Chapter 3.1

Deterministic methods

Bayesian computing

- Give the prior and data, the posterior is fixed and a Bayesian analysis boils down to summarizing the posterior
- We need point estimates, credible sets, etc
- Summarizing a p-dimensional posterior distribution is challenging for large p
- In the 80's, Bayesian computing was unable to do this for more than a few parameters
- In the 90's, new algorithms were developed that revolutionized Bayesian statistics
- Understanding these algorithms is obviously important

Approaches to Bayesian computing

Some approaches to dealing with complicated joint posteriors:

- Just use a point estimate (e.g., MAP), ignore uncertainty
- Approximate the posterior as Gaussian
- Numerical integration
- Markov Chain Monte Carlo (MCMC) sampling

Outline of Chapter 3

- Deterministic methods
  - MAP estimation
  - Numerical integration
  - Bayesian Central Limit Theorem
- MCMC algorithms
  - Gibbs sampling
  - Metropolis-Hastings sampling
- Just Another Gibbs Sampler (JAGS)
- Diagnostic and improving convergence
  - Setting initial values
  - Convergence diagnostics
  - Improving convergence
  - Dealing with large datasets

MAP estimation

- Sometimes you don't need an entire posterior distribution and a single point estimate will do
- Example: prediction in machine learning
- The Maximum a Posteriori (MAP) estimate is the posterior mode

$$\hat{\boldsymbol{\theta}}_{MAP} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ p(\boldsymbol{\theta}|\mathbf{Y}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ \log[f(\mathbf{Y}|\boldsymbol{\theta})] + \log[\pi(\boldsymbol{\theta})]$$

- This is similar to the maximum likelihood estimation but includes the prior

Univariate example

Say $Y|\theta \sim Binomial(n,\theta)$ *and* $\theta \sim Beta(0.5, 0.5)$, *find* $\hat{\theta}_{MAP}$

- The likelihood is $f(Y|\theta) \propto \theta^{Y}(1-\theta)^{n-Y}$
- The log likelihood is[1]

$$\log[f(Y|\theta)] = Y\log(\theta) + (n - Y)\log(1-\theta)$$

- The prior is $\pi(\theta) \propto \theta^{0.5-1}(\theta)^{0.5-1}$
- The log prior[1] is $\log[\pi(\theta)] = -0.5\log(\theta) - 0.5\log(1-\theta)$
- Therefore, the MAP estimator is

$$\hat{\theta} = \arg\max_{\theta}(Y - 0.5)\log(\theta) + (n - Y - 0.5)\log(1-\theta)$$

[1] ignoring constants that don't depend on θ

Univariate example

Say $Y|\theta \sim \textit{Binomial}(n, \theta)$ *and* $\theta \sim \textit{Beta}(0.5, 0.5)$*, find* $\hat{\theta}_{MAP}$

- The MAP estimator is

$$\hat{\theta} = \arg\max_{\theta} (Y - 0.5)\log(\theta) + (n - Y - 0.5)\log(1 - \theta)$$

- Taking the derivative and setting to zero gives

$$\frac{Y - 0.5}{\theta} - \frac{n - Y - 0.5}{1 - \theta} = 0$$

- The solution (assuming $Y, n - Y \geq 1$) is

$$\hat{\theta} = \frac{Y - 0.5}{n - 1}$$

Bayesian central limit theorem

- Another simplification is to approximate the posterior as Gaussian
- Berstein-Von Mises Theorem: As the sample size grows the posterior doesn't depend on the prior
- Frequentist result: As the sample size grows the likelihood function is approximately normal
- Bayesian CLT: For large n and some other conditions $\boldsymbol{\theta}|\mathbf{Y} \approx$ Normal

Bayesian central limit theorem

- Bayesian CLT: For large n and some other conditions

$$\boldsymbol{\theta} \sim \text{Normal}[\hat{\boldsymbol{\theta}}_{MAP}, \mathbf{I}(\hat{\boldsymbol{\theta}}_{MAP})^{-1}]$$

- $\mathbf{I}$ is Fisher's information matrix

- The (j, k) element of $\mathbf{I}$ is

$$-\frac{\partial^2}{\partial\theta_j \partial\theta_k} \log[p(\boldsymbol{\theta}|\mathbf{Y})]$$

  evaluated at $\hat{\boldsymbol{\theta}}_{MAP}$

- We have marginal and conditional means, standard deviations and intervals for the normal distribution

Univariate example

Say $Y|\theta \sim$ *Binomial*(n, θ) *and* $\theta \sim$ *Beta*$(0.5, 0.5)$*, find the Gaussian approximation for* $p(\theta|\mathbf{Y})$

- We have seen that (assuming $Y, n - Y \geq 1$),

$$\hat{\theta}_{MAP} = \frac{Y - 0.5}{n - 1}$$

- We have also seen (Jeffreys lecture) that

$$I(\theta) = n\theta^{-1}(1 - \theta)^{-1}$$

- Therefore,

$$\begin{aligned} \theta|Y &\approx \text{Normal}\left[\hat{\theta}_{MAP}, I(\hat{\theta}_{MAP})^{-1}\right] \\ &\approx \text{Normal}\left[\hat{\theta}_{MAP}, \hat{\theta}_{MAP}(1 - \hat{\theta}_{MAP})/n\right] \end{aligned}$$

