily binary data from 1983 and 1984.
- Dataset: Rain or No Rain
- 1984 was a leapyear - 731 Bernoulli experiments



Tokyo Rain Model

We assume:

- data is drawn from Bernoulli: $Y_{day,yr} \sim \mathcal{B}(p_{day})$.
- logit-link to latent variables: $p_{day} = logit^{-1}(\theta_{day})$
- latent field (GMRF) follows *circular RW2, with precision* κ

$$\boldsymbol{\theta}|\kappa \sim \mathcal{N}(\mathbf{0}, \kappa Q)$$

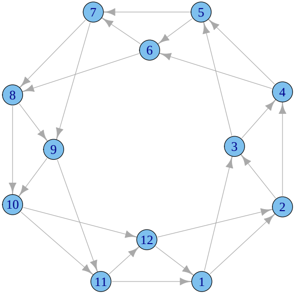
- $\kappa \sim$ *Gamma(1, 0.0001)*

We'd like to find posterior estimates for the latent variables

Tokyo Rain Model

For a RW2, we have

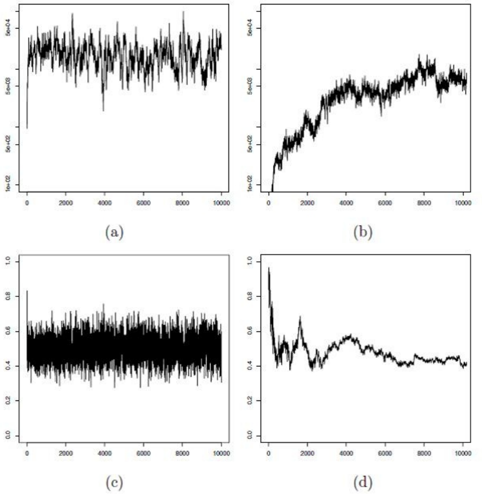
$$Q = \begin{pmatrix} 6 & -4 & 1 & * & * & * & * & \cdots & * & * & 1 & -4 \\ -4 & 6 & -4 & 1 & * & * & * & \cdots & * & * & * & 1 \\ 1 & -4 & 6 & -4 & 1 & * & * & \cdots & * & * & * & * \\ * & 1 & -4 & 6 & -4 & 1 & * & \cdots & * & * & * & * \\ * & * & 1 & -4 & 6 & -4 & 1 & \cdots & * & * & * & * \\ & \vdots & & & & \ddots & & \cdots & & & \vdots & \\ -4 & 1 & * & * & * & * & * & \cdots & * & 1 & -4 & 6 \\ 1 & * & * & * & * & * & * & \cdots & 1 & -4 & 6 & -4 \end{pmatrix}$$



- $\mathrm{E}[\theta_i | \boldsymbol{\theta}_{-i}, \kappa] = \frac{4}{6}(\theta_{i-1} + \theta_{i+1}) - \frac{1}{6}(\theta_{i-2} + \theta_{i+2})$
- $\mathrm{Prec}[\theta_i | \boldsymbol{\theta}_{-i}, \kappa] = 6 * \kappa$

MCMC

We could solve this with MCMC, but it can be tough



Rue and Held (2005)

Latent Gaussian Models

- Assume y_i belongs to exponential family
- Let $E[y_i] = \mu_i$ be linked to a η_i: $g(\mu_i) = \eta_i$

η_i: Structured Additive Predictor

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

- ϵ_is are unstructured terms
- $f^{(j)}$s are unknown functions of the u_{ji}s
- β_ks are linear effects of z_{ki}s
- Define $\boldsymbol{\theta}$ as all $\eta_i, \{f^{(j)}\}, \beta_k$, and α such that

$$\boldsymbol{\theta}|\boldsymbol{\gamma} \sim N(\mathbf{0}, \mathbf{Q}(\boldsymbol{\gamma}))$$

- $\boldsymbol{\gamma}$ are hyperparameters to which we can assign priors

Tokyo Series as LGM

- y_ts are all Bernoulli(p_t)
- $E[y_t] = p_t$ and with the logit link, we have

$$g(p_t) = logit(p_t) = \eta_t \rightarrow p_t = \frac{e^{\eta_t}}{1 + e^{\eta_t}}.$$

- Our structured additive predictor is simple, $\eta_t = f(\theta_t) = \theta_t$
- We have one hyperparameter $\kappa \sim$ Gamma(1, 0.0001)
- and our normal specification on the parameters: $\boldsymbol{\theta}|\kappa \sim \mathcal{N}(\mathbf{0}, \kappa Q)$

