Chapter 4

- 4.1 Normal approximation (Laplace's method)
- 4.2 Large-sample theory
- 4.3 Counter examples
  - includes examples of difficult posteriors for MCMC, too
- 4.4 Frequency evaluation*
- 4.5 Other statistical methods*

Normal approximation (Laplace approximation)

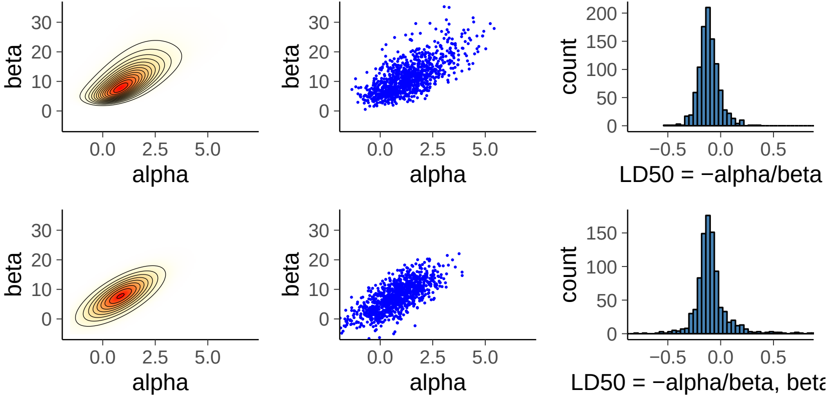
- Often posterior converges to normal distribution when $n \to \infty$
  - bounded, non-singular, the number of parameters don't grow with n
  - we can then approximate $p(\theta|y)$ with normal distribution

Normal approximation (Laplace approximation)

- Often posterior converges to normal distribution when $n \to \infty$
  - bounded, non-singular, the number of parameters don't grow with n
  - we can then approximate $p(\theta|y)$ with normal distribution
  - Laplace used this (before Gauss) to approximate the posterior of binomial model to infer ratio of girls and boys born

Normal approximation (Laplace approximation)

- Often posterior converges to normal distribution when $n \to \infty$
  - bounded, non-singular, the number of parameters don't grow with n
  - we can then approximate $p(\theta|y)$ with normal distribution



Taylor series

- We can approximate $p(\theta|y)$ with normal distribution

$$p(\theta|y) \approx \frac{1}{\sqrt{2\pi}\sigma_\theta} \exp\left(-\frac{1}{2\sigma_\theta^2}(\theta - \hat{\theta})^2\right)$$

  - i.e. log posterior $\log p(\theta|y)$ can be approximated with a quadratic function

$$\log p(\theta|y) \approx \alpha(\theta - \hat{\theta})^2 + C$$

Taylor series

- We can approximate $p(\theta|y)$ with normal distribution

$$p(\theta|y) \approx \frac{1}{\sqrt{2\pi}\sigma_\theta} \exp\left(-\frac{1}{2\sigma_\theta^2}(\theta - \hat{\theta})^2\right)$$

  - i.e. log posterior $\log p(\theta|y)$ can be approximated with a quadratic function

$$\log p(\theta|y) \approx \alpha(\theta - \hat{\theta})^2 + C$$

- Corresponds to Taylor series expansion around $\theta = \hat{\theta}$

$$f(\theta) = f(\hat{\theta}) + f'(\hat{\theta})(\theta - \hat{\theta}) + \frac{f''(\hat{\theta})}{2!}(\theta - \hat{\theta})^2 + \frac{f^{(3)}(\hat{\theta})}{3!}(\theta - \hat{\theta})^3 + \dots$$

Taylor series

- We can approximate $p(\theta|y)$ with normal distribution

$$p(\theta|y) \approx \frac{1}{\sqrt{2\pi}\sigma_\theta} \exp\left(-\frac{1}{2\sigma_\theta^2}(\theta - \hat{\theta})^2\right)$$

  - i.e. log posterior $\log p(\theta|y)$ can be approximated with a quadratic function

$$\log p(\theta|y) \approx \alpha(\theta - \hat{\theta})^2 + C$$

- Corresponds to Taylor series expansion around $\theta = \hat{\theta}$

$$f(\theta) = f(\hat{\theta}) + f'(\hat{\theta})(\theta - \hat{\theta}) + \frac{f''(\hat{\theta})}{2!}(\theta - \hat{\theta})^2 + \frac{f^{(3)}(\hat{\theta})}{3!}(\theta - \hat{\theta})^3 + \ldots$$

  - if $\hat{\theta}$ is at mode, then $f'(\hat{\theta}) = 0$

Taylor series

- We can approximate $p(\theta|y)$ with normal distribution

$$p(\theta|y) \approx \frac{1}{\sqrt{2\pi}\sigma_\theta} \exp\left(-\frac{1}{2\sigma_\theta^2}(\theta - \hat{\theta})^2\right)$$

  - i.e. log posterior $\log p(\theta|y)$ can be approximated with a quadratic function

$$\log p(\theta|y) \approx \alpha(\theta - \hat{\theta})^2 + C$$

- Corresponds to Taylor series expansion around $\theta = \hat{\theta}$

$$f(\theta) = f(\hat{\theta}) + f'(\hat{\theta})(\theta - \hat{\theta}) + \frac{f''(\hat{\theta})}{2!}(\theta - \hat{\theta})^2 + \frac{f^{(3)}(\hat{\theta})}{3!}(\theta - \hat{\theta})^3 + \ldots$$

  - if $\hat{\theta}$ is at mode, then $f'(\hat{\theta}) = 0$
  - often when $n \to \infty$, $\frac{f^{(3)}(\hat{\theta})}{3!}(\theta - \hat{\theta})^3 + \ldots$ is small

Multivariate Taylor series

- Multivariate series expansion

$$f(\theta) = f(\hat{\theta}) + \frac{df(\theta')}{d\theta'}_{\theta'=\hat{\theta}}(\theta - \hat{\theta}) + \frac{1}{2!}(\theta - \hat{\theta})^T \frac{d^2 f(\theta')}{d\theta'^2}_{\theta'=\hat{\theta}}(\theta - \hat{\theta}) + \dots$$

Normal approximation

- Taylor series expansion of the log posterior around the posterior mode $\hat{\theta}$

$$\log p(\theta|y) = \log p(\hat{\theta}|y) + \frac{1}{2}(\theta - \hat{\theta})^T \left[\frac{d^2}{d\theta^2} \log p(\theta'|y)\right]_{\theta'=\hat{\theta}} (\theta - \hat{\theta}) + \ldots$$

