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)

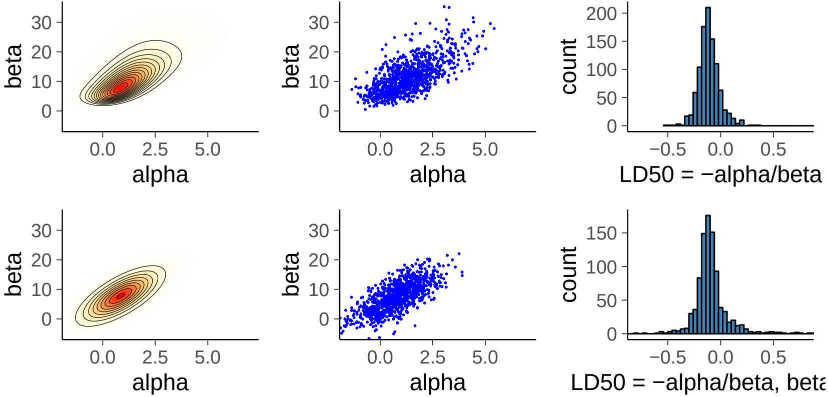
- Often posterior converges to normal distribution when $n \to \infty$
  - bounded, non-singular, the number of parameters don't grow with n
  - we can then approximate $p(\theta|y)$ with normal distribution

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 $\theta = \hat{\theta}$

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

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

Multivariate Taylor series

- Multivariate series expansion

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

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

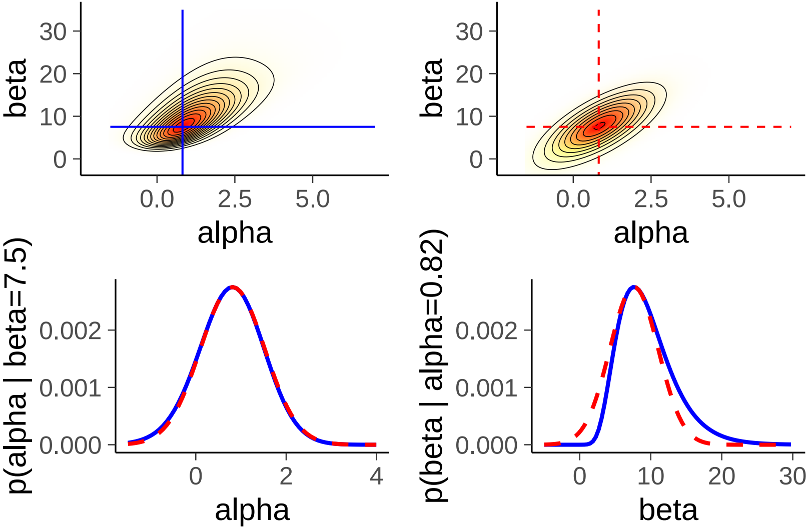
- Taylor series expansion of the log posterior around the posterior mode $\hat{\theta}$

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

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

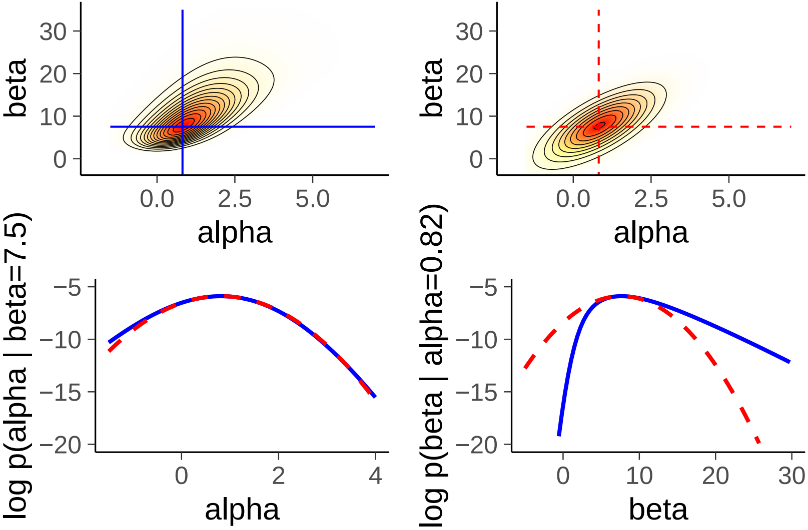
Normal approximation

- 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, hessian=T)
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 less density evaluations than MCMC

Normal approximation

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

Normal approximation

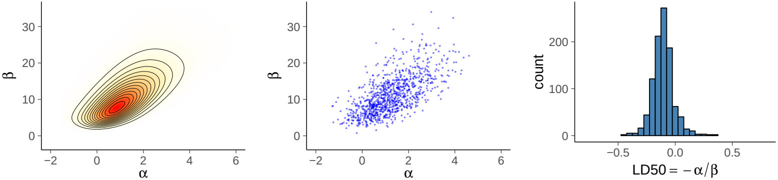
- Optimization and computation of Hessian requires usually much less density evaluations than MCMC
- In some cases accuracy is sufficient
- In some cases accuracy for a conditional distribution is sufficient (Ch 13)
  - e.g. Gaussian latent variable models, such as Gaussian processes (Ch 21) and Gaussian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

Normal approximation

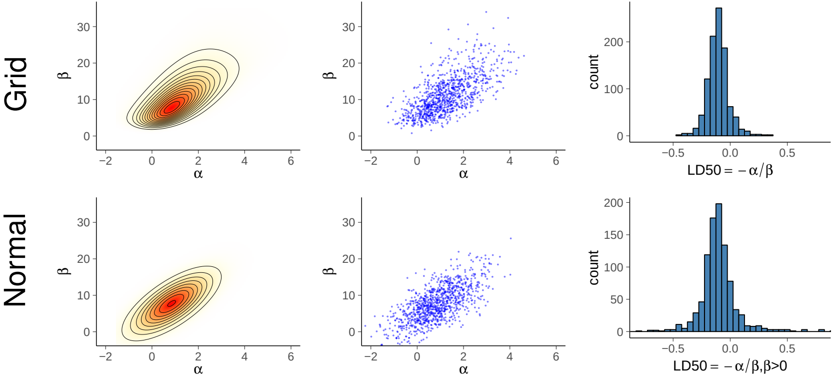
- Optimization and computation of Hessian requires usually much less density evaluations than MCMC
- In some cases accuracy is sufficient
- In some cases accuracy for a conditional distribution is sufficient (Ch 13)
  - e.g. Gaussian latent variable models, such as Gaussian processes (Ch 21) and Gaussian 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