INLA in 3 steps

Step 1

- Approximate $\pi(\gamma|\mathbf{y})$ with a Laplace approximation:

$$\tilde{\pi}(\gamma|\mathbf{y}) \propto \left.\frac{\pi(\gamma, \boldsymbol{\theta}, \mathbf{y})}{\tilde{\pi}_G(\boldsymbol{\theta}|\gamma, \mathbf{y})}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*(\gamma)}$$

- Approximate $\pi(\theta_i|\gamma, \mathbf{y})$ with a Laplace approximation
- Numerically integrate out γ from $\pi(\theta_i|\gamma, \mathbf{y})$ to approximate $\pi(\theta_i|\mathbf{y})$

Example - Gaussian Approximation

$$\tilde{\pi}_G(\theta_i|\gamma, \mathbf{y}) = \mathcal{N}(x_i; \mu_i(\gamma), \sigma_i^2(\gamma))$$

INLA in 3 steps

Step 2

- Approximate $\pi(\boldsymbol{\gamma}|\mathbf{y})$ with a Laplace approximation
- Approximate $\pi(\theta_i|\boldsymbol{\gamma}, \mathbf{y})$ with a Laplace approximation:

$$\tilde{\pi}(\theta_i|\boldsymbol{\gamma}, \mathbf{y}) \propto \left.\frac{\pi(\boldsymbol{\gamma}, \boldsymbol{\theta}, \mathbf{y})}{\tilde{\pi}_{GG}(\boldsymbol{\theta}_{-i}|\theta_i, \boldsymbol{\gamma}, \mathbf{y})}\right|_{\boldsymbol{\theta}_{-i}=\boldsymbol{\theta}^*_{-i}(\gamma)}$$

- Numerically integrate out $\boldsymbol{\gamma}$ from $\pi(\theta_i|\boldsymbol{\gamma}, \mathbf{y})$ to approximate $\pi(\theta_i|\mathbf{y})$

INLA in 3 steps

Step 3

- Approximate $\pi(\boldsymbol{\gamma}|\mathbf{y})$ with a Laplace approximation
- Approximate $\pi(\theta_i|\gamma, \mathbf{y})$ with a Laplace approximation
- Numerically integrate out γ from $\pi(\theta_i|\gamma, \mathbf{y})$ to approximate $\pi(\theta_i|\mathbf{y})$:

$$\tilde{\pi}(\theta_i|\mathbf{y}) = \sum_k \tilde{\pi}(\theta_i|\boldsymbol{\gamma}_k, \mathbf{y})\tilde{\pi}(\boldsymbol{\gamma}_k|\mathbf{y})\Delta_k$$

Obtaining $\tilde{\pi}(\gamma|\mathbf{y})$

Our Laplace approximation is:

$$\tilde{\pi}(\gamma|\mathbf{y}) \propto \left. \frac{\pi(\boldsymbol{\theta}, \gamma, \mathbf{y})}{\tilde{\pi}_G(\boldsymbol{\theta}|\gamma, \mathbf{y})} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*(\gamma)},$$

where $\boldsymbol{\theta}^*(\gamma)$ is the mode of the full conditional.

Expanding the joint yields

$$\tilde{\pi}(\gamma|\mathbf{y}) \propto \left. \frac{\pi(\mathbf{y}|\boldsymbol{\theta}, \gamma) * \pi(\boldsymbol{\theta}|\gamma) * \pi(\gamma)}{\tilde{\pi}_G(\boldsymbol{\theta}|\gamma, \mathbf{y})} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*(\gamma)}.$$

Obtain $\tilde{\pi}_G(\boldsymbol{\theta}|\gamma, \mathbf{y})$ for $\tilde{\pi}(\gamma|\mathbf{y})$

By our setup, the latent field is a GMRF, we'd like to keep it this way. As a result:

$$\pi(\boldsymbol{\theta}|\gamma, \mathbf{y}) \propto \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T[\kappa\mathbf{Q}]\boldsymbol{\theta} + \sum \log\pi(y_i|\theta_i, \gamma)\right),$$