Normal approximation

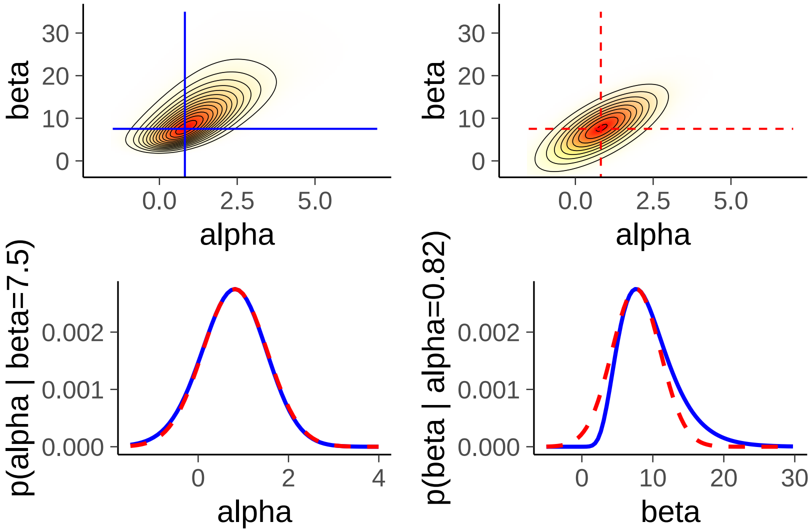
- Taylor series expansion of the log posterior around the posterior mode $\hat{\theta}$

$$\log p(\theta|y) = \log p(\hat{\theta}|y) + \frac{1}{2}(\theta-\hat{\theta})^T \left[\frac{d^2}{d\theta^2}\log p(\theta'|y)\right]_{\theta'=\hat{\theta}} (\theta-\hat{\theta}) + \ldots$$

- Multivariate normal $\propto |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(\theta - \hat{\theta}^T)\Sigma^{-1}(\theta - \hat{\theta})\right)$

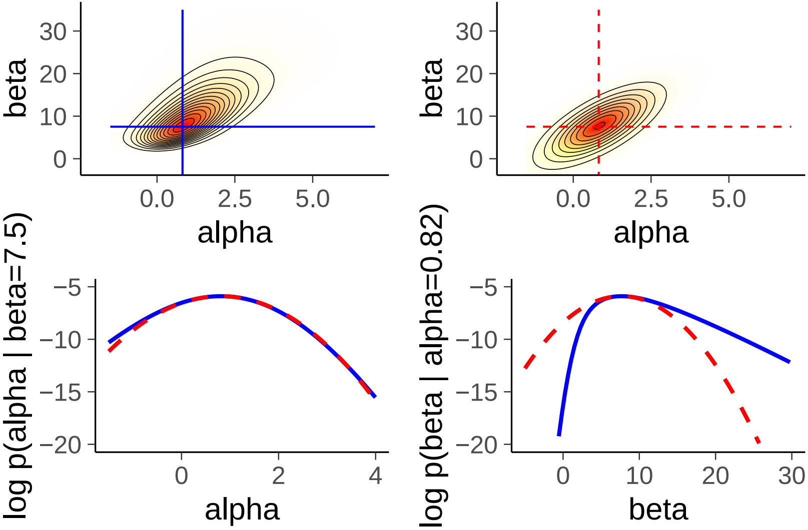
Normal approximation

- Multivariate normal $\propto |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(\theta - \hat{\theta}^T)\Sigma^{-1}(\theta - \hat{\theta})\right)$



Normal approximation

- Multivariate normal $\propto |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(\theta - \hat{\theta}^T)\Sigma^{-1}(\theta - \hat{\theta})\right)$



Normal approximation

- Normal approximation

$$p(\theta|y) \approx \mathrm{N}(\hat{\theta}, [I(\hat{\theta})]^{-1})$$

where $I(\theta)$ is called *observed information*

$$I(\theta) = -\frac{d^2}{d\theta^2} \log p(\theta|y)$$

Normal approximation

- Normal approximation

$$p(\theta|y) \approx \mathrm{N}(\hat{\theta}, [I(\hat{\theta})]^{-1})$$

where $I(\theta)$ is called *observed information*

$$I(\theta) = -\frac{d^2}{d\theta^2} \log p(\theta|y)$$

Hessian $H(\theta) = -I(\theta)$

Normal approximation

- $I(\theta)$ is called *observed information*

$$I(\theta) = -\frac{d^2}{d\theta^2} \log p(\theta|y)$$

  - $I(\hat{\theta})$ is the second derivatives at the mode and thus describes the curvature at the mode
  - if the mode is inside the parameter space, $I(\hat{\theta})$ is positive
  - if θ is a vector, then $I(\theta)$ is a matrix

Normal approximation

- BDA3 Ch 4 has an example where it is easy to compute first and second derivatives and there is easy analytic solution to find where the first derivatives are zero

Normal approximation – numerically

- Normal approximation can be computed numerically
  - iterative optimization to find a mode (may use gradients)
  - autodiff or finite-difference for gradients and Hessian

Normal approximation – numerically

- Normal approximation can be computed numerically
  - iterative optimization to find a mode (may use gradients)
  - autodiff or finite-difference for gradients and Hessian
  - e.g. in R, demo4_1.R:

```
bioassayfun <- function(w, df) {
  z <- w[1] + w[2]*df$x
  -sum(df$y*(z) - df$n*log1p(exp(z)))
}

theta0 <- c(0,0)
optimres <- optim(w0, bioassayfun, gr=NULL, df1, hessi
thetahat <- optimres$par
Sigma <- solve(optimres$hessian)
```

Normal approximation – numerically

- Normal approximation can be computed numerically
  - iterative optimization to find a mode (may use gradients)
  - autodiff or finite-difference for gradients and Hessian
- CmdStan(R) has Laplace algorithm

Normal approximation – numerically

- Normal approximation can be computed numerically
  - iterative optimization to find a mode (may use gradients)
  - autodiff or finite-difference for gradients and Hessian
- CmdStan(R) has Laplace algorithm
  - uses L-BFGS quasi-Newton optimization algorithm for finding the mode
  - uses autodiff for gradients
  - uses finite differences of gradients to compute Hessian

