Model criticism and selection

Three related issues to consider:
Model criticism and selection

Three related issues to consider:

- **Robustness:** Are any model assumptions having an undue impact on the results? (text, Sec. 4.2)
Model criticism and selection

Three related issues to consider:

- **Robustness**: Are any model assumptions having an undue impact on the results? *(text, Sec. 4.2)*
- **Assessment**: Does the model provide adequate fit to the data? *(text, Sec. 4.3)*
Model criticism and selection

Three related issues to consider:

- **Robustness**: Are any model assumptions having an undue impact on the results? (text, Sec. 4.2)
- **Assessment**: Does the model provide adequate fit to the data? (text, Sec. 4.3)
- **Selection**: Which model (or models) should we choose for final presentation? (text, Secs. 4.4–4.6)
Model criticism and selection

Three related issues to consider:

- **Robustness**: Are any model assumptions having an undue impact on the results? (text, Sec. 4.2)
- **Assessment**: Does the model provide adequate fit to the data? (text, Sec. 4.3)
- **Selection**: Which model (or models) should we choose for final presentation? (text, Secs. 4.4–4.6)

Consider each in turn...
Sensitivity analysis

Make modifications to an assumption and recompute the posterior; any impact on interpretations or decisions?

- **No:** The data are strongly informative with respect to this assumption *(robustness)*
Sensitivity analysis

Make modifications to an assumption and recompute the posterior; any impact on interpretations or decisions?

- **No:** The data are strongly informative with respect to this assumption *(robustness)*
- **Yes:** Document the sensitivity, think more carefully about it, and perhaps collect more data.
Sensitivity analysis

Make modifications to an assumption and recompute the posterior; any impact on interpretations or decisions?

- **No**: The data are strongly informative with respect to this assumption (robustness)

- **Yes**: Document the sensitivity, think more carefully about it, and perhaps collect more data.

Examples of assumptions to modify: increasing/decreasing a prior mean by one prior s.d.; doubling/halving a prior s.d.; case deletion.
Sensitivity analysis

Make modifications to an assumption and recompute the posterior; any impact on interpretations or decisions?

- **No:** The data are strongly informative with respect to this assumption (robustness)
- **Yes:** Document the sensitivity, think more carefully about it, and perhaps collect more data.

Examples of assumptions to modify: increasing/decreasing a prior mean by one prior s.d.; doubling/halving a prior s.d.; case deletion.

- **Importance sampling** and **asymptotic methods** can greatly reduce computational overhead, even if these methods were *not* used in analysis of original model.
Sensitivity analysis

Make modifications to an assumption and recompute the posterior; any impact on interpretations or decisions?

- **No:** The data are strongly informative with respect to this assumption (robustness)

- **Yes:** Document the sensitivity, think more carefully about it, and perhaps collect more data.

- Examples of assumptions to modify: increasing/decreasing a *prior mean* by one prior s.d.; doubling/halving a *prior s.d.*; case deletion.

- Importance sampling and asymptotic methods can greatly reduce computational overhead, even if these methods were *not* used in analysis of original model.

⇒ Run and diagnose convergence for “base” model; use approximate method for robustness study
Prior partitioning

– a “backwards” approach to robustness!

What if the range of plausible assumptions is unimaginably broad, as in the summary of a government-sponsored clinical trial?
Prior partitioning – a “backwards” approach to robustness!

What if the range of plausible assumptions is unimaginably broad, as in the summary of a government-sponsored clinical trial?

Potential solution: Determine the set of prior inputs that are consistent with a given conclusion, given the data observed so far. The consumer may then compare this prior class to his/her own personal prior beliefs.
Prior partitioning
– a “backwards” approach to robustness!

What if the range of plausible assumptions is unimaginably broad, as in the summary of a government-sponsored clinical trial?

Potential solution: Determine the set of prior inputs that are consistent with a given conclusion, given the data observed so far. The consumer may then compare this prior class to his/her own personal prior beliefs.

Thus we are partitioning the prior class based on possible outcomes.
Prior partitioning

– a “backwards” approach to robustness!

What if the range of plausible assumptions is unimaginably broad, as in the summary of a government-sponsored clinical trial?

