Some DIC slides

David Spiegelhalter
MRC Biostatistics Unit, Cambridge

with thanks to:
Nicky Best
Dave Lunn
Andrew Thomas

IceBUGS: Finland, 11^{th}-12^{th} February 2006

Model comparison

What is the 'deviance'?

- For a likelihood $p(y|\theta)$, we define the deviance as

$$D(\theta) = -2 \log p(y|\theta) \tag{1}$$

- In WinBUGS the quantity `deviance` is automatically calculated, where θ are the parameters that appear in the stated sampling distribution of y

- The full normalising constants for $p(y|\theta)$ are included in `deviance`

- *e.g.* for Binomial data `y[i]` `dbin(theta[i],n[i])`, the `deviance` is

$$-2 \left[\sum_i y_i \log \theta_i + (n_i - y_i) \log(1 - \theta_i) + \log \begin{pmatrix} n_i \\ r_i \end{pmatrix} \right]$$

Use of mean deviance as measure of fit

- Dempster (1974) suggested plotting posterior distribution of deviance $D = -2 \log p(y|\theta)$
- Many authors suggested using posterior mean deviance $\overline{D} = \mathbb{E}[D]$ as a measure of fit
- Invariant to parameterisation of θ
- Robust, generally converges well
- But more complex models will fit the data better and so will have smaller $\overline{D}$
- Need to have some measure of 'model complexity' to trade off against $\overline{D}$

Bayesian measures of model dimensionality (Spiegelhalter et al, 2002)

$$\begin{aligned} p_D &= E_{\theta|y}[d_\Theta(y,\theta,\tilde{\theta}(y))] \\ &= E_{\theta|y}[-2\log p(y|\theta)] + 2\log p(y|\tilde{\theta}(y)). \end{aligned}$$

If we take $\tilde{\theta} = E[\theta|y]$, then

$$p_D = \text{“posterior mean deviance - deviance of posterior means”}.$$

In normal linear hierarchical models:

$$p_D = tr(H)$$

where $Hy = \widehat{y}$. Hence H is the ‘hat’ matrix which projects data onto fitted values.

Thus $p_D = \sum h_{ii} = \sum$leverages.

In general, justification depends on asymptotic normality of posterior distribution.

Bayesian model comparison using DIC

- Natural way to compare models is to use criterion based on trade-off between the fit of the data to the model and the corresponding complexity of the model
- Spiegelhalter et al (2002) proposed a Bayesian model comparison criterion based on this principle:

  Deviance Information Criterion, DIC = 'goodness of fit' + 'complexity'

- They measure fit via the deviance

$$D(\theta) = -2 \log L(\text{data}|\theta)$$

- Complexity measured by estimate of the 'effective number of parameters':

$$\begin{aligned} p_D &= E_{\theta|y}[D] - D(E_{\theta|y}[\theta]) \\ &= \overline{D} - D(\overline{\theta}); \end{aligned}$$

  i.e. posterior mean deviance minus deviance evaluated at the posterior mean of the parameters

- The DIC is then defined analagously to AIC as

$$\begin{aligned} \text{DIC} &= D(\overline{\theta}) + 2p_D \\ &= \overline{D} + p_D \end{aligned}$$

  Models with smaller DIC are better supported by the data

- DIC can be monitored in `WinBUGS` from `Inference/DIC` menu

- These quantities are easy to compute in an MCMC run
- Aiming for Akaike-like, cross-validatory, behaviour based on ability to make short-term predictions of a repeat set of similar data.
- Not a function of the marginal likelihood of the data, so *not* aiming for Bayes factor behaviour.
- Do not believe there is any 'true' model.
- p_D is not invariant to reparameterisation (subject of much criticism).
- p_D can be negative!
- Alternative to p_D suggested

p_V: an alternative measure of complexity

- Suppose have non-hierarchical model with weak prior
- Then
$$D(\theta) \approx D(\overline{\theta}) + \chi^2_I :$$
so that $\mathbb{E}(D(\theta)) \approx D(\overline{\theta}) + I$ (leading to $p_D \approx I$ as shown above), and $\text{Var}(D(\theta)) \approx 2I$.
- Thus with negligible prior information, half the variance of the deviance is an estimate of the number of free parameters in the model
- This estimate generally turns out to be remarkably robust and accurate
- Invariant to parameterisation
- This might suggest using $p_V = \text{Var}(D)/2$ as an estimate of the effective number of parameters in a model in more general situations: this was originally tried in a working paper by Spiegelhalter et al (1997), and has since been suggested by Gelman et al (2004).

- Working through distribution theory for simple Normal random-effects model with I groups suggests

$$p_V \approx p_D(2 - p_D/I)$$

, but many assumptions

- So may expect p_V to be larger than p_D when there is moderate shrinkage.

Schools example - Gelman et al

Exam results in 8 schools

Model	$\overline{D}$	p_D	Var(D)	p_V	DIC
Common effect	55.62	1.00	1.41	0.99	56.62
Fixed effects	56.85	7.99	3.98	7.92	64.77
Random effects	55.16	2.92	2.31	2.67	58.08

In this case give similar results, even though considerable shrinkage.