Normal approximation – numerically

- Normal approximation can be computed numerically
  - iterative optimization to find a mode (may use gradients)
  - autodiff or finite-difference for gradients and Hessian
- CmdStan(R) has Laplace algorithm
  - uses L-BFGS quasi-Newton optimization algorithm for finding the mode
  - uses autodiff for gradients
  - uses finite differences of gradients to compute Hessian
    - second order autodiff in progress

Normal approximation

- Optimization and computation of Hessian requires usually much fewer density evaluations than MCMC

Normal approximation

- Optimization and computation of Hessian requires usually much fewer density evaluations than MCMC
- In some cases accuracy is sufficient

Normal approximation

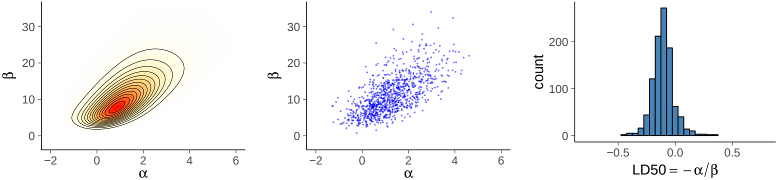
- Optimization and computation of Hessian requires usually much fewer density evaluations than MCMC
- In some cases accuracy is sufficient
- In some cases accuracy for a conditional distribution is sufficient (Ch 13)
  - e.g. Gaussian latent variable models, such as Gaussian processes (Ch 21) and Gaussian Markov random fields
  - Rasmussen & Williams: Gaussian Processes for Machine Learning
  - CS-E4895 - Gaussian Processes (in spring)

Normal approximation

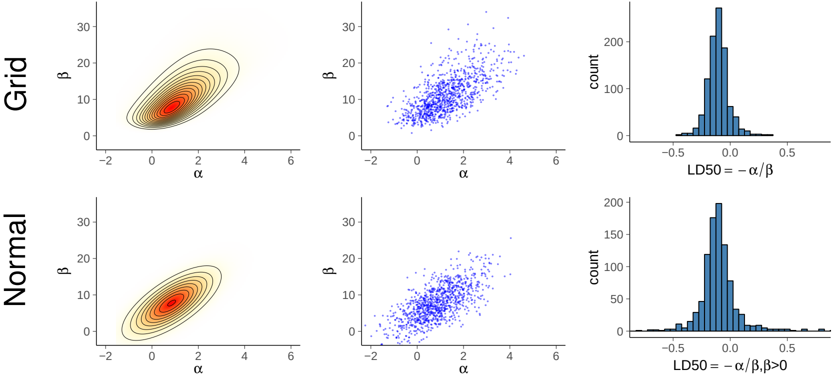
- Optimization and computation of Hessian requires usually much fewer density evaluations than MCMC
- In some cases accuracy is sufficient
- In some cases accuracy for a conditional distribution is sufficient (Ch 13)
  - e.g. Gaussian latent variable models, such as Gaussian processes (Ch 21) and Gaussian Markov random fields
  - Rasmussen & Williams: Gaussian Processes for Machine Learning
  - CS-E4895 - Gaussian Processes (in spring)
- Accuracy can be improved by importance sampling (Ch 10)