Potential solution: Determine the set of prior inputs that are consistent with a given conclusion, given the data observed so far. The consumer may then compare this prior class to his/her own personal prior beliefs.

Thus we are partitioning the prior class based on possible outcomes.

Example: Find set of all prior means $\mu$ such that

$$P(\theta \geq 0 | y) > .025$$

(for otherwise, we will decide $\theta < 0$).
Model assessment

Many of the tools mentioned in Chapter 2 are now easy to compute via Monte Carlo methods!

Example: Find the cross-validation residual

\[ r_i = y_i - E(y_i | y_{(i)}) \]

where \( y_{(i)} \) denotes the vector of all the data except the \( i^{th} \) value, i.e.

\[ y_{(i)} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)' \]
Model assessment

Using MC draws \( \theta^{(g)} \sim p(\theta|y) \), we have

\[
E(y_i|y_{(i)}) = \int \int y_i f(y_i|\theta) p(\theta|y_{(i)}) dy_i d\theta
\]

\[
= \int E(y_i|\theta) p(\theta|y_{(i)}) d\theta
\]

\[
\approx \int E(y_i|\theta) p(\theta|y) d\theta
\]

\[
\approx \frac{1}{G} \sum_{g=1}^{G} E(y_i|\theta^{(g)})
\]
Model assessment

Using MC draws $\theta^{(g)} \sim p(\theta|y)$, we have

$$E(y_i|y_{(i)}) = \int \int y_i f(y_i|\theta) p(\theta|y_{(i)}) dy_i d\theta$$

$$= \int E(y_i|\theta) p(\theta|y_{(i)}) d\theta$$

$$\approx \int E(y_i|\theta) p(\theta|y) d\theta$$

$$\approx \frac{1}{G} \sum_{g=1}^{G} E(y_i|\theta^{(g)}) .$$

Approximation should be adequate unless the dataset is small and $y_i$ is an extreme outlier.
Model assessment

Using MC draws $\theta^{(g)} \sim p(\theta|y)$, we have

$$E(y_i|y_{(i)}) = \int \int y_i f(y_i|\theta)p(\theta|y_{(i)})dy_id\theta$$

$$\approx \int E(y_i|\theta)p(\theta|y_{(i)})d\theta$$

$$\approx \frac{1}{G} \sum_{g=1}^{G} E(y_i|\theta^{(g)}) .$$

Approximation should be adequate unless the dataset is small and $y_i$ is an extreme outlier.

Same $\theta^{(g)}$'s may be used for each $i = 1, \ldots, n$. 

Chapter 4: Model Criticism and Selection – p. 5/17
Bayes factors

the most basic Bayesian model choice tool!

Given models $M_1$ and $M_2$, computable as

$$BF = \frac{p(y \mid M_1)}{p(y \mid M_2)}.$$
Bayes factors

the most basic Bayesian model choice tool!

Given models $M_1$ and $M_2$, computable as

$$BF = \frac{p(y | M_1)}{p(y | M_2)}.$$ 

Sadly, unlike posteriors and predictives, marginal distributions are not easily estimated via MCMC! So...
Bayes factors

the most basic Bayesian model choice tool!

Given models $M_1$ and $M_2$, computable as

$$BF = \frac{p(y \mid M_1)}{p(y \mid M_2)}.$$

Sadly, unlike posteriors and predictives, marginal distributions are not easily estimated via MCMC! So...

◊ **Direct methods:** Since $p(y) = \int f(y \mid \theta) p(\theta) d\theta$, we could draw $\theta^{(g)} \sim p(\theta)$ and compute

$$\hat{p}(y) = \frac{1}{G} \sum_{g=1}^{G} f(y \mid \theta^{(g)}).$$

Easy, but terribly inefficient.
Bayes factors

Better: Draw $\theta^{(g)} \sim p(\theta|y)$ and compute the harmonic mean estimate

$$\hat{p}(y) = \left[ \frac{1}{G} \sum_{g=1}^{G} \frac{1}{f(y | \theta^{(g)})} \right]^{-1},$$

But this is terribly unstable (division by 0)!
Bayes factors

Better: Draw $\theta^{(g)} \sim p(\theta|y)$ and compute the harmonic mean estimate

$$\hat{p}(y) = \left[ \frac{1}{G} \sum_{g=1}^{G} \frac{1}{f(y | \theta^{(g)})} \right]^{-1},$$

But this is terribly unstable (division by 0)!