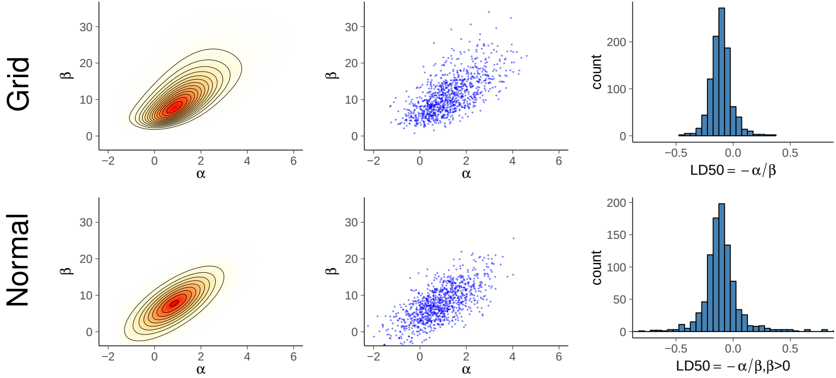
Grid



Example: Importance sampling in Bioassay

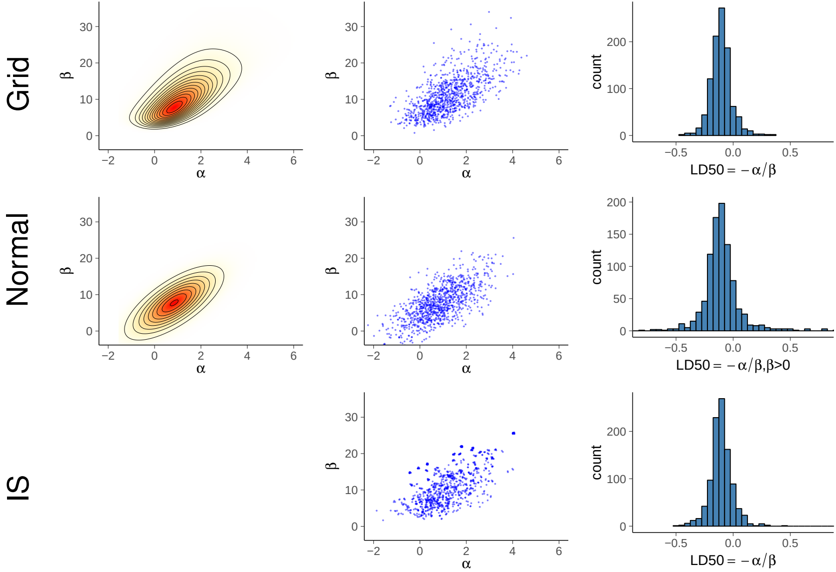


Example: Importance sampling in Bioassay

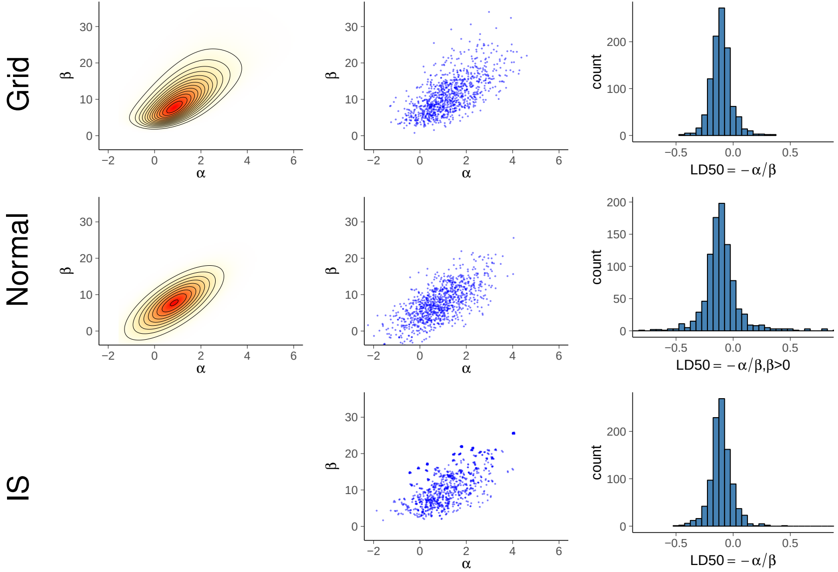


But the normal approximation is not that good here:
Grid sd(LD50) $\approx$ 0.1, Normal sd(LD50) $\approx$ .75!

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. $\theta \in [0, 1]$ presentin probability

Normal approximation and parameter transformations

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

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

Normal approximation and parameter transformations

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

Normal approximation and parameter transformations

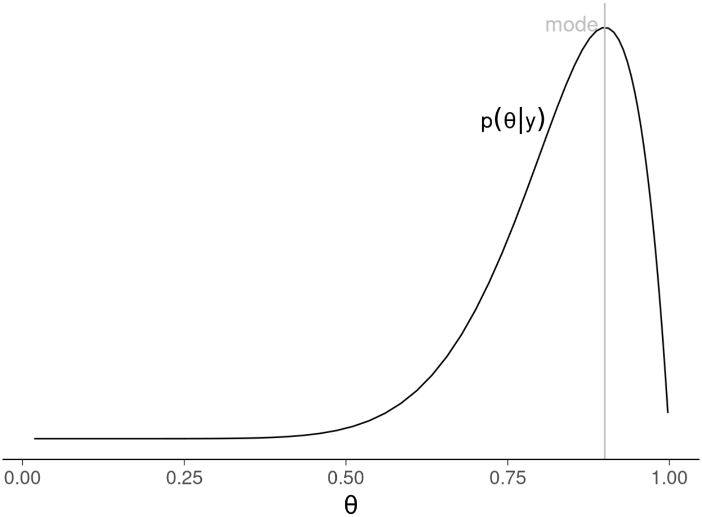
- Normal approximation is not good for parameters with bounded or half-bounded support
    - e.g. $\theta \in [0, 1]$ presentin probability
    - Stan code can include constraints