Example: Importance sampling in Bioassay

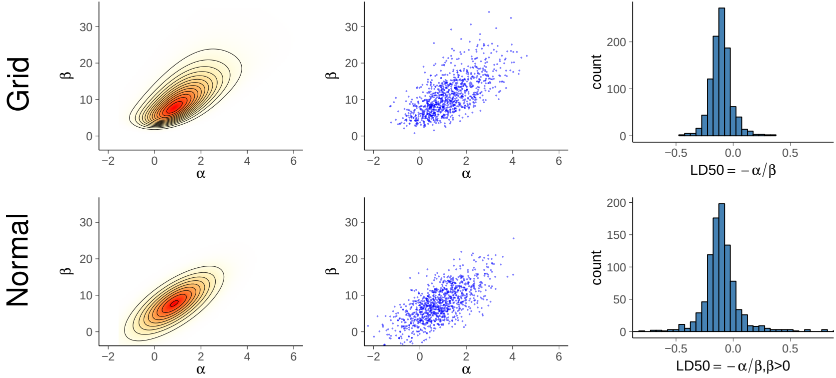
Grid



Example: Importance sampling in Bioassay

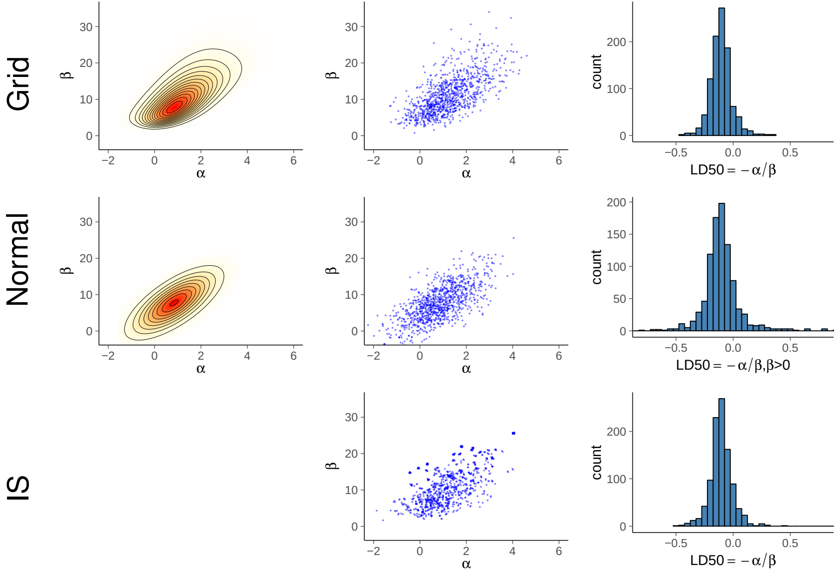


Example: Importance sampling in Bioassay

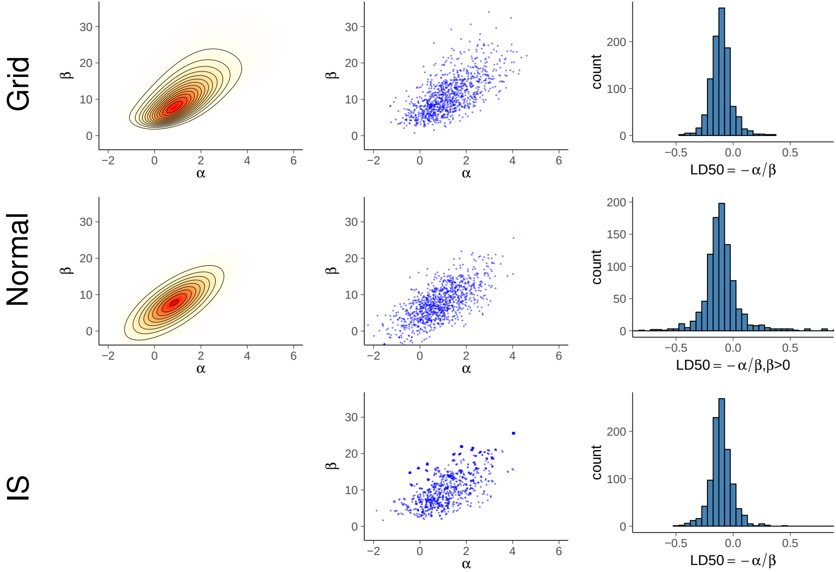


But the normal approximation is not that good here:
Grid sd(LD50) $\approx$ 0.1, Normal sd(LD50) $\approx$ .75!

Example: Importance sampling in Bioassay



Example: Importance sampling in Bioassay



Grid sd(LD50) $\approx$ 0.1, IS sd(LD50) $\approx$ 0.1

Normal approximation

- Accuracy can be improved by importance sampling
- Pareto-k diagnostic of importance sampling weights can be used for diagnostic
  - in Bioassay example $k = 0.57$, which is ok

Normal approximation

- Accuracy can be improved by importance sampling
- Pareto-*k* diagnostic of importance sampling weights can be used for diagnostic
  - in Bioassay example $k = 0.57$, which is ok
- CmdStan(R) has Laplace algorithm
  - since version 2.33 (2023)
    - + Pareto-*k* diagnostic via posterior package
    - + importance resampling (IR) via posterior package

Normal approximation and parameter transformations

- Normal approximation is not good for parameters with bounded or half-bounded support
  - e.g. $\theta \in [0, 1]$ presenting probability

Normal approximation and parameter transformations

- Normal approximation is not good for parameters with bounded or half-bounded support
  - e.g. $\theta \in [0, 1]$ presenting probability
  - Stan code can include constraints

```
real<lower=0,upper=0> theta;
```

Normal approximation and parameter transformations

- Normal approximation is not good for parameters with bounded or half-bounded support
  - e.g. $\theta \in [0, 1]$ presenting probability
  - Stan code can include constraints
    ```
    real<lower=0,upper=0> theta;
    ```
  - for this, Stan does the inference in unconstrained space using logit transformation

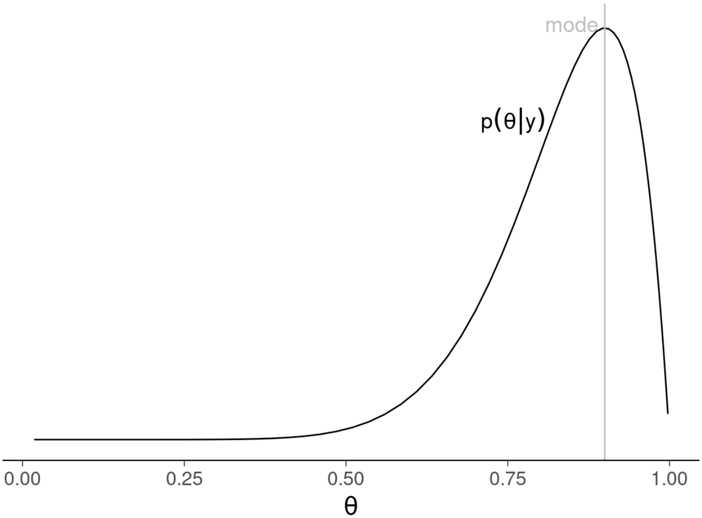
Normal approximation and parameter transformations

- Normal approximation is not good for parameters with bounded or half-bounded support
  - e.g. $\theta \in [0, 1]$ presenting probability
  - Stan code can include constraints
    ```
    real<lower=0,upper=0> theta;
    ```
  - for this, Stan does the inference in unconstrained space using logit transformation
  - density of the transformed parameter needs to include Jacobian of the transformation (BDA3 p. 21)

Normal approximation and parameter transformations

Binomial model $y \sim \mathrm{Bin}(\theta, N)$, with data $y = 9, N = 10$

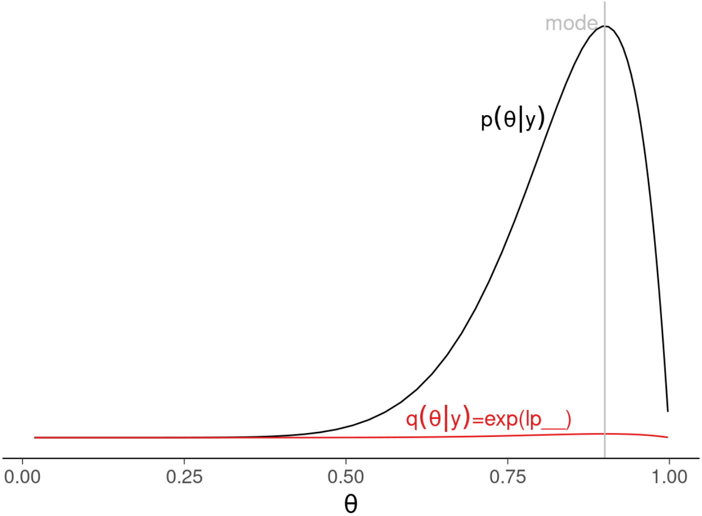
With $\mathrm{Beta}(1, 1)$ prior, the posterior is $\mathrm{Beta}(9 + 1, 1 + 1)$



