Deviance Information Criterion

PUBH 8442: Bayes Decision Theory and Data Analysis

Eric F. Lock
UMN Division of Biostatistics, SPH
elock@umn.edu

04/21/2021

BIC and AIC

- Define the *deviance* function for a model with parameters θ:

$$D(\theta) = -2 \log p(\mathbf{y} \mid \theta)$$

- Recall: Bayesian information criterion

$$BIC : D(\hat{\theta}) + p \log n$$

  - $\hat{\theta}$ is the maximum likelihood estimate
  - p is model dimension, $\theta = (\theta_1, \ldots, \theta_p)$,
  - n is sample size, $\mathbf{y} = y_1, \ldots, y_n$
  - Motivated by asymptotic approximation of Bayes factor

- Akaike information criterion

$$AIC = D(\hat{\theta}) + 2p$$

  - Motivated by asymptotic approximation to Kullback-Leibler divergence

BIC and AIC

- What if choice of p and n is not clear?
- This is common in Bayesian hierarchical models.
- Example: Consider the multi-level normal model

$$y_{ij} \sim \text{Normal}(\theta_i, \sigma^2)\} \text{ for } i = 1, \ldots, m \text{ and } j = 1, \ldots, n_i$$

$$\theta_i \sim \text{Normal}(\mu, \tau^2)\}$$

  - If θ_i are all nearly identical ($\tau^2 \to 0$), model depends only on estimation of μ ($p \approx 1$)
  - If θ_i are estimated independently ($\tau^2 \to \infty$), $p \approx m$ makes sense.
  - The choice of "sample size" is similarly unclear

Effective number of parameters

- Define the effective number of parameters by

$$p_D = E_{\theta \mid \mathbf{y}} D(\theta) - D(\hat{\theta})$$

  where typically $\hat{\theta} = E_{\theta \mid \mathbf{y}} \theta$.

  - The "expected" deviance minus the "fitted" deviance

- Higher p_D implies more over-fitting with estimate $\hat{\theta}$
- For a non-hierarchical model, the Bayesian CLT implies $p \approx p_D$ for large n

Define $L(\theta) = \log p(y|\theta)$

and $D(\theta) = -2L(\theta)$

By the Bayesian CLT,

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

where $\hat{\theta}$ is posterior mean or MLE

$L'(\hat{\theta}) = 0$

2nd order Taylor approx about $\hat{\theta}$:

$$D(\theta) \approx D(\hat{\theta}) + 0 - \underbrace{(\theta - \hat{\theta})^T L''(\hat{\theta})(\theta - \hat{\theta})}_{\sim \chi^2_p}$$

$$\rightarrow p \approx E_{\theta|y} D(\theta) - D(\hat{\theta})$$

Deviance information criteria

- The *Deviance information criteria* (DIC) is

$$DIC = E_{\theta \mid \mathbf{y}} D(\theta) + p_D = D(\hat{\theta}) + 2p_D$$

$$= 2E_{\theta|y} D(\theta) - D(\hat{\theta})$$

- Approximates AIC for a non-hierarchical model
- Similar asymptotic justification as AIC
- Used for model comparison
  - Lower DIC values are better
- Can estimate DIC from posterior samples:

$$DIC = 2\bar{D} - D(\bar{\theta})$$

where $\bar{\theta} = \frac{1}{N}\sum_{t=1}^{N} \theta^{(t)}$,

$$\bar{D} = \frac{1}{N}\sum_{t=1}^{N} -2\log p(\mathbf{y} \mid \theta^{(t)})$$

DIC comments

- DIC values are not very informative on their own
  - Used for comparisons
- Includes a "goodness-of-fit" term $\overset{D(\hat{\theta})}{\cancel{\bar{D}}}$ with a penalty for "complexity" (p_D)
  - Like BIC, AIC, and other model selection criteria
- More appropriate for hierarchical models than AIC, BIC
- p_D can be negative if $D(\bar{\theta})$ is relatively large.
  - Implies Bayesian CLT does not hold and $\bar{\theta}$ is a poor estimate
- Compute in winBUGS and openBUGS: http://www.openbugs.net/Manuals/InferenceMenu.html

Example: gene testing

- 40 mice are given a given a dose of alcohol, 40 are kept as control
- Expression levels are subsequently measured for 500 genes in liver
- Y_{ij}^{g} is expression level for gene i, mouse j, group g
- Measurements are normally distributed with variance 1:

$$Y_{ij}^{g} \sim \text{Normal}(\mu_i^{g}, 1)$$

- Consider the group differences

$$Y_i^{\text{diff}} = \underbrace{\bar{Y}_i^{\text{alc}}}_{N(\mu_i^{alc}, \frac{1}{40})} - \underbrace{\bar{Y}_i^{\text{con}}}_{N(\mu_i^{con}, \frac{1}{40})} \sim \text{Normal}\left(\mu_i^{\text{alc}} - \mu_i^{\text{con}}, \frac{1}{20}\right)$$