      ```
      real<lower=,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 $y = 9, N = 10$

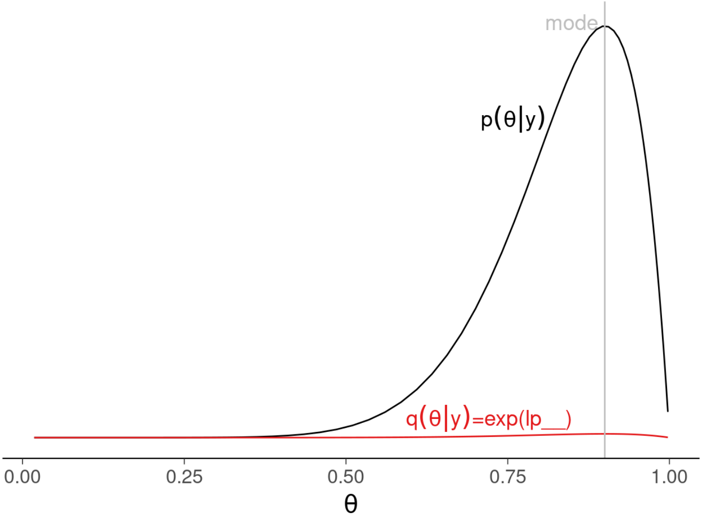
With $\mathrm{Beta}(1, 1)$ prior, the posterior is $\mathrm{Beta}(9 + 1, 1 + 1)$



Normal approximation and parameter transformations

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

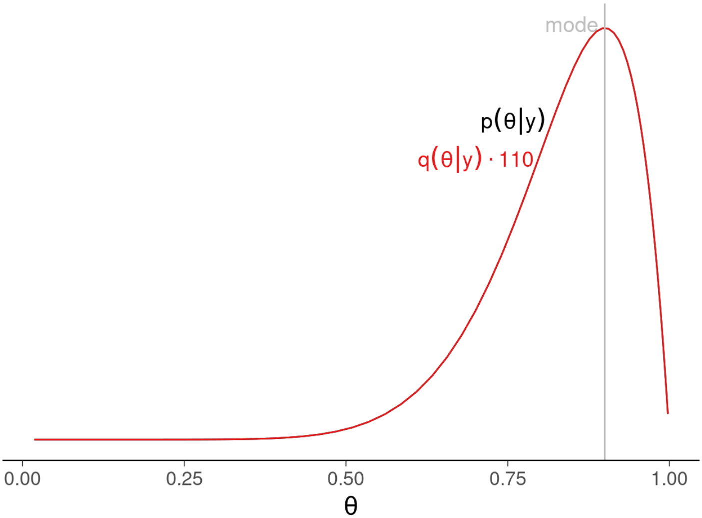
Stan computes only the unnormalized posterior $q(\theta|y)$



Normal approximation and parameter transformations

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

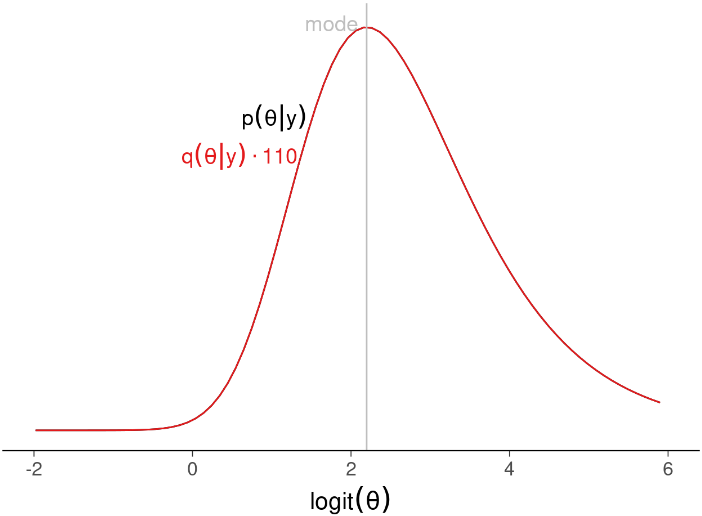
For illustration purposes we normalize Stan result $q(\theta|y)$



Normal approximation and parameter transformations

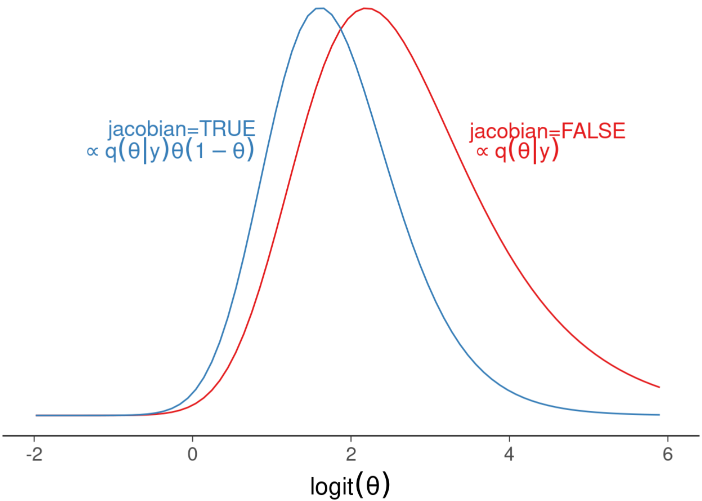
With $\mathrm{Beta}(1,1)$ prior, the posterior is $\mathrm{Beta}(9+1,1+1)$

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



Normal approximation and parameter transformations

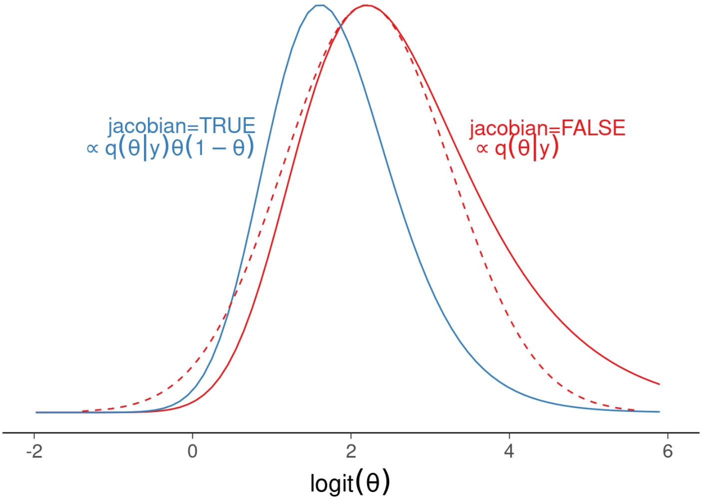
...but we need to take into account the absolute value of the determinant of the Jacobian of the transformation $\theta(1-\theta)$