Normal approximation and parameter transformations

With $\mathrm{Beta}(1,1)$ prior, the posterior is $\mathrm{Beta}(9+1,1+1)$

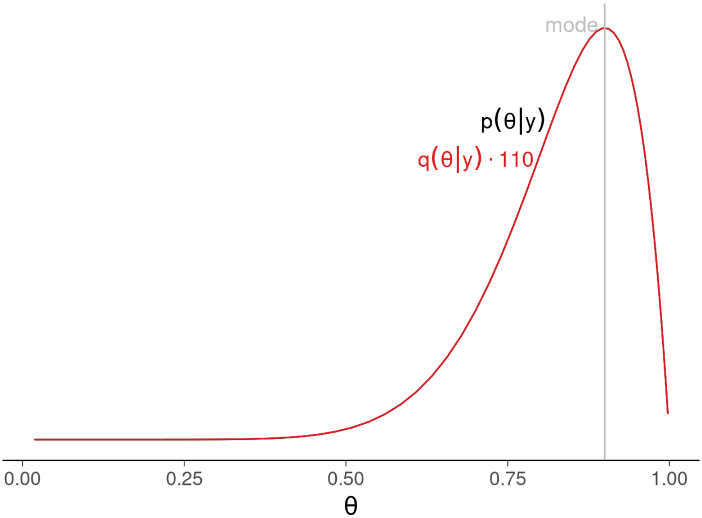
Stan computes only the unnormalized posterior $q(\theta|y)$



Normal approximation and parameter transformations

With $\mathrm{Beta}(1,1)$ prior, the posterior is $\mathrm{Beta}(9+1,1+1)$

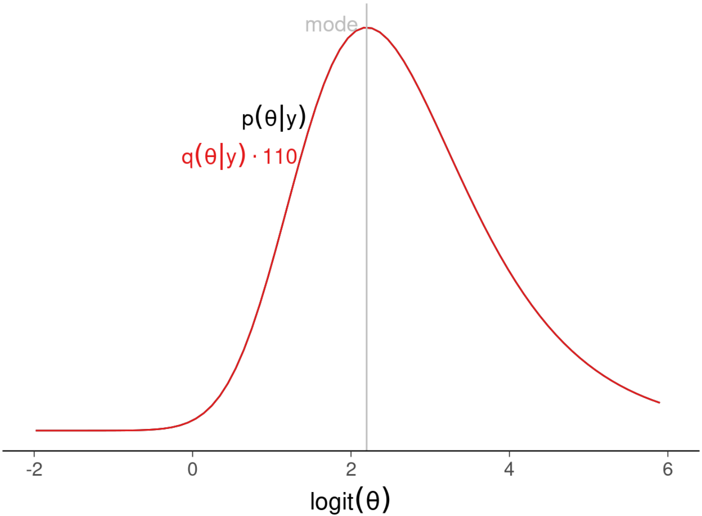
For illustration purposes we normalize Stan result $q(\theta|y)$



Normal approximation and parameter transformations

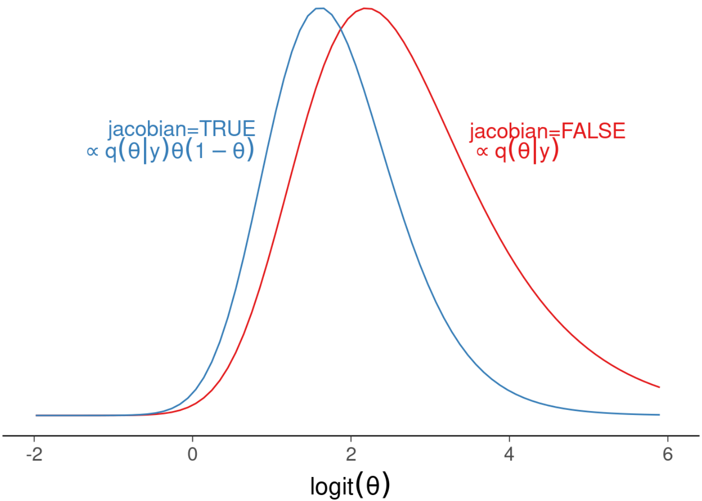
With $\text{Beta}(1, 1)$ prior, the posterior is $\text{Beta}(9 + 1, 1 + 1)$

$\text{Beta}(9 + 1, 1 + 1)$, but x-axis shows the unconstrained $\text{logit}(\theta)$



Normal approximation and parameter transformations

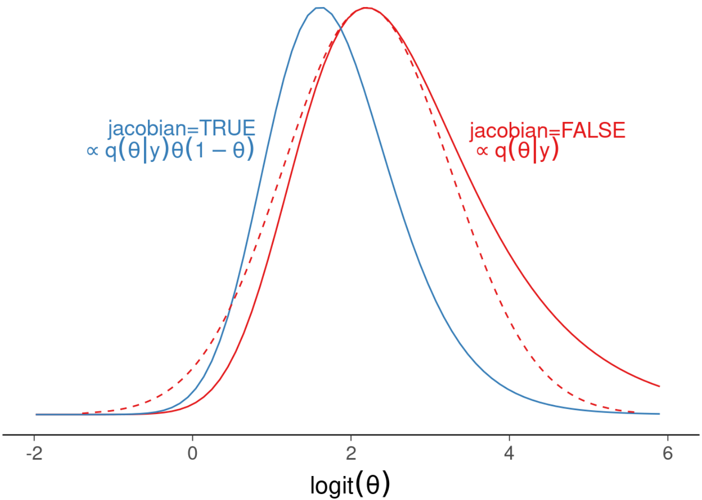
...but we need to take into account the absolute value of the determinant of the Jacobian of the transformation $\theta(1-\theta)$



Normal approximation and parameter transformations

...but we need to take into account Jacobian $\theta(1-\theta)$

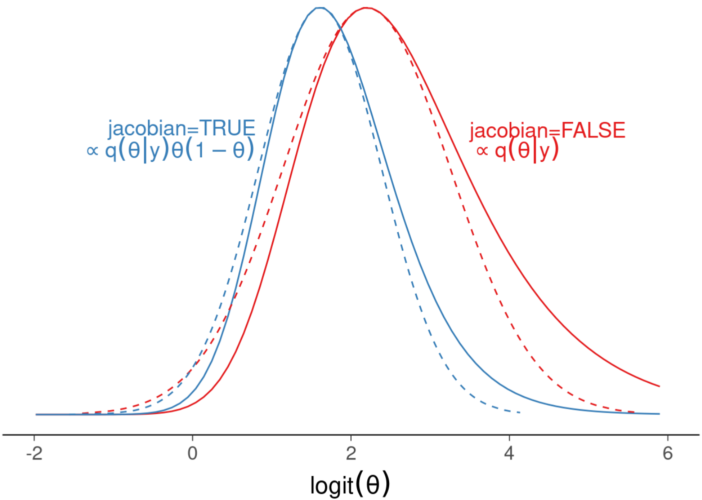
Let's compare a wrong normal approximation...



