Model Selection:

The Red-haired stepchild of Bayesian Analysis

Why model selection?

- Inference in Bayes (like in ML) is conditional on the model being true
- We don't know the model is true
  - Affects reliability of posterior distribution, precision, etc

Frequentist/ Classical approaches

- Likelihood ratio testing

- AIC

Bayes- the dog's lunch

- Bayes Factors'
- Information criteria
  - AIC
  - BIC
  - DIC
- Cross validation
- Reversible jump MCMC

Bayes statement of problem

The problem till now:

$$[\underline{\theta}, X] = [\underline{\theta} \mid X][X] = [X \mid \underline{\theta}][\underline{\theta}]$$

$$[\underline{\theta} \mid X] \propto [X \mid \underline{\theta}][\underline{\theta}]$$

likelihood

prior

Averaging by simulation to get individual parameters

With multiple models--

$$[\{\underline{\theta}_m\}, m, X] = [\{\underline{\theta}_m\}, m \mid X][X] = [X \mid \{\underline{\theta}_m\}][\{\underline{\theta}_m\} \mid m][m]$$

$$[\{\underline{\theta}_m\}, m \mid X] \propto [X \mid \{\underline{\theta}_m\}][\{\underline{\theta}_m\} \mid m][m]$$

Likelihood for
Model m

Priors
For model m

Prior that model m is true

What we want…

Posterior inference on parameters, averaged across models

$$[\underline{\theta} \mid X] \propto \int_m [X \mid \{\underline{\theta}_m\}][\{\underline{\theta}_m\} \mid m][m]dm$$

Posterior probability that model *m* is true

$$[m \mid X] \propto \int_\theta [X \mid \{\underline{\theta}_m\}][\{\underline{\theta}_m\} \mid m][m]d\underline{\theta}_m$$

Consider 2 models..

$$[m_1 \mid X] \propto [X \mid m_1][m_1]$$

$$[m_2 \mid X] \propto [X \mid m_2][m_2]$$

Posterior odds

Bayes factor

Prior odds

$$\frac{[m_1 \mid X]}{[m_2 \mid X]} = \frac{[X \mid m_1]}{[X \mid m_2]} \times \frac{[m_1]}{[m_2]}$$

Bayes factor/ aka “likelihood ratio”

$$BF = \frac{[X \mid m_1]}{[X \mid m_2]}$$

$$\log(BF) = \log([X \mid m_1]) - \log([X \mid m_2])$$

>>1: evidence for model 1 (log BF>>0)

<<1: evidence for model 2 (log BF<<0)

~1: evidence for either (neither) model (log BF~0)

>2 models

$$[m_i \mid X] = \frac{[X \mid m_i][m_i]}{\sum_{k=1}^{m} [X \mid m_k][m_k]}$$

Bayesian Information Criterion (ΔBIC)

- "Penalty" for number of parameters
- Related to Bayes Factor
- Similar to ΔAIC

$$\Delta BIC = -2\ln(BF) - (p_1 - p_2)\log(n)$$

Asymptotically = BF

($\rightarrow$ 0 as n $\rightarrow \infty$)

AIC

$$AIC_i = -2\ln([X \mid m_i]) + 2p_i$$

Deviance for model i

Complexity penalty

$$BIC_i = -2\ln([X \mid m_i]) + p_i \log(n)$$

Implementation in Winbugs

- Deviance can be monitored as a node or calculated in model

$$Deviance_i = -2\ln([X \mid m_i])$$

So:

$$\Delta BIC = Deviance_1 - Deviance_2 - (p_1 - p_2)\log(n)$$

$$BF = \exp[Deviance_1 - Deviance_2] \times 0.5)$$

$$AIC_i = Deviance_i + 2p_i$$

Winbugs example

- binomial_models.odc

DIC

- Proposed by Spiegelhalter et al
- Estimates “effective number of parameters”
- Enabled by DIC ‘tool’ in WINBUGS

$$DIC_i = \overline{Deviance_i} + p_D$$

$$p_D = \overline{Deviance_i} - Deviance(\bar{\theta})$$

Mean of posterior deviances (averaged over parameters, data)– the usual

Deviance evaluated at posterior parameter means

Issues

- BF ok for posterior prediction of data (see DIC notes)
- AIC counts parameters ok in non-hierarchical models but not in hierarchical ones
- BIC can overstate model strength as n gets big
- DIC
  - Sometimes doesn't work without intervention
  - Can results in negative (nonsensical) p_D

Alternative: RJMCMC

- Reversible Jump MCMC
- Treats model probability as another sampling problem
  - Jumps between parameter spaces of alternative models
  - Time spent in model *i* used to calculate marginal probability of model

RJ-MCMC

Want to sample from

$$[\{\underline{\theta}_m\}, m \mid X] \propto [X \mid \{\underline{\theta}_m\}][\{\underline{\theta}_m\} \mid m][m]$$

Essentially we need a rule for jumping back and forth between models j and k

$$[\{\underline{\theta}_j\}, j \mid X] \Leftrightarrow [\{\underline{\theta}_k\}, k \mid X]$$

Jumping distribution $J(k, j)$

Acceptance probability

$$\alpha(j,k) = \min\left[1, \frac{[\{\underline{\theta}_k\}, k \mid X] J(k, j)}{[\{\underline{\theta}_j\}, j \mid X] J(j,k)}\right]$$

Proceeds iteratively in a similar manner to Metropolis-Hastings

RJ-MCMC

- Provides a fully Bayesian treatment of model probability using MCMC
- In theory should work for any identifiable model
- As yet no general algorithm
- Potentially very slow to converge

Information theoretic distance measures (Jim Peterson)

Akaike information criteria

AIC = -2ln(likelihood) + 2*K

- *Assumes no true model*
- *Tends to overfit models*

Bayesian or Schwartz's information criteria

BIC = -2ln(likelihood) + KlnN

- *Assumes true model is contained in the set*
- *Tends to overstate evidence with large datasets*

Consistent Akaike information criteria

CAIC = -2ln(likelihood) + Kln(N+1)

- *Similar to BIC*

Deviance information criteria

DIC = used in MCMC modeling (more later)

RJ-MCMC

Assignment

- For your examples or other data construct 2-3 alternative models (increasing or decreasing complexity but same data)
- Compare models using BF, BIC, AIC, and DIC
- Make a recommendation based on model selection criteria and/or GOF