Normal approximation and parameter transformations

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

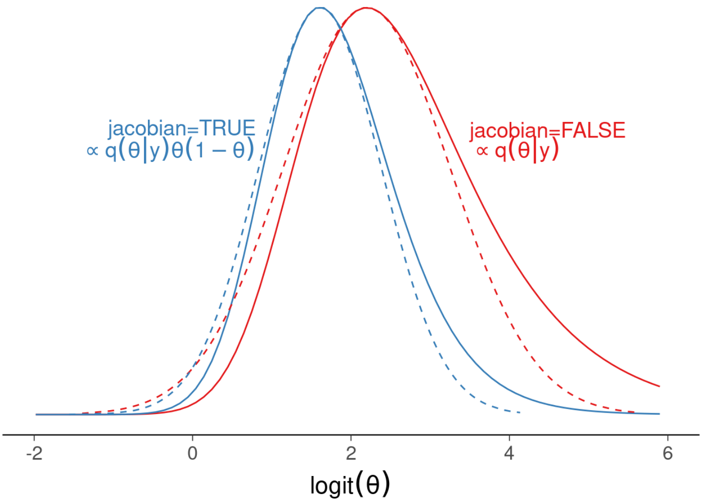
Let's compare a wrong normal approximation...



Normal approximation and parameter transformations

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

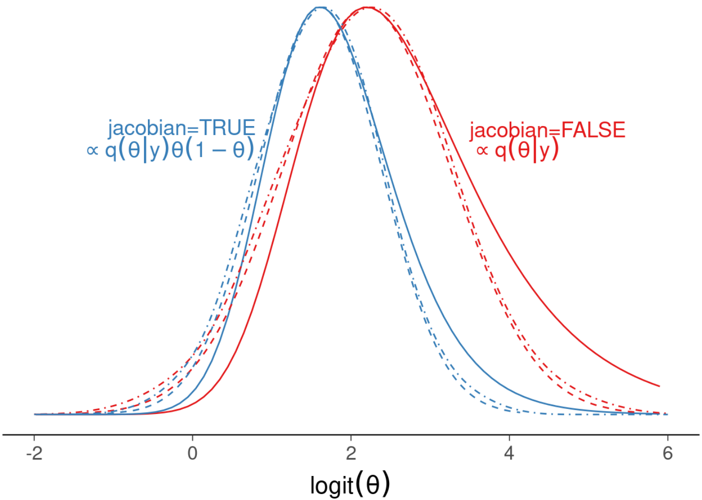
Let's compare a wrong normal approximation and correct one



Normal approximation and parameter transformations

Let's compare a wrong normal approximation and correct one

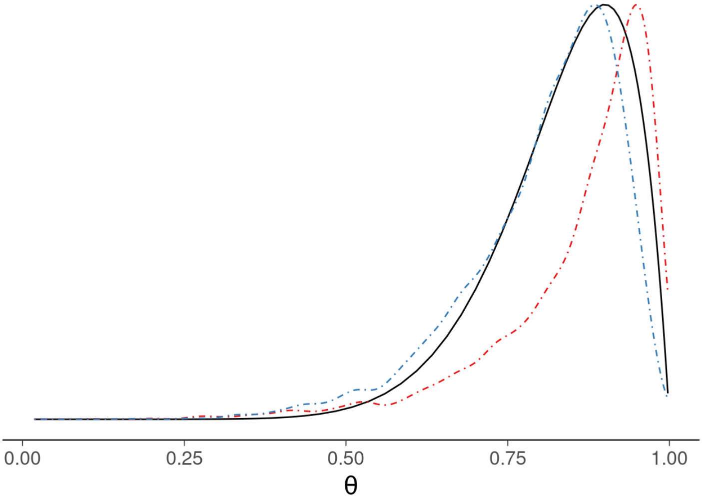
Sample from both approximations and show KDEs for draws



Normal approximation and parameter transformations

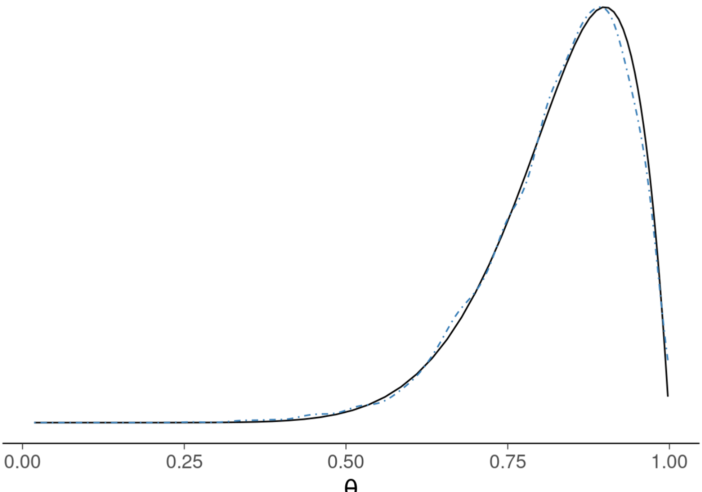
Let's compare a wrong normal approximation and correct one