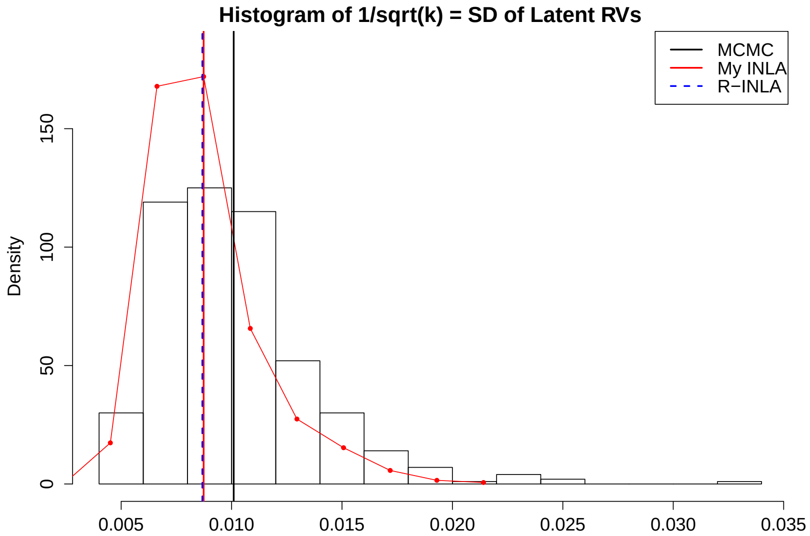
we match the mode and the curvature at the mode to produce the Gaussian approximation:

$$\tilde{\pi}_G(\boldsymbol{\theta}|\gamma, \mathbf{y}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T(\kappa\mathbf{Q} + \mathrm{diag}(c))(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\right),$$

where $\mathbf{Q}$, c, and $\boldsymbol{\theta}^*$ (the mode), are functions of γ.

Exploring $\tilde{\pi}(\gamma|\mathbf{y})$

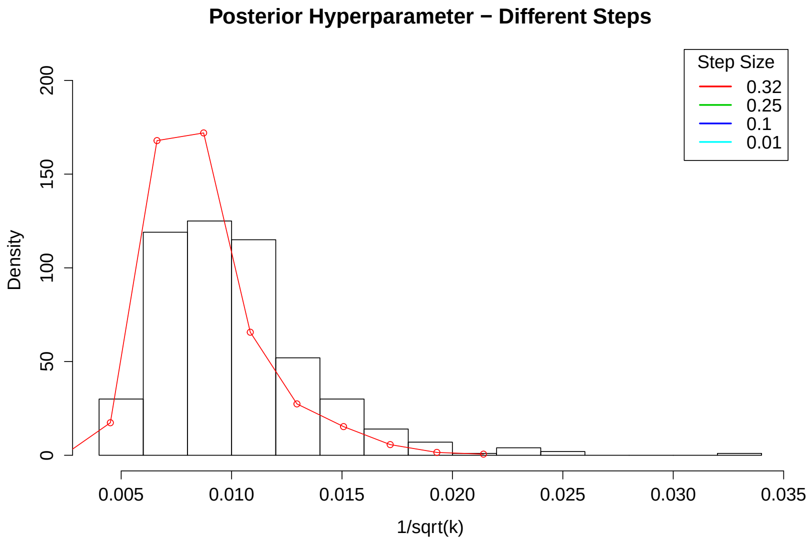
Locate the mode of $log[\tilde{\pi}(\gamma|\mathbf{y})]$ numerically. Move out until $log[\tilde{\pi}(\mathbf{0}|\mathbf{y})] - log[\tilde{\pi}(\gamma|\mathbf{y})] < \delta_\pi$. We save grid points as γ_k.



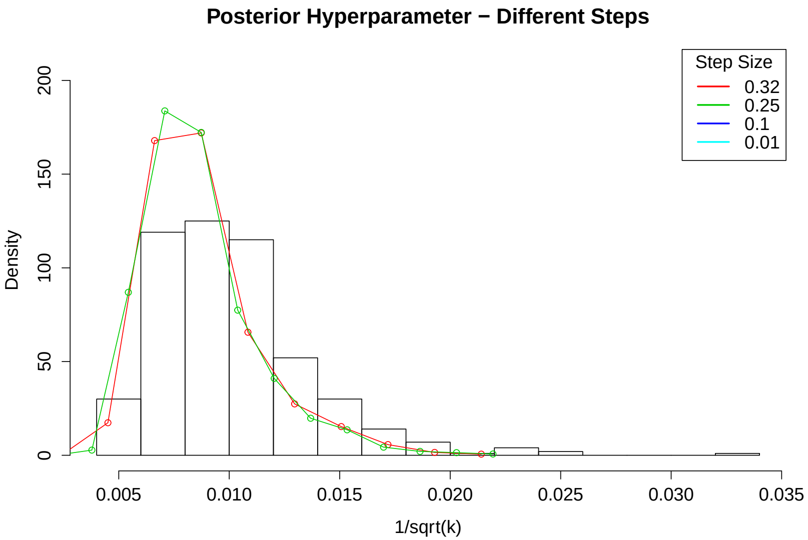
Exploring $\tilde{\pi}(\gamma|\mathbf{y})$