Seeds example

Random-effects logistic regression of $I = 21$ binomial observations, with 3 covariates

	Dbar	Dhat	pD	DIC
r	100.0	87.6	12.4	112.4

$p_D = 12.4$

3 are regression coefficients, so estimated dimensionality of 21 random effects is 9.4.

node	mean	sd	2.5%	median	97.5%	start	sample
deviance	100.0	6.428	89.19	99.45	113.8	1001	10000

Hence $p_V = \text{Var}(D)/2 = 20.7$ parameters: 17.7 is estimated dimensionality of 21 random effects. Seems rather high.

$p_D(2 - p_D/I) \approx 17.5$, which is not a very good approximation to p_V

Which plug-in estimate to use in p_D?

- p_D is not invariant to reparameterisation, *i.e.* which estimate is used in $D(\tilde{\theta})$
- WinBUGS currently uses posterior mean of stochastic parents of θ, *i.e.* if there are stochastic nodes ψ such that $\theta = f(\psi)$, then $D(\tilde{\theta}) = D(f(\overline{\psi}))$
- p_D can be negative if posterior of ψ is very non-normal and so $f(\overline{\psi})$ does not provide a very good estimate of θ.
- Also can get negative p_D if non-log-concave sampling distribution and strong prior-data conflict

Example

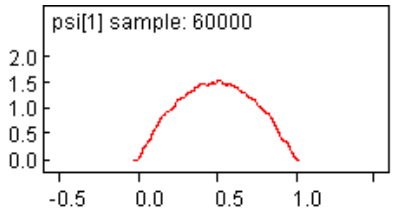
- If $\theta \sim U[0,1]$, then $\psi = \theta^a$ is beta$(a^{-1}, 1)$.
- Suppose we observe $r = 1$ successes out of $n = 2$ Bernoulli trials, so that $r \sim$ Bin$[\theta, n]$
- Consider putting prior on $\psi = \theta$, θ^5 and θ^{20}, each equivalent to uniform prior on θ
- Hence $\theta = \psi^{1/a}, \psi \sim$ Beta$(1/a, 1)$
- Also consider logit$(\theta) \sim$ N$(0, 2)$ (implies $\theta \approx$ U$(0, 1)$).

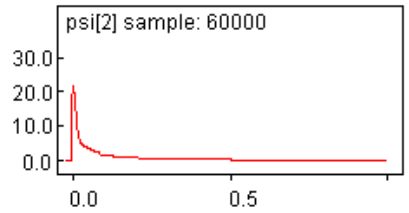
```
r <- 1;  n<- 2 a[1]<-1 ; a[2] <- 5; a[3] <- 20
for (i in 1:3){
     a.inv[i]<- 1/a[i]
     theta[i] <- pow(psi[i], a.inv[i])
     psi[i]  ~ dbeta(a.inv[i] , 1)
     }
r1<- r; r2<-r ; r3 <- r
r1 ~ dbin(theta[1],n)
r2 ~ dbin(theta[2],n)
r3 ~ dbin(theta[3],n)
```

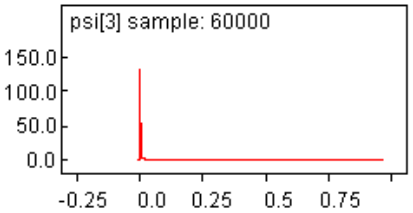
	Dbar	pD	pV	DIC
Uniform	1.94	0.56	0.30	2.50
a=5	1.94	0.41	0.30	2.35
a=20	1.94	-0.39	0.30	1.55
logit	1.88	0.49	0.21	2.36

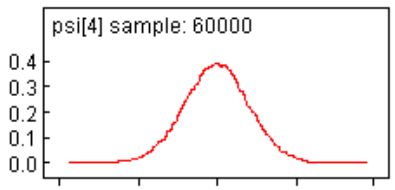
Mean deviances (Dbar) and posteriors for all θ's are the same, but using $\overline{\psi}$ as a plug-in is clearly a bad idea.

Posterior distributions whose means are plugged in









What should we do about it?

- It would be better if WinBUGS used the posterior mean of the 'direct parameters' (eg those that appear in the WinBUGS distribution syntax) to give a 'plug-in' deviance, rather than the posterior means of the stochastic parents.

- Users are free to calculate this themselves: could dump out posterior means of 'direct' parameters in likelihood, then calculate deviance outside WinBUGS or by reading posterior means in as data and checking `deviance` in `node info`

- Lesson: need to be careful with highly non-linear models, where posterior means may not lead to good predictive estimates

- Same problem arises with mixture models

DIC is allowed to be negative - not a problem!

- A probability density $p(y|\theta)$ can be greater than 1 if has a small standard deviation
- Hence a deviance can be negative, and a DIC negative
- Only *differences* in DIC are important: its absolute size is irrelevant
- Suppose observe data (-0.01, 0.01)
- Unknown mean (uniform prior), want to choose between three models with $\sigma = 0.001, 0.01, 0.1$.