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

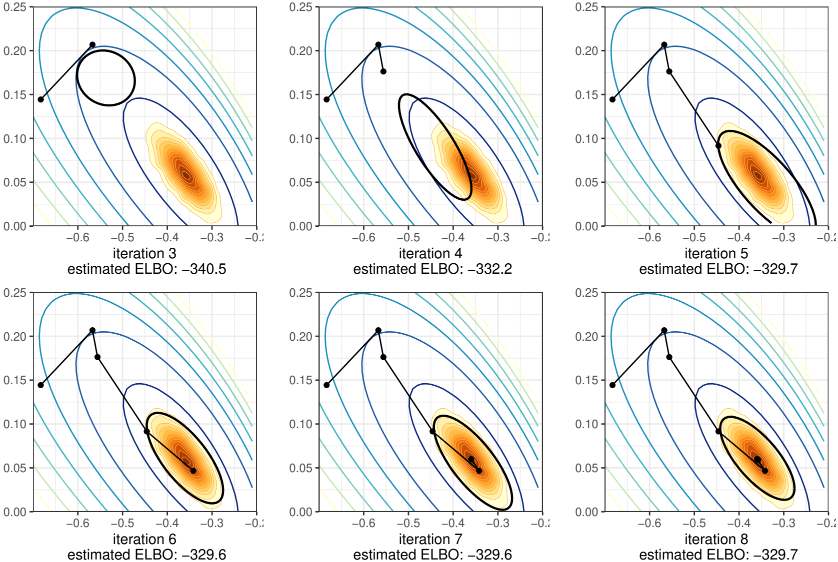
Other distributional approximations

- 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

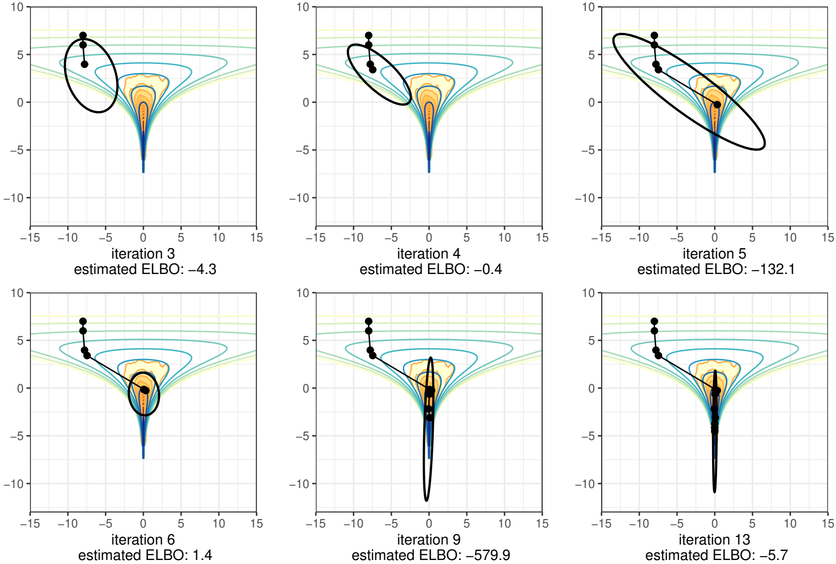
- Higher order derivatives at the mode can be used
- Split-normal and split-*t* by Geweke (1989) use additional scaling along different principal axes
- Other distributions can be used (e.g. *t*-distribution)
- Instead of mode and Hessian at mode, e.g.
    - variational inference (Ch 13)
        - CS-E4820 - Machine Learning: Advanced Probabilistic Methods
        - CS-E4895 - Gaussian Processes
        - Stan has the ADVI algorithm (not very good implementaion)
        - Stan has Pathfinder algorithm (CmdStanR github version)
        - 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 eeven 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.

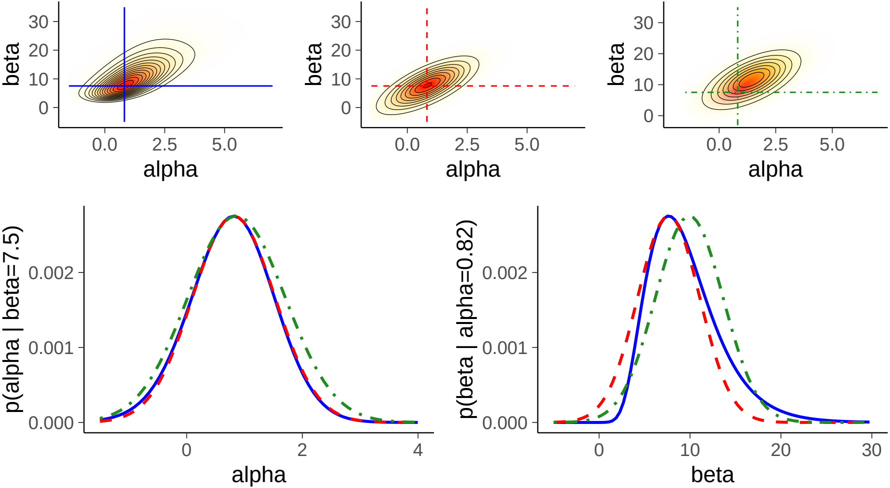
Pathfinder: Parallel quasi-Newton variational inference.