	MCMC	My INLA	R-INLA
κ est.	12978.21	13308.64	13287.47
κ SD	9971.059	-	8962.27
$\frac{1}{\sqrt{\kappa}}$ est.	0.008675	0.008777	0.008668
$\frac{1}{\sqrt{\kappa}}$ SD	0.003432	0.002449	-
total time (sec)	57.14	slow	1.87

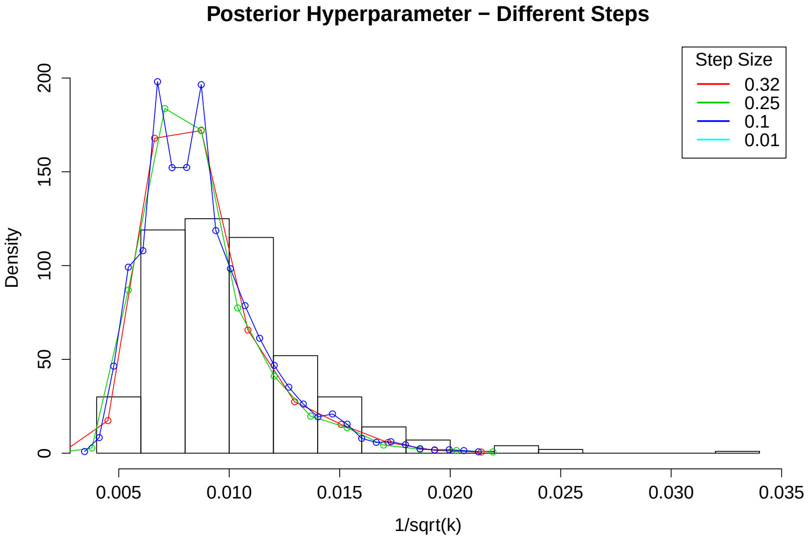
Numerical Stability



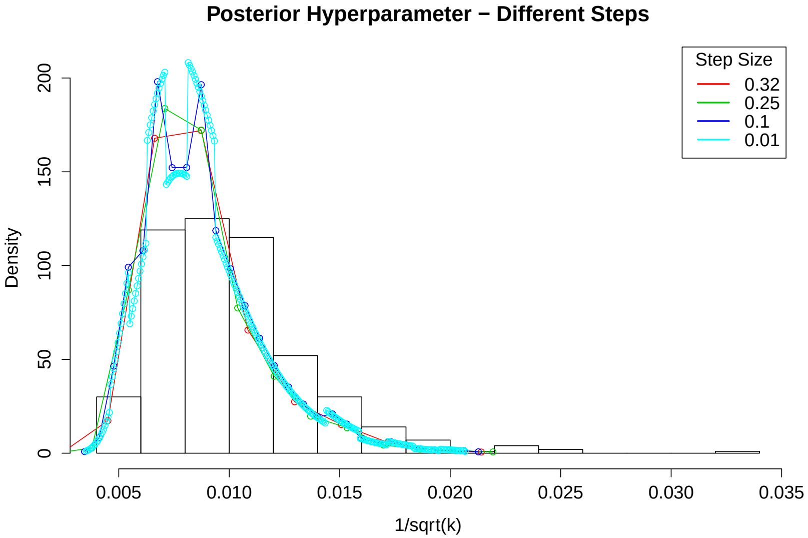
Numerical Stability



Numerical Stability



Numerical Stability



Obtaining $\tilde{\pi}(\theta_i|\gamma_k, \mathbf{y})$

The equivalent Laplace approximation

$$\tilde{\pi}(\theta_i|\gamma_k, \mathbf{y}) \propto \left.\frac{\pi(\boldsymbol{\theta}, \gamma_k, \mathbf{y})}{\tilde{\pi}_{GG}(\boldsymbol{\theta}_{-i}|\theta_i, \gamma_k, \mathbf{y})}\right|_{\boldsymbol{\theta}_{-i}=\boldsymbol{\theta}^*_{-i}(\theta_i,\gamma_k)}$$

is expensive, though it can be done. The denominator is easy in our GMRF. We already have

$$\tilde{\pi}_G(\boldsymbol{\theta}|\gamma_k, \mathbf{y}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T(\kappa\mathbf{Q} + \mathrm{diag}(c))(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\right).$$