Example: gene testing

- We are interested in effect of alcohol on each gene i:

$$\mu_i^{\text{diff}} = \mu_i^{\text{alc}} - \mu_i^{\text{con}}$$

- Use normal prior for effects:

$$\mu_i^{\text{diff}} \overset{iid}{\sim} \text{Normal}(0, \tau^2)$$

- Jeffrey's prior for effect variance:

$$p(\tau^2) \propto \frac{1}{\tau^2}$$

- Full distribution for y_i^{diff} s:

$$\frac{1}{\tau^2} \prod_{i=1}^{500} N(\mu_i^{\text{diff}} \mid 0, \tau^2) \underline{N(y_i^{\text{diff}} \mid \mu_i^{\text{diff}}, 1/20)}$$

Example: gene testing

- Gibbs sample conditionals for $\mu_i^{\text{diff}}s$ and τ^2:

$$p(\mu_i^{\text{diff}} \mid \tau^2, \mathbf{y}) = \text{Normal}\left(\frac{\tau^2 y_i^{\text{diff}}}{\tau^2 + 1/20}, \frac{(1/20)\tau^2}{\tau^2 + 1/20}\right)$$

$$p(\tau^2 \mid \mu^{\text{diff}}, \mathbf{y}) = IG\left(250, \frac{1}{2}\sum_{i=1}^{500} \mu_i^2\right)$$

- Initialize $\tau^2 = 1/20$, run 10000 iterations with 2000 burn-in
- Compute

$$D(\mu^{\text{diff}}, \tau^2) = -2\sum_{i=1}^{500} \log[N(y_i^{\text{diff}} \mid \mu_i^{\text{diff}}, 1/20)]$$

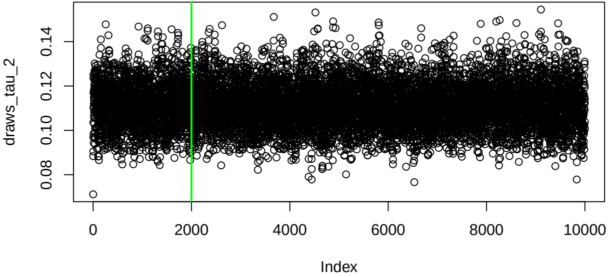
at each iteration.

R CODE

```
T=10000
BurnIn = 2000
N=T-BurnIn
draws_tau_2 = rep(0,T)
draws_mu_diff = matrix(nrow = T, ncol = 500)
Ds = rep(0,T)
tau_2 = 1/20 ### initialize
for(t in 1:T){ ##Run gibbs sampler
  mus = rnorm(500, tau_2*y_diffs/(tau_2+0.05),
      sqrt(0.05*tau_2/tau_2+0.05)))
 tau_2 =1/rgamma(1,250, 0.5*sum(mus^2))
  draws_tau_2[t] = tau_2
  draws_mu_diff[t,] = mus
  Ds[t] = -2*sum(log(dnorm(y_diffs,mus,sqrt(0.05))))
}
```

Example: gene testing

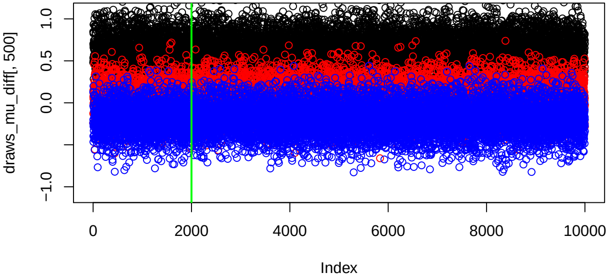
- Gibbs draws for τ^2:



http://www.ericfrazerlock.com/Deviance_Information_Criteria_Rcode1.R

Example: gene testing

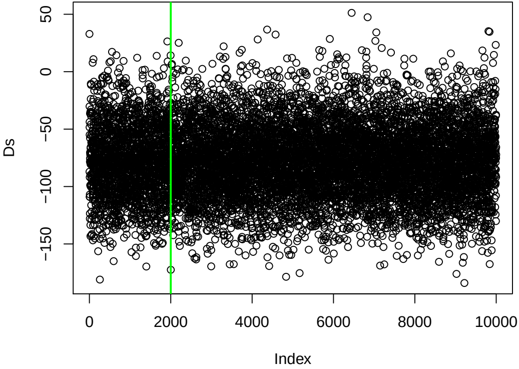
- Gibbs draws for μ_{diff}, three genes:



http://www.ericfrazerlock.com/Deviance_Information_Criteria_Rcode1.R

Example: gene testing

- Plot of deviance over Gibbs draws



R CODE

```
###compute DIC
mean_mus = colMeans(draws_mu_diff[2001:T,])
D_mean =  -2*sum(log(dnorm(y_diffs,mean_mus,sqrt(0.05))))
p_d = mean(Ds[2001:T])-D_mean
DIC = 2*mean(Ds[2001:T])-D_mean
DIC_null = -2*sum(log(dnorm(y_diffs,0,sqrt(0.05))))
```