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

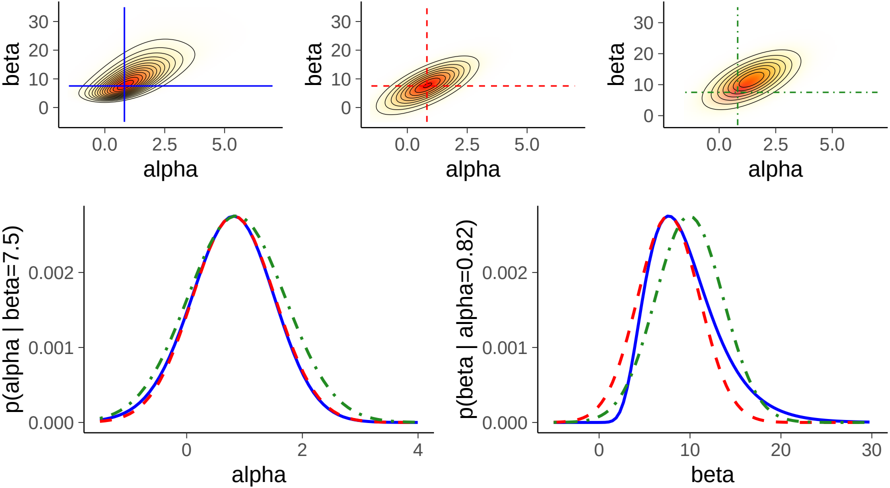
Distributional approximations

Exact, Normal at mode, Normal with variational inference



Distributional approximations

Exact, Normal at mode, Normal with variational inference

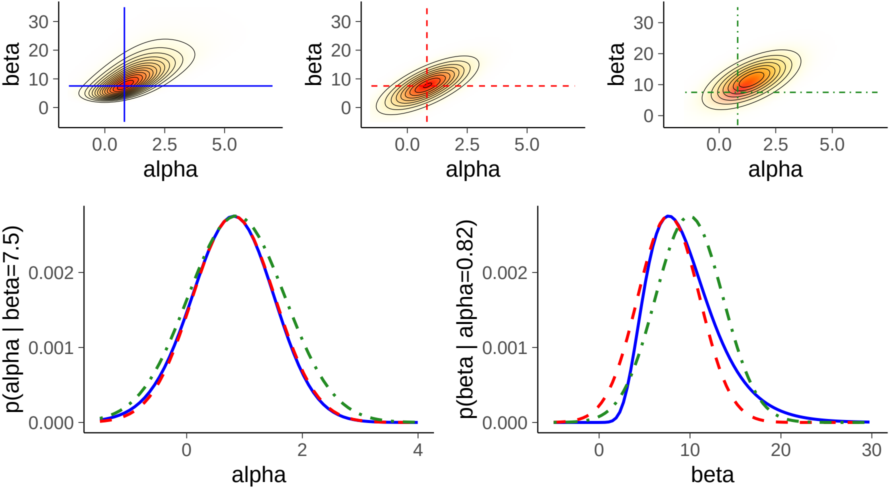


Grid sd(LD50) ≈ 0.090,

Normal sd(LD50) ≈ .75, Normal + IR sd(LD50) ≈ 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)

Large sample theory

- Asymptotic normality
  - as n the number of observations y_i increases the posterior converges to normal distribution

Large sample theory

- Asymptotic normality
  - as n the number of observations y_i increases the posterior converges to normal distribution
  - can be shown by showing that
    - eventually likelihood dominates the prior
    - the higher order terms in Taylor series increase slower than the second order term

Large sample theory

- Asymptotic normality
  - as n the number of observations y_i increases the posterior converges to normal distribution
  - can be shown by showing that
    - eventually likelihood dominates the prior
    - the higher order terms in Taylor series increase slower than the second order term
  - see counter examples

Large sample theory

- Assume "true" underlying data distribution $f(y)$
  - observations $y_1, \ldots, y_n$ are independent samples from the joint distribution $f(y)$
  - "true" data distribution $f(y)$ is not always well defined
  - in the following we proceed as if there were true underlying data distribution
  - for the theory the exact form of $f(y)$ is not important as long at it has certain regularity conditions

Large sample theory

- Consistency
  - if true distribution is included in the parametric family, so that $f(y) = p(y|\theta_0)$ for some θ_0, then posterior converges to a point θ_0, when $n \to \infty$

Large sample theory

- Consistency
    - if true distribution is included in the parametric family, so that $f(y) = p(y|\theta_0)$ for some θ_0, then posterior converges to a point θ_0, when $n \to \infty$
    - the same result as for maximum likelihood estimate

Large sample theory

- Consistency
  - if true distribution is included in the parametric family, so that $f(y) = p(y|\theta_0)$ for some θ_0, then posterior converges to a point θ_0, when $n \to \infty$
  - the same result as for maximum likelihood estimate
- If true distribution is not included in the parametric family, then there is no true θ_0
  - true θ_0 is replaced with θ_0 which minimizes the Kullback-Leibler divergence from $f(y)$ to $p(y_i|\theta_0)$

Large sample theory

- Consistency
  - if true distribution is included in the parametric family, so that $f(y) = p(y|\theta_0)$ for some θ_0, then posterior converges to a point θ_0, when $n \to \infty$
  - the same result as for maximum likelihood estimate
- If true distribution is not included in the parametric family, then there is no true θ_0
  - true θ_0 is replaced with θ_0 which minimizes the Kullback-Leibler divergence from $f(y)$ to $p(y_i|\theta_0)$
  - the same result as for maximum likelihood estimate

Large sample theory – counter examples

- Under- and non-identifiability
  - a model is under-identifiable, if the model has parameters or parameter combinations for which there is no information in the data
  - then there is no single point θ_0 where posterior would converge

Large sample theory – counter examples

- Under- and non-identifiability
  - a model is under-identifiable, if the model has parameters or parameter combinations for which there is no information in the data
  - then there is no single point θ_0 where posterior would converge
  - e.g. if the model is

$$y \sim \mathrm{N}(a + b + cx, \sigma)$$

Large sample theory – counter examples

- Under- and non-identifiability
  - a model is under-identifiable, if the model has parameters or parameter combinations for which there is no information in the data
  - then there is no single point θ_0 where posterior would converge
  - e.g. if the model is

$$y \sim \mathrm{N}(a + b + cx, \sigma)$$

    - posterior would converge to a line with prior determining the density along the line

Large sample theory – counter examples

- Under- and non-identifiability
  - a model is under-identifiable, if the model has parameters or parameter combinations for which there is no information in the data
  - then there is no single point θ_0 where posterior would converge
  - e.g. if the model is

$$y \sim \mathrm{N}(a + b + cx, \sigma)$$

    - posterior would converge to a line with prior determining the density along the line
  - e.g. if we never observe u and v at the same time and the model is

$$\begin{pmatrix} u \\ v \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

    then correlation ρ is non-identifiable

Large sample theory – counter examples

- Under- and non-identifiability
  - a model is under-identifiable, if the model has parameters or parameter combinations for which there is no information in the data
  - then there is no single point θ_0 where posterior would converge
  - e.g. if the model is

$$y \sim \mathrm{N}(a + b + cx, \sigma)$$

    - posterior would converge to a line with prior determining the density along the line
  - e.g. if we never observe u and v at the same time and the model is

$$\begin{pmatrix} u \\ v \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

    then correlation ρ is non-identifiable
    - e.g. u and v could be length and weight of a student; if only one of them is measured for each student, then ρ is non-identifiable

Large sample theory – counter examples

- Under- and non-identifiability
  - a model is under-identifiable, if the model has parameters or parameter combinations for which there is no information in the data
  - then there is no single point θ_0 where posterior would converge
  - e.g. if the model is

$$y \sim \mathrm{N}(a + b + cx, \sigma)$$

    - posterior would converge to a line with prior determining the density along the line
  - e.g. if we never observe u and v at the same time and the model is

$$\begin{pmatrix} u \\ v \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

    then correlation ρ is non-identifiable
    - e.g. u and v could be length and weight of a student; if only one of them is measured for each student, then ρ is non-identifiable
- Problem also for other inference methods like MCMC