Normal approximation and parameter transformations

...but we need to take into account Jacobian $\theta(1-\theta)$

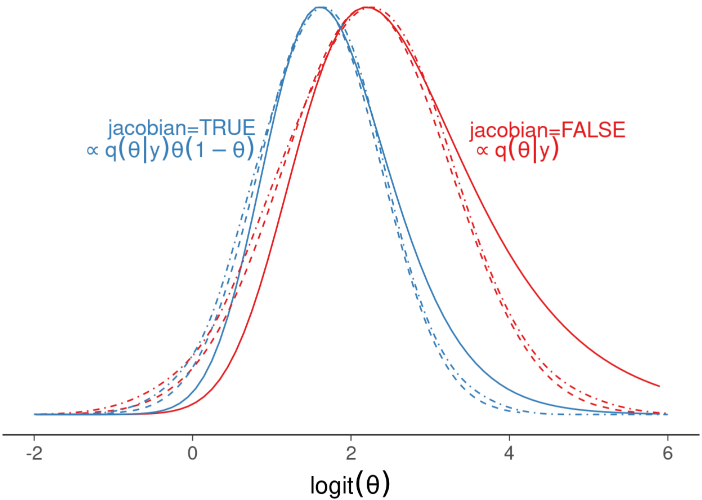
Let's compare a wrong normal approximation and correct one



Normal approximation and parameter transformations

Let's compare a wrong normal approximation and correct one

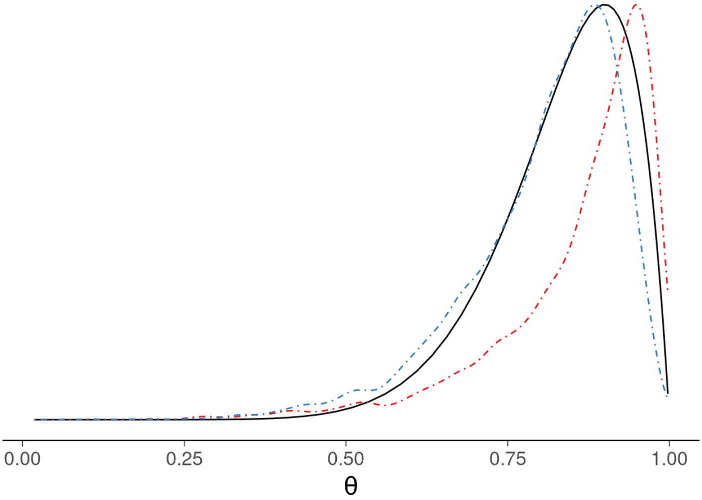
Sample from both approximations and show KDEs for draws



Normal approximation and parameter transformations

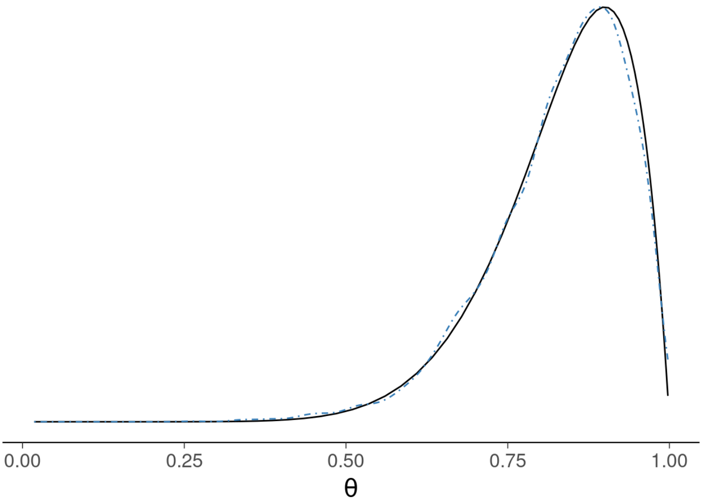
Let's compare a wrong normal approximation and correct one

Inverse transform draws and show KDEs



Normal approximation and parameter transformations

Laplace approximation can be further improved with importance resampling



Other distributional approximations

- Higher order derivatives at the mode can be used

Other distributional approximations

- Higher order derivatives at the mode can be used
- Split-normal and split-t by Geweke (1989) use additional scaling along different principal axes

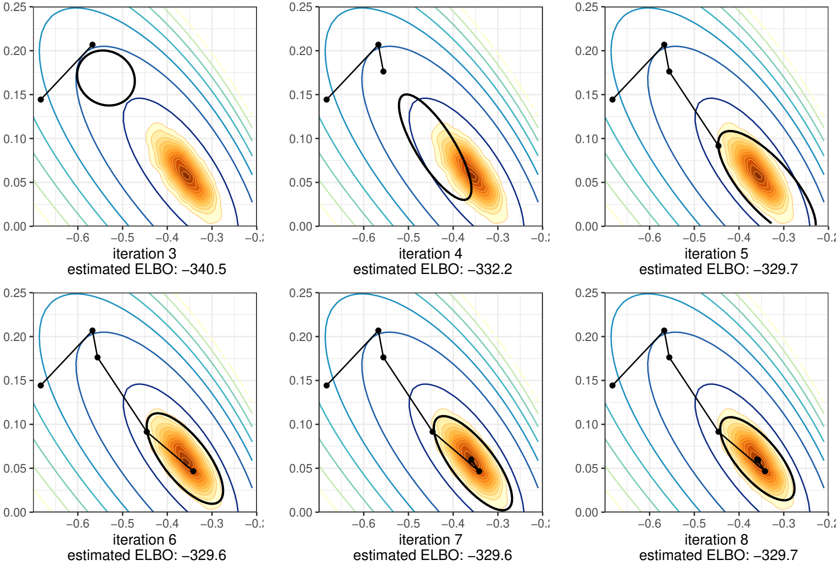
Other distributional approximations

- Higher order derivatives at the mode can be used
- Split-normal and split-t by Geweke (1989) use additional scaling along different principal axes
- Other distributions can be used (e.g. t-distribution)