Illustration of the Bayesian CLT

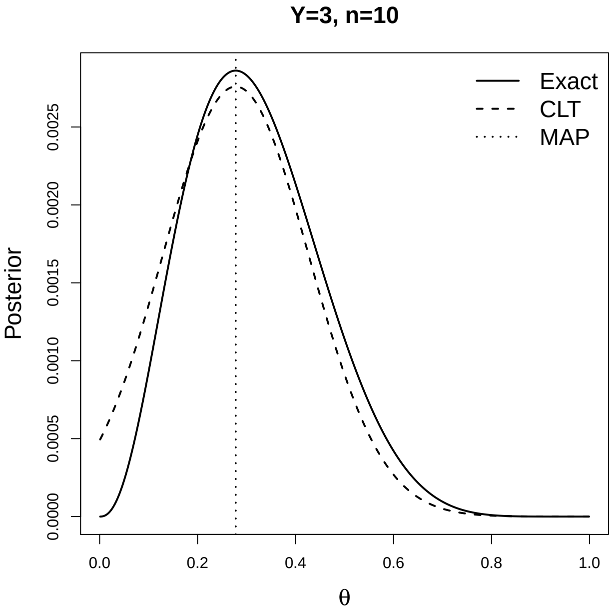


Illustration of the Bayesian CLT

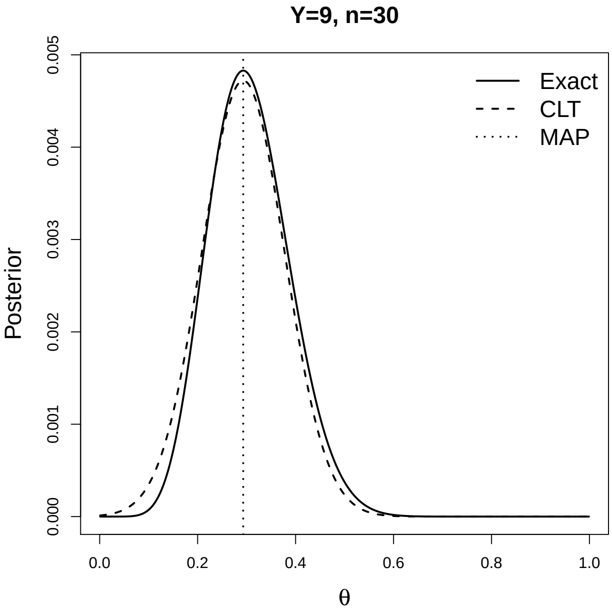
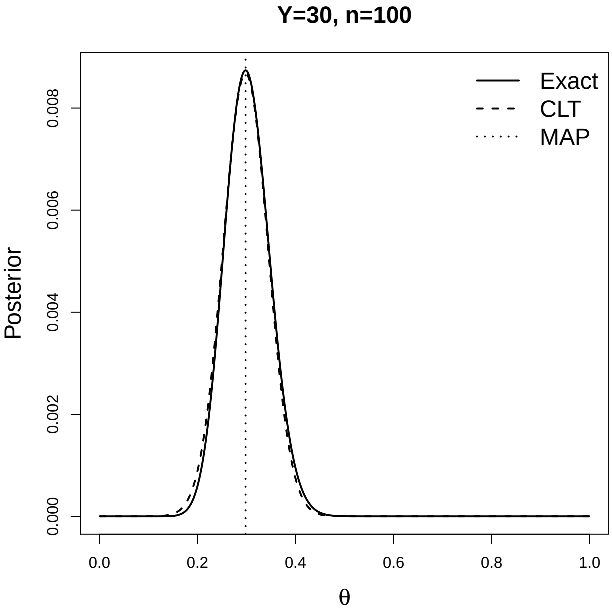


Illustration of the Bayesian CLT



Bayesian central limit theorem

- For large datasets with a small number of parameters evoking the Bayes CLT is probably the best approach
- The approximate posterior can be computing using standard software (e.g., `glm` in `R`)
- The numerical values (e.g., intervals) will equal the frequentist values, but the interpretation remains Bayesian
- Why not just do a frequentist analysis? Well, why not just do a Bayesian analysis?

Numerical integration

- Many posterior summaries of interest are integrals over the posterior
- Ex: $\text{E}(\theta_j|\mathbf{Y}) = \int \theta_j p(\boldsymbol{\theta}) d\boldsymbol{\theta}$
- Ex: $\text{V}(\theta_j|\mathbf{Y}) = \int [\theta_j - \text{E}(\theta|\mathbf{Y})]^2 p(\boldsymbol{\theta}) d\boldsymbol{\theta}$
- These are *p* dimensional integrals that we usually can't solve analytically
- A grid approximation is a crude approach
- Gaussian quadrature is better

Numerical integration

- Numerical integration is only feasible for small p
- The Iteratively Nested Laplace Approximation (INLA) is an even more sophisticated method
- INLA combines Gaussian approximations with numerical integration
- This works well if most of the parameters are approximately normal and only a few are non-Gaussian and require numerical integration