Asymptotic identifiability vs finite data case

- If we randomly would measure both height and weight, asymptotically the correlation ρ would be identifiable

Asymptotic identifiability vs finite data case

- If we randomly would measure both height and weight, asymptotically the correlation ρ would be identifiable
- But a finite data from this data generating process may lack the joint height and weight observations, and thus the the finite data likelihood doesn't have information about ρ

Asymptotic identifiability vs finite data case

- If we randomly would measure both height and weight, asymptotically the correlation ρ would be identifiable
- But a finite data from this data generating process may lack the joint height and weight observations, and thus the the finite data likelihood doesn't have information about ρ
- If the likelihood is weakly informative for some parameters, priors and integration are more important

Large sample theory – counter examples

- If the number of parameter increases as the number of observation increases
  - in some models number of parameters depends on the number of observations
  - e.g. time series models $y_t \sim \mathrm{N}(\theta_t, \sigma^2)$ and θ_t has prior in time
  - posterior of θ_t does not converge to a point, if additional observations do not bring enough information

Large sample theory – counter examples

- Aliasing (valetoisto in Finnish)
  - special case of under-identifiability where likelihood repeats in separate points
  - e.g. mixture of normals

$$p(y_i|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \lambda) = \lambda\,\mathrm{N}(\mu_1, \sigma_1^2) + (1-\lambda)\,\mathrm{N}(\mu_2, \sigma_2^2)$$

Large sample theory – counter examples

- Aliasing (valetoisto in Finnish)
  - special case of under-identifiability where likelihood repeats in separate points
  - e.g. mixture of normals

$$p(y_i|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \lambda) = \lambda \, \mathrm{N}(\mu_1, \sigma_1^2) + (1 - \lambda) \, \mathrm{N}(\mu_2, \sigma_2^2)$$

  if (μ_1, μ_2) are switched, (σ_1^2, σ_2^2) are switched and replace λ with $(1 - \lambda)$, model is equivalent; posterior would usually have two modes which are mirror images of each other and the posterior does not converge to a single point

Large sample theory – counter examples

- Aliasing (valetoisto in Finnish)
  - special case of under-identifiability where likelihood repeats in separate points
  - e.g. mixture of normals

$$p(y_i|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \lambda) = \lambda\,\mathrm{N}(\mu_1, \sigma_1^2) + (1-\lambda)\,\mathrm{N}(\mu_2, \sigma_2^2)$$

  if (μ_1, μ_2) are switched, (σ_1^2, σ_2^2) are switched and replace λ with $(1-\lambda)$, model is equivalent; posterior would usually have two modes which are mirror images of each other and the posterior does not converge to a single point
- For MCMC makes the convergence diagnostics more difficult, as it is difficult to identify aliasing from other multimodality

Large sample theory – counter examples

- Unbounded (rajoittamaton in Finnish) likelihood
  - if likelihood is unbounded it is possible that there is no mode in the posterior

Large sample theory – counter examples

- Unbounded (rajoittamaton in Finnish) likelihood
  - if likelihood is unbounded it is possible that there is no mode in the posterior
  - e.g. previous normal mixture model; assume λ to be known (and not 0 or 1); if we set $\mu_1 = y_i$ for any i and $\sigma_1^2 \to 0$, then likelihood $\to \infty$

Large sample theory – counter examples

- Unbounded (rajoittamaton in Finnish) likelihood
  - if likelihood is unbounded it is possible that there is no mode in the posterior
  - e.g. previous normal mixture model; assume λ to be known (and not 0 or 1); if we set $\mu_1 = y_i$ for any i and $\sigma_1^2 \to 0$, then likelihood $\to \infty$
  - if prior for σ_1^2 does not go to zero when $\sigma_1^2 \to 0$, then the posterior is unbounded

Large sample theory – counter examples

- Unbounded (rajoittamaton in Finnish) likelihood
  - if likelihood is unbounded it is possible that there is no mode in the posterior
  - e.g. previous normal mixture model; assume λ to be known (and not 0 or 1); if we set $\mu_1 = y_i$ for any i and $\sigma_1^2 \to 0$, then likelihood $\to \infty$
  - if prior for σ_1^2 does not go to zero when $\sigma_1^2 \to 0$, then the posterior is unbounded
  - when $n \to \infty$ the number of likelihood modes increases

Large sample theory – counter examples

- Unbounded (rajoittamaton in Finnish) likelihood
  - if likelihood is unbounded it is possible that there is no mode in the posterior
  - e.g. previous normal mixture model; assume λ to be known (and not 0 or 1); if we set $\mu_1 = y_i$ for any i and $\sigma_1^2 \to 0$, then likelihood $\to \infty$
  - if prior for σ_1^2 does not go to zero when $\sigma_1^2 \to 0$, then the posterior is unbounded
  - when $n \to \infty$ the number of likelihood modes increases
- Problem for any inference method including MCMC
  - can be avoided with good priors

Large sample theory – counter examples

- Unbounded (rajoittamaton in Finnish) likelihood
  - if likelihood is unbounded it is possible that there is no mode in the posterior
  - e.g. previous normal mixture model; assume λ to be known (and not 0 or 1); if we set $\mu_1 = y_i$ for any i and $\sigma_1^2 \to 0$, then likelihood $\to \infty$
  - if prior for σ_1^2 does not go to zero when $\sigma_1^2 \to 0$, then the posterior is unbounded
  - when $n \to \infty$ the number of likelihood modes increases
- Problem for any inference method including MCMC
  - can be avoided with good priors
  - a prior close to a prior allowing unbounded posterior may produce almost unbounded posterior

Large sample theory – counter examples

- Improper posterior
  - asymptotic results assume that probability sums to 1
  - e.g. Binomial model, with $\mathrm{Beta}(0, 0)$ prior and observation $y = n$
    - posterior $p(\theta|n, 0) = \theta^{n-1}(1-\theta)^{-1}$
    - when $\theta \to 1$, then $p(\theta|n, 0) \to \infty$