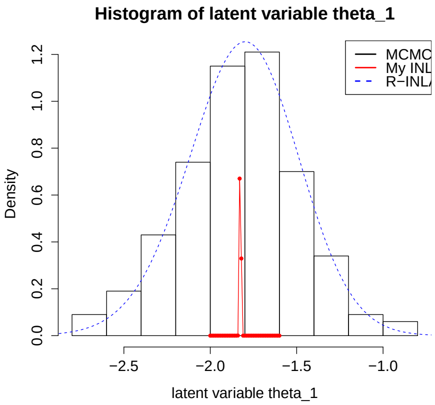
Putting it all together

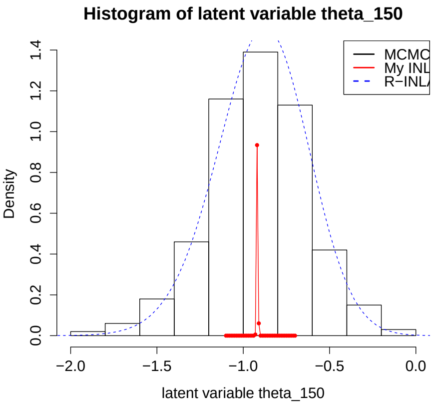
Finally we obtain the marginals of interest for the latent field by numerical integration.

$$\tilde{\pi}(\theta_i|\mathbf{y}) = \sum_k \tilde{\pi}(\theta_i|\boldsymbol{\gamma}_k, \mathbf{y})\tilde{\pi}(\boldsymbol{\gamma}_k|\mathbf{y})\Delta_k$$

will do it.

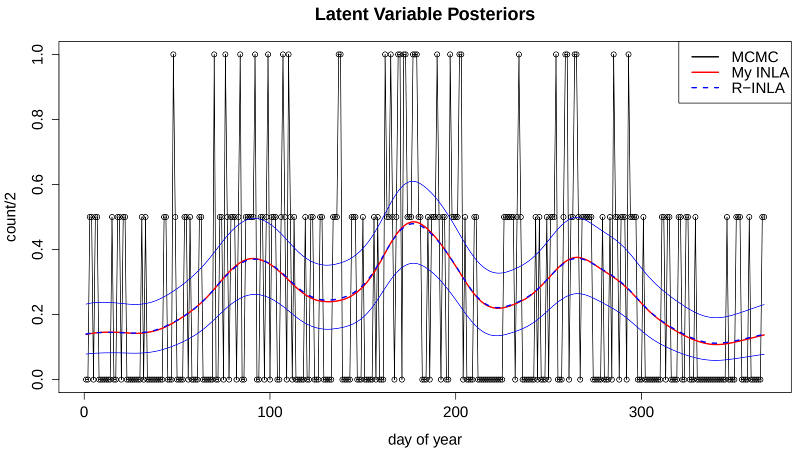
Posterior Results





Remember $\kappa = 13108$, and $\text{Prec}(\theta_i | \boldsymbol{\theta}_{-i}, \kappa) = 6 * \kappa$...

Posterior Latent Variable



In Summary

- Brings the "numerics back into statistics" - Held and Riebler
- INLA can be a very quick tool to avoid MCMC in some situations
- Is it needed? - maybe for exploratory modeling
- Computation cost is exponential in $|\gamma|$
- Numerical stability and speed issues when coding by hand (at least I have them)
- R-INLA is a black box, but I think that it's working, just keep in mind that it has limitations

Thanks everyone!

References I

Besag et al. (1991) *Bayesian Image Restoration with two applications in spatial statistics*

Dey et al. (2000) *Generalized linear models: a bayesian perspective*

Hinton and van Camp. (1993) *Keeping the neural networks simple by minimizing the description length of the weights*

Kitagawa and Gersch. (1996) *Smoothness priors analysis of time series*

Minka. (2001) *Expectation propagation for approximate bayesian inference*

Rue, Martino, Chopin. (2008) *Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations*

Tierney and Kadane. (1986) *Accurate approximations for posterior moments and marginal densities*