Other distributional approximations

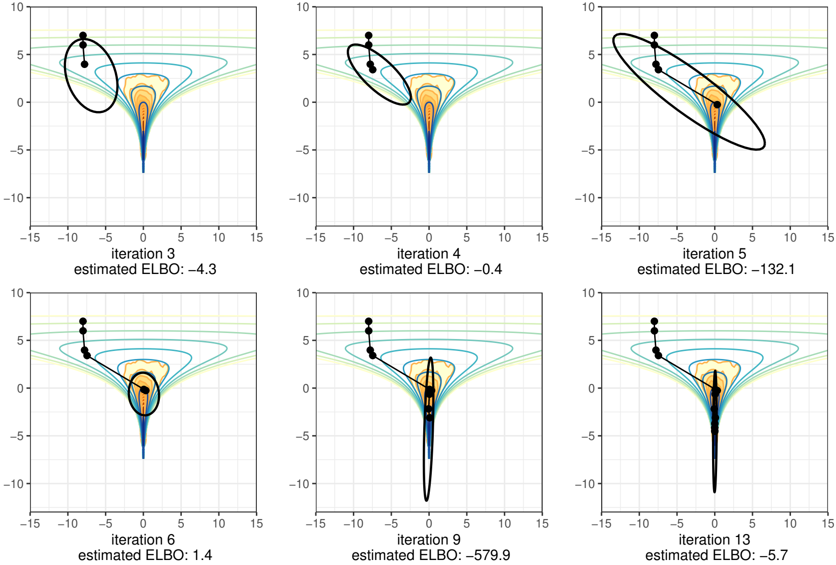
- Higher order derivatives at the mode can be used
- Split-normal and split-t by Geweke (1989) use additional scaling along different principal axes
- Other distributions can be used (e.g. t-distribution)
- Instead of mode and Hessian at mode, e.g.
  - variational inference (Ch 13)
    - CS-E4820 - Machine Learning: Advanced Probabilistic Methods
    - CS-E4895 - Gaussian Processes
    - Stan has the ADVI algorithm (not very good implementation)
    - Stan has Pathfinder algorithm (CmdStanR, brms)
    - instead of normal, methods with flexible flow transformations
  - expectation propagation (Ch 13)
  - speed of these is usually between optimization and MCMC
    - stochastic variational inference can be even slower than MCMC

Pathfinder: Parallel quasi-Newton variational inference.



Zhang, Carpenter, Gelman, and Vehtari (2022). Pathfinder: Parallel quasi-Newton variational inference. *Journal of Machine Learning Research*, 23(306):1–49.

Pathfinder: Parallel quasi-Newton variational inference.

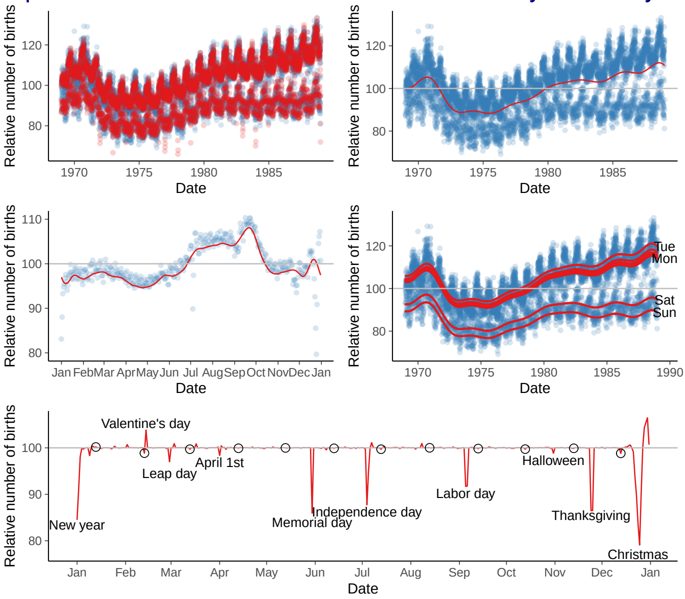


Zhang, Carpenter, Gelman, and Vehtari (2022). Pathfinder: Parallel quasi-Newton variational inference. *Journal of Machine Learning Research*, 23(306):1–49.

Pathfinder: Parallel quasi-Newton variational inference.

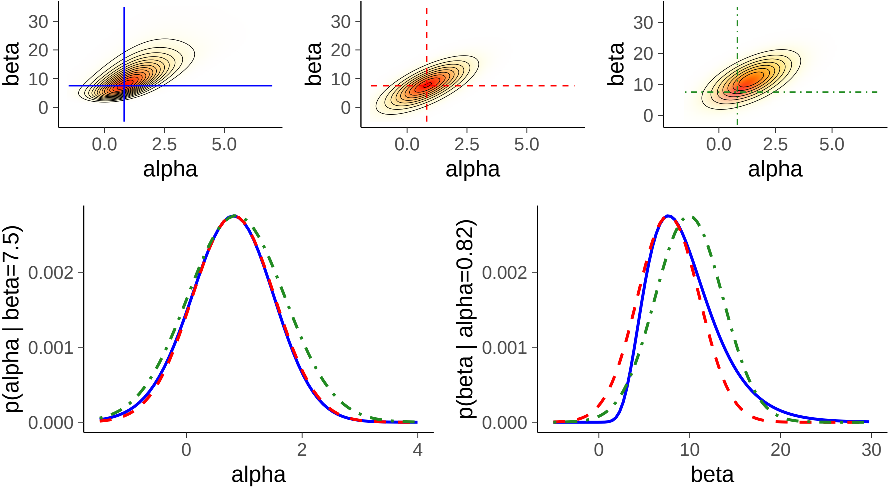
Birthdays case study uses Pathfinder to speed up workflow