Better yet: try

$$\hat{p}(y) = \left[ \frac{1}{G} \sum_{g=1}^{G} \frac{h(\theta^{(g)})}{f(y | \theta^{(g)}) p(\theta^{(g)})} \right]^{-1},$$

where $\theta^{(g)} \sim p(\theta|y)$ and $h(\theta) \approx p(\theta|y)$. (If $h$ equals the prior, we get the harmonic mean again.)
Other Classes of Alternatives

◊ Marginal densities from the Gibbs sampler: Use the identity 

\[ p(y) = \frac{f(y|\theta)p(\theta)}{p(\theta|y)} \]

with a MCMC-based estimate of 

\[ p(\theta|y) \]

evaluated at \( \theta = \theta_0 \).

⇒ text, Sec. 4.4; Chib (1995), Chib and Jeliazkov (2001)
Other Classes of Alternatives

◊ Marginal densities from the Gibbs sampler: Use the identity \( p(y) = f(y|\theta)p(\theta)/p(\theta|y) \) with a MCMC-based estimate of \( p(\theta|y) \), evaluated at \( \theta = \theta_0 \).

⇒ text, Sec. 4.4; Chib (1995), Chib and Jeliazkov (2001)

◊ Sampling over model space alone: Integrate the \( \theta_j \) out of the model before sampling begins (feasible for regression and some discrete data settings).

⇒ text, Sec. 4.5
Other Classes of Alternatives

♦ Marginal densities from the Gibbs sampler: Use the identity \( p(y) = f(y|\theta)p(\theta)/p(\theta|y) \) with a MCMC-based estimate of \( p(\theta|y) \), evaluated at \( \theta = \theta_0 \).

⇒ text, Sec. 4.4; Chib (1995), Chib and Jeliazkov (2001)

♦ Sampling over model space alone: Integrate the \( \theta_j \) out of the model before sampling begins (feasible for regression and some discrete data settings).

⇒ text, Sec. 4.5

♦ Sampling over model and parameter space: Include the model indicator \( M \) in the sampling algorithm, producing a stream of samples \( \{M(g)\}_{g=1}^{G} \) from \( p(M|y) \) which in turn produces \( BF \).

⇒ text, Sec. 4.5; typically requires reversible jump MCMC, or a product space search....
Product Space Search

Suppose model \( j \) has parameter vector \( \theta_j \in \mathbb{R}^{n_j} \) and prior probability \( \pi_j, j = 1, \ldots, K \).
(Note the models need not be nested!)
Product Space Search

Suppose model $j$ has parameter vector $\theta_j \in \mathbb{R}^{n_j}$ and prior probability $\pi_j, j = 1, \ldots, K$.
(Note the models need not be nested!)

Carlin and Chib (1995): Sample over the product space $\prod_{j=1}^{K} \mathbb{R}^{n_j}$. for model $j$, specify:

- likelihood: $f(y | \theta_j, j)$
- prior: $p(\theta_j | j)$
- pseudopriors: $p(\theta_j | M \neq j)$
Product Space Search

Suppose model \( j \) has parameter vector \( \theta_j \in \mathbb{R}^{n_j} \) and prior probability \( \pi_j, j = 1, \ldots, K \).
(Note the models need not be nested!)

Carlin and Chib (1995): Sample over the product space \( \prod_{j=1}^{K} \mathbb{R}^{n_j} \). for model \( j \), specify:

- **likelihood:** \( f(y|\theta_j, j) \)
- **prior:** \( p(\theta_j|j) \)
- **pseudopriors:** \( p(\theta_j|M \neq j) \)

Full conditional distributions for \( \theta_j \):

\[
p(\theta_j|\theta_{i \neq j}, M, y) \propto \begin{cases} 
  f(y|\theta_j, j)p(\theta_j|j), & M = j \\
  p(\theta_j|M \neq j), & M \neq j 
\end{cases}
\]
Product Space Search

- Full conditional distribution for $M$:

$$p(M = j | \theta, y) = A_j / \sum_{k=1}^{K} A_k$$

where $A_k = f(y | \theta_k, k) \left\{ \prod_{i=1}^{K} p(\theta_i | k) \right\} \pi_k.$
Product Space Search

- Full conditional distribution for $M$:

$$p(M = j | \theta, y) = A_j / \sum_{k=1}^{K} A_k$$

where $A_k = f(y | \theta_k, k) \left\{ \prod_{i=1}^{K} p(\theta_i | k) \right\} \pi_k$.