Wakefield. (2007) *Disease mapping and spatial regression with count data*

Errors in INLA

- If $p = |\boldsymbol{\theta}| + |\boldsymbol{\gamma}| = n + m$ is fixed and $n_{data} \to \infty$, asymptotic validity of the Laplace approximations hold. e.g. Binomial$(n_i, logit^{-1}(x_i))$ for large n_i.
- If n, and therefore p, grows with n_{data}, then the error rate is $\mathcal{O}(\frac{n}{n_{data}})$Think of the unnormalized version of $\tilde{\pi}(\boldsymbol{\gamma}|\mathbf{y})$ which is defined by an n-dimensional integral. e.g. regression models with individual effects.
- If n/n_{data} is constant (often it's 1), and if as $n_{data} \to \infty$, the true latent field $\boldsymbol{\theta}$ converges to a degenerate Gaussian random distribution of rank q, then the asymptotic error is $\mathcal{O}(q/n_{data})$.

Errors in INLA

- The accuracy of $\tilde{\pi}(\gamma|\mathbf{y})$ seems to depend on the 'actual' dimension of $\boldsymbol{\theta}$. They recommend using the $p_D(\gamma) \equiv$ *effective number of parameters* from DIC theory.
- For instance, if $p_D(\gamma) = 0$, then the approximation rate is exactly zero because the data is non-informative and the posterior is exactly the Gaussian prior.
- In most cases, when you normalize the approximated densities, dominating terms in the numerator and denominator cancel out and this reduces the error rate from $\mathcal{O}(n_{data}^{-1})$ to $\mathcal{O}(n_{data}^{-3/2})$.

Assessing Errors in INLA

- Only thing to do is to compare results to MCMC
- Can also compare different levels accurate approximations (e.g. with SKLD)

$$\text{Gaussian} \rightarrow \text{Simplified Laplace} \rightarrow \text{Full Laplace}$$

Real Data Again

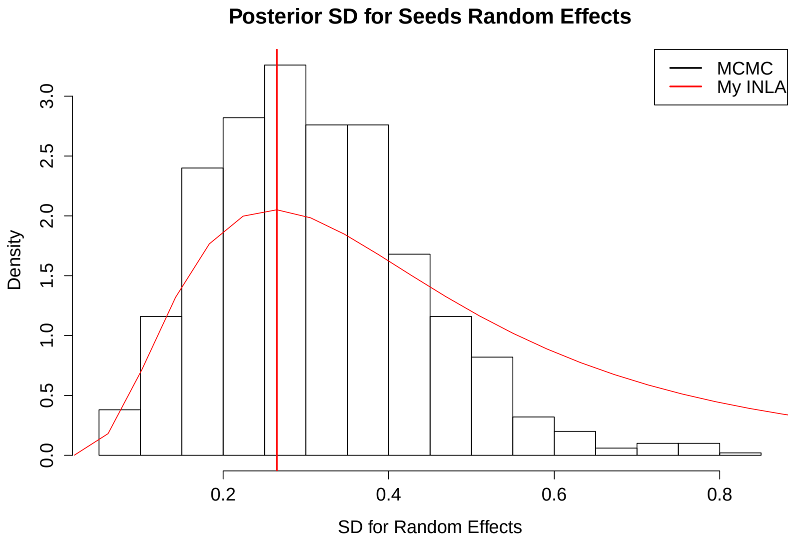
21 plates of seeds. Tracking germination rates.

seed *O. aegyptiaco 75*						seed *O. aegyptiaco 73*					
Bean			Cucumber			Bean			Cucumber		
r	n	r/n	r	n	r/n	r	n	r/n	r	n	r/n
10	39	0.26	5	6	0.83	8	16	0.50	3	12	0.25
23	62	0.37	53	74	0.72	10	30	0.33	22	41	0.54
23	81	0.28	55	72	0.76	8	28	0.29	15	30	0.50
26	51	0.51	32	51	0.63	23	45	0.51	32	51	0.63
17	39	0.44	46	79	0.58	0	4	0.00	3	7	0.43
			10	13	0.77						

- We assume $y_{plate} \sim Binom(n_{plate}, p_{plate})$
- $p_i = logit^{-1}(\alpha_0 + \alpha_1 * x_{1,i} + \alpha_2 * x_{2,i} + \alpha_1 2 * x_{12,i} + r_i)$
- $\alpha. \sim N(0, 1e6)$, and $r_i | \sigma_r \sim N(0, \sigma_r^2)$
- $\tau_r = \frac{1}{\sigma_r^2} \sim \Gamma(0.001, 0.001)$

Crowder (1978).

Posterior Hyperparameter



Posterior Latent