https://users.aalto.fi/~ave/casestudies/Birthdays/birthdays.html



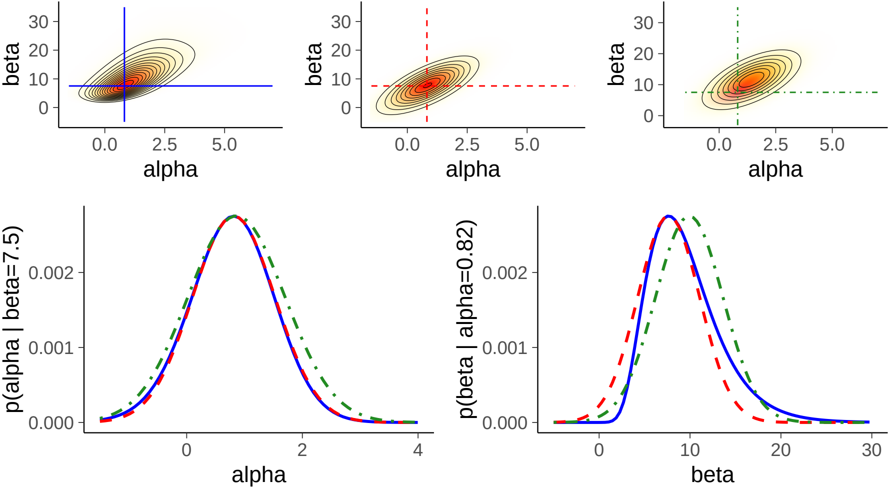
Distributional approximations

Exact, Normal at mode, Normal with variational inference



Distributional approximations

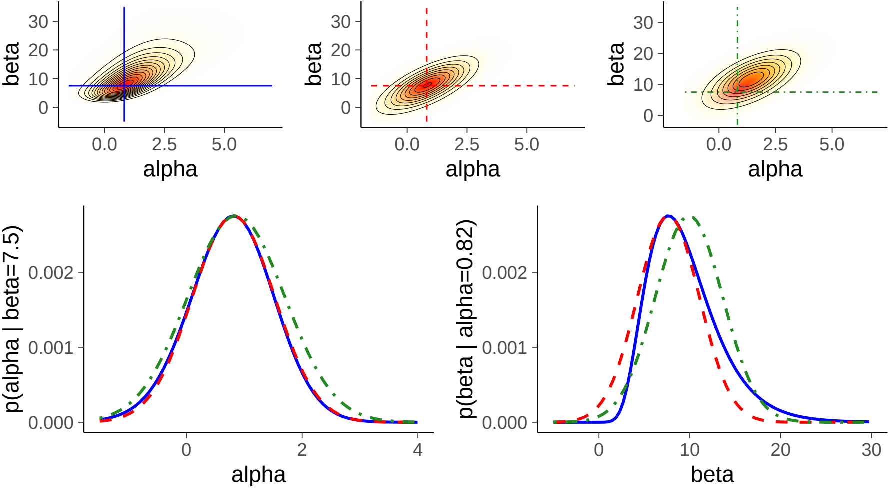
Exact, Normal at mode, Normal with variational inference



Grid sd(LD50) $\approx$ 0.090,
Normal sd(LD50) $\approx$ .75, Normal + IR sd(LD50) $\approx$ 0.096 (Pareto-k = 0.57)

Distributional approximations

Exact, Normal at mode, Normal with variational inference



Grid sd(LD50) $\approx$ 0.090,
Normal sd(LD50) $\approx$ .75, Normal + IR sd(LD50) $\approx$ 0.096 (Pareto-k = 0.57)
VI sd(LD50) $\approx$ 0.13, VI + IR sd(LD50) $\approx$ 0.095 (Pareto-k = 0.17)

Variational inference

- Variational inference includes a large number of methods

Variational inference

- Variational inference includes a large number of methods
- For a restricted set of models, possible to derive deterministic algorithms
  - can be fast and can be relatively accurate

Variational inference

- Variational inference includes a large number of methods
- For a restricted set of models, possible to derive deterministic algorithms
  - can be fast and can be relatively accurate
- Using stochastic (Monte Carlo) estimation of the divergence, possible to derive generic black box algorithms

Variational inference

- Variational inference includes a large number of methods
- For a restricted set of models, possible to derive deterministic algorithms
  - can be fast and can be relatively accurate
- Using stochastic (Monte Carlo) estimation of the divergence, possible to derive generic black box algorithms
  - possible to use use also mini-batching

Variational inference

- Variational inference includes a large number of methods
- For a restricted set of models, possible to derive deterministic algorithms
  - can be fast and can be relatively accurate
- Using stochastic (Monte Carlo) estimation of the divergence, possible to derive generic black box algorithms
  - possible to use use also mini-batching
  - can be fast and provide better predictive distribution than Laplace approximation if the posterior is far from normal

Variational inference

- Variational inference includes a large number of methods
- For a restricted set of models, possible to derive deterministic algorithms
  - can be fast and can be relatively accurate
- Using stochastic (Monte Carlo) estimation of the divergence, possible to derive generic black box algorithms
  - possible to use use also mini-batching
  - can be fast and provide better predictive distribution than Laplace approximation if the posterior is far from normal
  - in general, unlikely to achieve accuracy of HMC with the same computation cost

Variational inference

- Variational inference includes a large number of methods
- For a restricted set of models, possible to derive deterministic algorithms
  - can be fast and can be relatively accurate
- Using stochastic (Monte Carlo) estimation of the divergence, possible to derive generic black box algorithms
  - possible to use use also mini-batching
  - can be fast and provide better predictive distribution than Laplace approximation if the posterior is far from normal
  - in general, unlikely to achieve accuracy of HMC with the same computation cost
  - with increasing number of posterior dimensions, the obtained approximation gets worse (Dhaka, Catalina, Andersen, Magnusson, Huggins, and Vehtari, 2020)

Variational inference

- Variational inference includes a large number of methods
- For a restricted set of models, possible to derive deterministic algorithms
  - can be fast and can be relatively accurate
- Using stochastic (Monte Carlo) estimation of the divergence, possible to derive generic black box algorithms
  - possible to use use also mini-batching
  - can be fast and provide better predictive distribution than Laplace approximation if the posterior is far from normal
  - in general, unlikely to achieve accuracy of HMC with the same computation cost
  - with increasing number of posterior dimensions, the obtained approximation gets worse (Dhaka, Catalina, Andersen, Magnusson, Huggins, and Vehtari, 2020)
  - with increasing number of posterior dimensions, the stochastic divergence estimate gets worse and flows have problems, too (Dhaka, Catalina, Andersen, Welandawe, Huggins, and Vehtari, 2021)

brms supports Laplace / Pathfinder / ADVI

These might be useful for initializng MCMC or big data. The ADVI implementation is not very good.

```
fit1 <- brm(..., algorithm = "laplace")

fit1 <- brm(..., algorithm = "pathfinder")

fit1 <- brm(..., algorithm = "meanfield")

fit1 <- brm(..., algorithm = "fullrank")
```