- If $M^{(g)} = j$, sample $\theta_j^{(g)}$ from the model $j$ full conditional;
- If $M^{(g)} \neq j$, sample $\theta_j^{(g)}$ from the pseudoprior.
Product Space Search

- Full conditional distribution for $M$:

$$p(M = j | \theta, y) = \frac{A_j}{\sum_{k=1}^{K} A_k}$$

where $A_k = f(y | \theta_k, k) \left\{ \prod_{i=1}^{K} p(\theta_i | k) \right\} \pi_k$.

- If $M^{(g)} = j$, sample $\theta_j^{(g)}$ from the model $j$ full conditional;
- If $M^{(g)} \neq j$, sample $\theta_j^{(g)}$ from the pseudoprior.

- To insure good mixing in the $M^{(g)}$ chain, select pseudopriors close to the true posteriors, i.e.

$$p(\theta_j | M \neq j) \approx p(\theta_j | j, y)$$
**Product Space Search**

- Full conditional distribution for $M$:

  \[
p(M = j | \theta, y) = A_j / \sum_{k=1}^{K} A_k
  \]

  where $A_k = f(y | \theta_k, k) \left\{ \prod_{i=1}^{K} p(\theta_i | k) \right\} \pi_k$.

- If $M^{(g)} = j$, sample $\theta_j^{(g)}$ from the model $j$ full conditional;
  
  If $M^{(g)} \neq j$, sample $\theta_j^{(g)}$ from the pseudoprior.

- To insure good mixing in the $M^{(g)}$ chain, select pseudopriors close to the true posteriors, i.e.

  \[
p(\theta_j | M \neq j) \approx p(\theta_j | j, y)
  \]

  ⇒ A nuisance, especially if $K$ is large!
“Metropolized” Product Space Search

Use reversible jump to “Metropolize” the model choice step above:
“Metropolized” Product Space Search

- Use reversible jump to “Metropolize” the model choice step above:
- Propose a move $j \rightarrow j'$ with probability $S(j, j')$
“Metropolized” Product Space Search

- Use reversible jump to “Metropolize” the model choice step above:
  - Propose a move \( j \rightarrow j' \) with probability \( S(j, j') \)
  - Draw \( \theta_{j'} \) from the pseudoprior \( p(\theta_{j'} | M \neq j') \)
“Metropolized” Product Space Search

Use reversible jump to “Metropolize” the model choice step above:

- Propose a move \( j \rightarrow j' \) with probability \( S(j, j') \)
- Draw \( \theta_{j'} \) from the pseudoprior \( p(\theta_{j'} | M \neq j') \)
- Accept this move with probability

\[
\begin{align*}
r & = \min \left( 1, \frac{A_{j'} S(j', j)}{A_j S(j, j')} \right) \\
& = \min \left( 1, \frac{f(y|\theta_{j'}, j') p(\theta_{j'} | j') p(\theta_j | j') \pi_{j'} S(j', j)}{f(y|\theta_j, j) p(\theta_j | j) p(\theta_{j'} | j) \pi_j S(j, j')} \right)
\end{align*}
\]

since all other pseudopriors cancel in this ratio.
“Metropolized” Product Space Search