Large sample theory – counter examples

- Improper posterior
  - asymptotic results assume that probability sums to 1
  - e.g. Binomial model, with $\mathtt{Beta}(0, 0)$ prior and observation $y = n$
    - posterior $p(\theta|n, 0) = \theta^{n-1}(1-\theta)^{-1}$
    - when $\theta \to 1$, then $p(\theta|n, 0) \to \infty$
- Problem for any inference method including MCMC
  - can be avoided with proper priors

Large sample theory – counter examples

- Improper posterior
  - asymptotic results assume that probability sums to 1
  - e.g. Binomial model, with $\mathtt{Beta}(0, 0)$ prior and observation $y = n$
    - posterior $p(\theta|n, 0) = \theta^{n-1}(1-\theta)^{-1}$
    - when $\theta \to 1$, then $p(\theta|n, 0) \to \infty$
- Problem for any inference method including MCMC
  - can be avoided with proper priors
  - a prior close to a improper prior may produce almost improper posterior

Large sample theory – counter examples

- Prior distribution does not include the convergence point
  - if in discrete case $p(\theta_0) = 0$ or in continuous case $p(\theta) = 0$ in the neighborhood of θ_0, then the convergence results based on the dominance of the likelihood do not hold

Large sample theory – counter examples

- Prior distribution does not include the convergence point
  - if in discrete case $p(\theta_0) = 0$ or in continuous case $p(\theta) = 0$ in the neighborhood of θ_0, then the convergence results based on the dominance of the likelihood do not hold
- Should have a positive prior probability/density where needed

Large sample theory – counter examples

- Convergence point at the edge of the parameter space
  - if θ_0 is on the edge of the parameter space, Taylor series expansion has to be truncated, and normal approximation does not necessarily hold

Large sample theory – counter examples

- Convergence point at the edge of the parameter space
  - if θ_0 is on the edge of the parameter space, Taylor series expansion has to be truncated, and normal approximation does not necessarily hold
  - e.g. $y_i \sim \mathrm{N}(\theta, 1)$ with a restriction $\theta \geq 0$ and assume that $\theta_0 = 0$
    - posterior of θ is left truncated normal distribution with $\mu = \bar{y}$
    - in the limit $n \to \infty$ posterior is half normal distribution
- Can be easy or difficult for MCMC

Frequency evaluations

- Bayesian theory has epistemic and aleatory probabilities
- Frequency evaluations focus on frequency properties given aleatoric repetition of an observation and modeling
    - It is useful to examine these for Bayesian inference, too

Frequency evaluations

- Bayesian theory has epistemic and aleatory probabilities
- Frequency evaluations focus on frequency properties given aleatoric repetition of an observation and modeling
    - It is useful to examine these for Bayesian inference, too
    - Asymptotic unbiasedness
        - not that important in Bayesian inference, small and decreasing error more important

Frequency evaluations

- Bayesian theory has epistemic and aleatory probabilities
- Frequency evaluations focus on frequency properties given aleatoric repetition of an observation and modeling
  - It is useful to examine these for Bayesian inference, too
  - Asymptotic unbiasedness
    - not that important in Bayesian inference, small and decreasing error more important
  - Asymptotic efficiency
    - no other point estimate with smaller squared error
    - useful also in Bayesian inference, but should consider which utility/loss is important

Frequency evaluations

- Bayesian theory has epistemic and aleatory probabilities
- Frequency evaluations focus on frequency properties given aleatoric repetition of an observation and modeling
  - It is useful to examine these for Bayesian inference, too
  - Asymptotic unbiasedness
    - not that important in Bayesian inference, small and decreasing error more important
  - Asymptotic efficiency
    - no other point estimate with smaller squared error
    - useful also in Bayesian inference, but should consider which utility/loss is important
  - Calibration
    - α%-posterior interval has the true value in α% cases
    - α%-predictive interval has the true future values in α% cases
    - approximate calibration with shorter intervals for likely true values more important than exact calibration with very bad intervals for all possible values.

Frequentist statistics

- Frequentist statistics accepts only aleatory probabilities
  - Estimates are based on data
  - Uncertainty of estimates are based on all possible data sets which could have been generated by the data generating mechanism

Frequentist statistics

- Frequentist statistics accepts only aleatory probabilities
    - Estimates are based on data
    - Uncertainty of estimates are based on all possible data sets which could have been generated by the data generating mechanism
        - inference is based also on data we did not observe

Frequentist statistics

- Frequentist statistics accepts only aleatory probabilities
  - Estimates are based on data
  - Uncertainty of estimates are based on all possible data sets which could have been generated by the data generating mechanism
    - inference is based also on data we did not observe
- Estimates are derived to fulfill frequency properties
  - Maximum likelihood (often) fulfills asymptotic frequency properties
  - Common finite data desiderata are 1) unbiasedness, 2) minimum variance, 3) calibration of confidence interval

Frequentist statistics

- Estimates are derived to fulfill frequency properties
  - Maximum likelihood fulfills just asymptotic frequency properties
  - Common desiderata are 1) unbiasedness, 2) minimum variance, 3) calibration of confidence interval
- Requirement of unbiasedness may lead to higher variance or silly estimates
  - unbiased estimate for strictly positive parameter can be negative

Frequentist statistics

- Estimates are derived to fulfill frequency properties
  - Maximum likelihood fulfills just asymptotic frequency properties
  - Common desiderata are 1) unbiasedness, 2) minimum variance, 3) calibration of confidence interval
- Requirement of unbiasedness may lead to higher variance or silly estimates
  - unbiased estimate for strictly positive parameter can be negative
- Confidence interval is defined to have true value inside the interval in $\alpha\%$ cases of repeated data generation from the data generating mechanism
  - doesn't need be useful to have perfect calibration

Frequentist vs Bayes vs others

- There is a great amount of very useful frequentist statistics
  - also for simple models and lot's of data there is not much difference

Frequentist vs Bayes vs others

- There is a great amount of very useful frequentist statistics
  - also for simple models and lot's of data there is not much difference
- Bayesian inference
  - easier for complex, e.g. hierarchical, models
  - easier when model changes
  - a consistent way to add prior information

Frequentist vs Bayes vs others

- There is a great amount of very useful frequentist statistics
  - also for simple models and lot's of data there is not much difference
- Bayesian inference
  - easier for complex, e.g. hierarchical, models
  - easier when model changes
  - a consistent way to add prior information
- A lot of machine learning is not pure frequentist or Bayesian