Example: gene testing

- The deviance for $\hat{\mu}^{\text{diff}}$, the mean vector over draws, is

$$D(\hat{\mu}^{\text{diff}}) = -422.7$$

- Thus $p_D = \bar{D} - D(\hat{\mu}^{\text{diff}}) = 344.9$
- DIC is $DIC = \bar{D} + p_D = 267.1$
- Consider the null model $\mu_i^{\text{diff}} = 0 \ \forall i$
  - The effective number of parameters is $p_D = 0$
  - DIC is

$$DIC = -2 \sum_{i=1}^{500} \log[N(y_i^{\text{diff}} \mid 0, 1/20)] = 1029$$

- Evidence there are alcohol effects (for at least some genes)

Example: gene testing

- Consider a third model, that allows "no effect" for some genes.
- P_1 is shared probability that $\mu_i^{\text{diff}} \neq 0$ for a given gene:

$$\mu_i^{\text{diff}} \sim \begin{cases} 0 \text{ with probability } 1 - P_1 \\ N(0, \tau^2) \text{ with probability } P_1 \end{cases}$$

- Again, $p(\tau^2) = 1/\tau^2$
- Use a uniform prior for P_1

$$P_1 \sim \text{Beta}(1, 1)$$

- Let $\zeta_i = \mathbb{1}\{\mu_i^{\text{diff}} \neq 0\}$

Gibbs sampling

- Draw from conditional for $(\zeta, \mu^{\text{diff}})$ for each gene i:

  $$y_i^{\text{diff}} = \underbrace{\mu_i^{\text{diff}}}_{N(0,\tau^2)} + \underbrace{e_i^{\text{diff}}}_{N(0,\frac{1}{20})}$$

  - Draw $\zeta_i \in \{0, 1\}$ by

    $$P(\zeta_i = 1 | \mathbf{y}, \tau^2, P_1) = \frac{P_1 N(y_i^{\text{diff}} \mid 0, \tau^2 + \frac{1}{20})}{P_1 N(y_i^{\text{diff}} \mid 0, \tau^2 + \frac{1}{20}) + (1 - P_1) N(y_i^{\text{diff}} \mid 0, \frac{1}{20})}$$

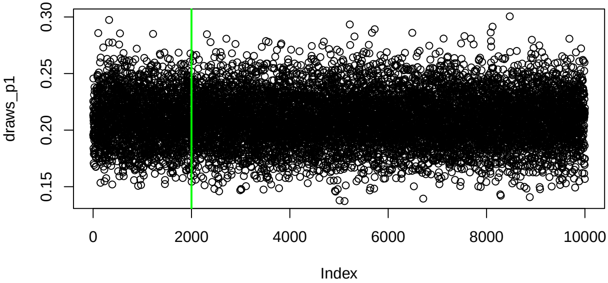
  - If $\zeta_i = 0$, set $\mu_i^{\text{diff}} = 0$
  - Otherwise, generate $\mu_i^{\text{diff}} \sim \text{Normal}\left(\frac{\tau^2 y_i^{\text{diff}}}{\tau^2+1/20}, \frac{(1/20)\tau^2}{\tau^2+1/20}\right)$

- Draw τ^2 from $P(\tau^2 \mid \mu^{\text{diff}}, \mathbf{y}, \zeta) = IG\left(\frac{1}{2} \underbrace{\sum \zeta_i}, \frac{1}{2} \underbrace{\sum \zeta_i \overset{\text{diff}}{\mu_i^2}}\right)$
- Draw P_1 from

  $$P(P_1 \mid \mathbf{y}, \zeta, \mu, \tau^2) = \text{Beta}(1 + \sum \zeta_i, 1 + 500 - \sum \zeta_i)$$

Example: gene testing

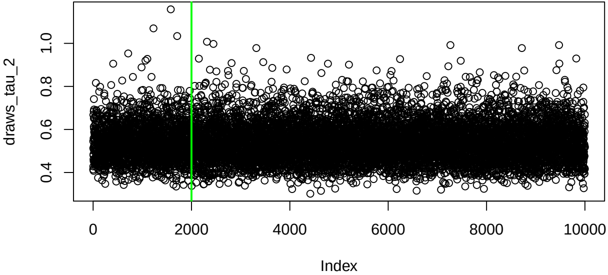
- Gibbs draws for P_1:



- Estimate $\approx 21\%$ of genes show an alcohol effect

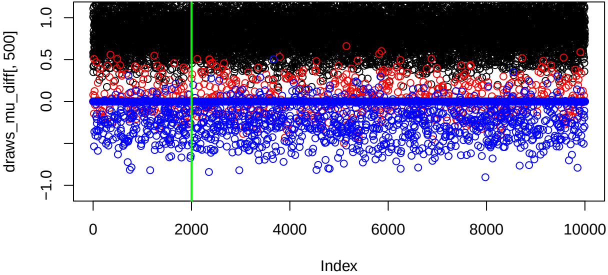
Example: gene testing

- Gibbs draws for τ^2:



Example: gene testing

- Gibbs draws for μ^{diff}, three genes:



- Estimated probability of an effect for the red gene: 0.06
- For the blue gene: 0.12
- For the **black** gene: 0.99

Example: gene testing

- p_D for the present model is 179.8
- DIC is 106.57
- Suggests this is a good compromise between
  - Null model ($DIC = 1029$)
  - Model with an effect in every gene ($DIC = 267.1$)