- Use reversible jump to “Metropolize” the model choice step above:
  - Propose a move \( j \rightarrow j' \) with probability \( S(j, j') \)
  - Draw \( \theta_{j'} \) from the pseudoprior \( p(\theta_{j'} | M \neq j') \)
  - Accept this move with probability

\[
\begin{align*}
r & = \min \left( 1, \frac{A_{j'} S(j', j)}{A_j S(j, j')} \right) \\
& = \min \left( 1, \frac{f(y | \theta_{j'}, j') p(\theta_{j'} | j') p(\theta_j | j) \pi_j S(j', j)}{f(y | \theta_j, j) p(\theta_j | j) p(\theta_{j'} | j') \pi_{j'} S(j, j')} \right)
\end{align*}
\]

since all other psuedopriors cancel in this ratio.
- This approach often works nearly as well as full-blown reversible jump, yet easier to understand/implement!
Predictive Model Selection

Less formal approaches, useful when Bayes factor is unavailable or inappropriate (e.g., when using improper priors). These include:

- **Cross-validatory checks**, such as $\sum_i \log f(y_i^{obs}|y(i))$ or $\sum_i [y_i - E(y_i|y(i))]^2$. 
Predictive Model Selection

Less formal approaches, useful when Bayes factor is unavailable or inappropriate (e.g., when using improper priors). These include:

- **Cross-validatory checks**, such as \( \sum_i \log f(y_{i \text{obs}}|y_{(i)}) \) or \( \sum_i [y_i - E(y_{y_{(i)}})]^2 \).

- **Expected predicted “model discrepancy,”**

\[
E[d(y_{\text{new}}, y_{\text{obs}})|y_{\text{obs}}, M_i],
\]

where \( d(y_{\text{new}}, y_{\text{obs}}) \) is an appropriate discrepancy function, e.g.,

\[
d(y_{\text{new}}, y_{\text{obs}}) = (y_{\text{new}} - y_{\text{obs}})^T(y_{\text{new}} - y_{\text{obs}}).
\]

Choose the model that minimizes discrepancy!
Predictive Model Selection

Likelihood criteria: think of $\ell \equiv \log L(\theta)$ as a parametric function of interest, and compute

$$\hat{\ell} \equiv E[\log L(\theta)|y] \approx \frac{1}{G} \sum_{g=1}^{G} \log L(\theta^{(g)})$$

as an overall measure of model fit.
Predictive Model Selection

**Likelihood criteria:** think of $\ell \equiv \log L(\theta)$ as a parametric function of interest, and compute

$$\hat{\ell} \equiv E[\log L(\theta)|y] \approx \frac{1}{G} \sum_{g=1}^{G} \log L(\theta^{(g)})$$

as an overall measure of model fit.

**Penalized likelihood criteria:** Subtract a “penalty” from the likelihood score, in order to avoid flooding unhelpful predictors into the model. Most common example: the Bayesian Information (Schwarz) Criterion,

$$\hat{BIC} = 2\hat{\ell} - p \log n$$

where $p$ is the number of parameters in the model, and $n$ is the number of datapoints.
Extension to Hierarchical Models

- Penalized likelihood criteria (BIC, AIC) trade off “fit” against “complexity”
Extension to Hierarchical Models

- Penalized likelihood criteria (BIC, AIC) trade off “fit” against “complexity”
- But what is the “complexity” of a hierarchical model?
Extension to Hierarchical Models

- Penalized likelihood criteria (BIC, AIC) trade off “fit” against “complexity”
- But what is the “complexity” of a hierarchical model?
- **Example**: One-way ANOVA model

\[
Y_i | \theta_i \sim^\text{ind} N(\theta_i, 1/\tau_i) \quad \text{and} \quad \theta_i \sim^\text{iid} N(\mu, 1/\lambda), \quad i = 1, \ldots, p
\]

Suppose \(\mu, \lambda,\) and the \(\tau_i\) are known. How many parameters are in this model?
Extension to Hierarchical Models

- Penalized likelihood criteria (BIC, AIC) trade off “fit” against “complexity”
- But what is the “complexity” of a hierarchical model?
- **Example:** One-way ANOVA model

\[
Y_i | \theta_i \sim \text{ind } \mathcal{N}(\theta_i, 1/\tau_i) \quad \text{and} \quad \theta_i \sim \text{iid } \mathcal{N}(\mu, 1/\lambda), \ i = 1, \ldots, p
\]

Suppose \( \mu, \lambda, \) and the \( \tau_i \) are known. How many parameters are in this model?
- If \( \lambda = \infty \), all \( \theta_i = \mu \) and there are 0 free parameters
**Extension to Hierarchical Models**

- Penalized likelihood criteria (BIC, AIC) trade off “fit” against “complexity”
- But what is the “complexity” of a hierarchical model?
- **Example:** One-way ANOVA model

\[
Y_i | \theta_i \sim \text{iid } N(\theta_i, 1/\tau_i) \quad \text{and} \quad \theta_i \sim N(\mu, 1/\lambda), \quad i = 1, \ldots, p
\]

Suppose \( \mu, \lambda, \) and the \( \tau_i \) are known. How many parameters are in this model?

- If \( \lambda = \infty \), all \( \theta_i = \mu \) and there are 0 free parameters
- If \( \lambda = 0 \), the \( \theta_i \) are unconstrained and there are \( p \) free parameters
Extension to Hierarchical Models

Penalized likelihood criteria (BIC, AIC) trade off “fit” against “complexity”

But what is the “complexity” of a hierarchical model?

Example: One-way ANOVA model

\[ Y_i | \theta_i \overset{ind}{\sim} N(\theta_i, 1/\tau_i) \quad \text{and} \quad \theta_i \overset{iid}{\sim} N(\mu, 1/\lambda), \quad i = 1, \ldots, p \]

Suppose \( \mu, \lambda, \) and the \( \tau_i \) are known. How many parameters are in this model?

- If \( \lambda = \infty \), all \( \theta_i = \mu \) and there are 0 free parameters
- If \( \lambda = 0 \), the \( \theta_i \) are unconstrained and there are \( p \) free parameters

In practice, \( 0 < \lambda < \infty \) so the “effective number of parameters” is somewhere in between! How to define?....
**Hierarchical model complexity**

Proposal: use the *effective number of parameters*,

\[
D_p = E_{\theta|y}[D] - D(E_{\theta|y}[\theta]) = \bar{D} - D(\bar{\theta}),
\]

where \(D(\theta) = -2 \log f(y|\theta) + 2 \log h(y)\) is the *deviance* score, computed from the likelihood \(f(y|\theta)\) and a standardizing function \(h(y)\).
Hierarchical model complexity

Proposal: use the effective number of parameters,

\[ p_D = E_{\theta|y}[D] - D(E_{\theta|y}[\theta]) = \bar{D} - D(\bar{\theta}), \]

where \( D(\theta) = -2 \log f(y|\theta) + 2 \log h(y) \)

is the deviance score, computed from the likelihood \( f(y|\theta) \) and a standardizing function \( h(y) \).

Example: For the one-way ANOVA model,

\[ p_D = \sum_{i=1}^{p} \frac{\tau_i}{\tau_i + \lambda}, \]
Hierarchical model complexity

Proposal: use the effective number of parameters,

\[ p_D = E_{\theta|y}[D] - D(E_{\theta|y}[\theta]) = \bar{D} - D(\bar{\theta}) , \]

where \( D(\theta) = -2 \log f(y|\theta) + 2 \log h(y) \)

is the deviance score, computed from the likelihood \( f(y|\theta) \) and a standardizing function \( h(y) \).

Example: For the one-way ANOVA model,

\[ p_D = \sum_{i=1}^{p} \frac{\tau_i}{\tau_i + \lambda} , \]

Clearly \( 0 \leq p_D \leq p \) as desired
Hierarchical model complexity

Proposal: use the effective number of parameters,

\[ p_D = E_{\theta|y}[D] - D(E_{\theta|y}[\theta]) = \bar{D} - D(\bar{\theta}) , \]

where \( D(\theta) = -2 \log f(y|\theta) + 2 \log h(y) \)

is the deviance score, computed from the likelihood \( f(y|\theta) \) and a standardizing function \( h(y) \).

Example: For the one-way ANOVA model,

\[ p_D = \sum_{i=1}^{p} \frac{\tau_i}{\tau_i + \lambda} , \]

Clearly \( 0 \leq p_D \leq p \) as desired

If we place a hyperprior on \( \lambda \), the effective model size \( p_D \) will depend on the dataset!
Model selection via DIC

Given the $p_D$ measure of model complexity, suppose we now summarize fit of a model by

$$\bar{D} = E_{\theta|y}[D],$$

Chapter 4: Model Criticism and Selection – p. 16/17
Model selection via DIC

- Given the $p_D$ measure of model complexity, suppose we now summarize fit of a model by

$$\bar{D} = E_{\theta|y}[D],$$

- Compare models via the **Deviance Information Criterion**, 

$$DIC = \bar{D} + p_D = D(\bar{\theta}) + 2p_D,$$

a generalization of the Akaike Information Criterion (AIC), since $AIC \approx \bar{D} + p$ for nonhierarchical models.
Model selection via DIC

Given the $p_D$ measure of model complexity, suppose we now summarize fit of a model by

$$\bar{D} = E_{\theta|y}[D] ,$$

Compare models via the Deviance Information Criterion,

$$DIC = \bar{D} + p_D = D(\bar{\theta}) + 2p_D ,$$

a generalization of the Akaike Information Criterion (AIC), since $AIC \approx \bar{D} + p$ for nonhierarchical models.

Smaller values of DIC indicate preferred models.
Model selection via DIC

Given the $p_D$ measure of model complexity, suppose we now summarize fit of a model by

$$\bar{D} = E_{\theta|y}[D] ,$$

Compare models via the Deviance Information Criterion,

$$DIC = \bar{D} + p_D = D(\bar{\theta}) + 2p_D ,$$

a generalization of the Akaike Information Criterion (AIC), since $AIC \approx \bar{D} + p$ for nonhierarchical models.

Smaller values of DIC indicate preferred models.

While $p_D$ has a scale (effective model size), DIC does not, so only differences in DIC across models matter.
Issues in using DIC

- $p_D$ and DIC are very broadly applicable provided $p(y|\theta)$ is available in closed form
Issues in using DIC

- $p_D$ and DIC are very broadly applicable provided $p(y|\theta)$ is available in closed form.
- Both building blocks of DIC and $p_D$, $E_{\theta|y}[D]$ and $D(E_{\theta|y}[\theta])$, are easily estimated via MCMC methods.
Issues in using DIC

- $p_D$ and DIC are very broadly applicable provided $p(y|\theta)$ is available in closed form.
- Both building blocks of DIC and $p_D, E_{\theta|y}[D]$ and $D(E_{\theta|y}[\theta])$, are easily estimated via MCMC methods.
- ...and in fact are directly available within WinBUGS!
Issues in using DIC

- $p_D$ and DIC are very broadly applicable provided $p(y|\theta)$ is available in closed form.

- Both building blocks of DIC and $p_D, E_{\theta|y}[D]$ and $D(E_{\theta|y}[\theta])$, are easily estimated via MCMC methods.

- ...and in fact are directly available within WinBUGS!

- $p_D$ and DIC may not be invariant to reparametrization.
Issues in using DIC

- $p_D$ and DIC are very broadly applicable provided $p(y|\theta)$ is available in closed form.

- Both building blocks of DIC and $p_D$, $E_{\theta|y}[D]$ and $D(E_{\theta|y}[\theta])$, are easily estimated via MCMC methods.

- ...and in fact are directly available within WinBUGS!

- $p_D$ and DIC may not be invariant to reparametrization.

- $p_D$ can be negative for non-log-concave likelihoods, or when there is strong prior-data conflict.
Issues in using DIC

- $p_D$ and DIC are very broadly applicable provided $p(y|\theta)$ is available in closed form
- Both building blocks of DIC and $p_D$, $E_{\theta|y}[D]$ and $D(E_{\theta|y}[\theta])$, are easily estimated via MCMC methods
- ...and in fact are directly available within WinBUGS!

- $p_D$ and DIC may not be invariant to reparametrization
- $p_D$ can be negative for non-log-concave likelihoods, or when there is strong prior-data conflict
- $p_D$ and DIC will depend on our “focus” (i.e., what is considered to be part of the likelihood):
  - $f(y|\theta)$: “focused on $\theta$”
  - $p(y|\eta) = \int f(y|\theta)p(\theta|\eta)d\theta$: “focused on $\eta$”