	Dbar	Dhat	pD	DIC
y1	177.005	176.046	0.959	177.964
y2	-11.780	-12.740	0.961	-10.819
y3	-4.423	-5.513	1.090	-3.332

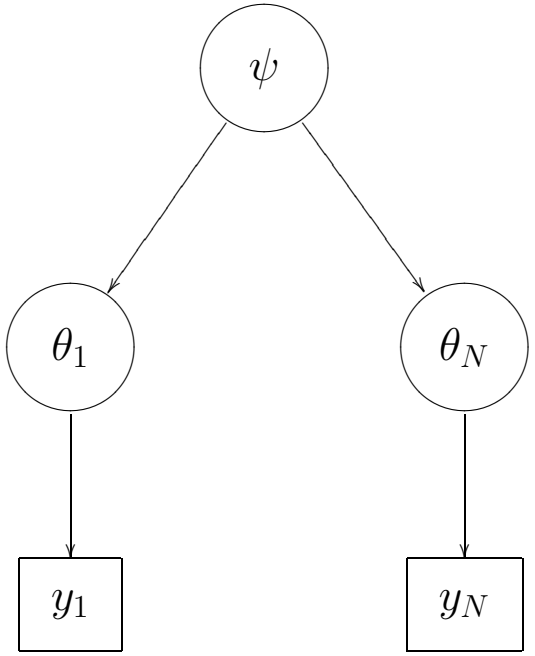
- Each correctly estimates the number of unknown parameters.
- The middle model ($\sigma = 0.01$) has the smallest DIC, which is negative.

Why won't DIC work with mixture likelihoods?

- WinBUGS currently 'greys out' DIC if the likelihood depends on any discrete parameters
- So cannot be used for mixture likelihoods
- Not clear what estimate to plug in for class membership indicator – mode?
- If mixture is represented marginally (ie not using an explicit indicator for class membership), could use $\overline{\theta}$ but could be taking mean of bimodal distribution and get poor estimate
- Celeux et al (2003) have made many suggestions
- Can still be used if prior (random effects) is a mixture

But what is the 'likelihood' in a hierarchical model?

The importance of 'focus'.



- Consider hierarchical model

$$p(y, \theta, \psi) = p(y|\theta)p(\theta|\psi)p(\psi)$$

$$p(y) = \int_{\Theta} p(y|\theta)p(\theta)d\theta = \int_{\Psi} p(y|\psi)p(\psi)d\psi$$

  depending on whether 'focus' is Θ or Ψ.

- The likelihood might be $p(y|\theta)$ or $p(y|\psi)$ depending on focus of analysis

- Prediction is not well-defined in a hierarchical model without stating the focus, which is what remains fixed when making predictions (See later)

What about alternatives for model comparison?

- **DIC**

$$\text{DIC} = D(\overline{\theta}) + 2p_D$$

- **AIC**

$$\text{AIC} = -2\log p(y|\widehat{\psi}) + 2p_\psi$$

  where p_ψ is the number of hyperparameters

- **BIC**

$$\text{BIC} = -2\log p(y|\widehat{\psi}) + p_\psi \log n$$

  An approximation to $-2\log p(y)$, where

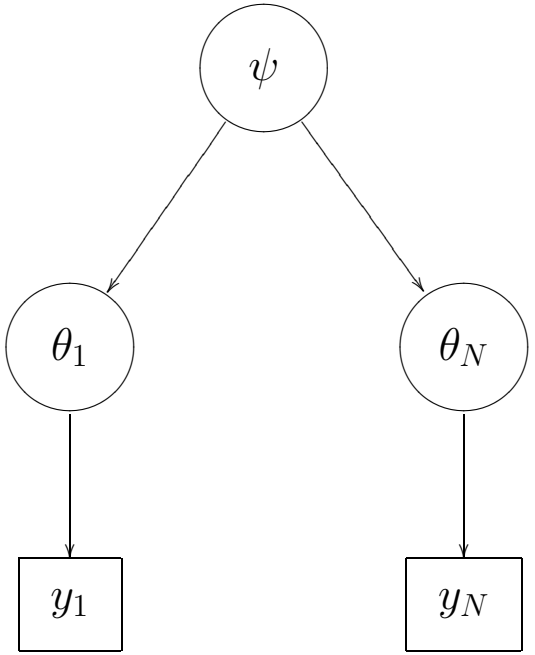
$$p(y) = \int_\Theta p(y|\theta)p(\theta)d\theta = \int_\Psi p(y|\psi)p(\psi)d\psi$$

Depends on objective of analysis Vaida and Blanchard (2005) develop 'conditional' AIC for when focus is random effects - this counts parameters using the 'hat' matrix dimensionality $p = tr(H)$, and so is restricted to normal linear models.

- Interested in predicting Y^{rep} *with* θ*'s fixed*?
  - **DIC:**
  - $\log p(Y^{\text{rep}}|\widehat{\theta})$ estimated by $\log p(y|\widehat{\theta})$,
  - penalised by $2p_D$

- Interested in predicting Y^{rep} *with* ψ *fixed*?
  - **AIC:**
  - integrate out θ's
  - $\log p(Y^{\text{rep}}|\widehat{\psi})$ estimated by $\log p(y|\widehat{\psi})$,
  - penalised by $2k$

- Interested in predicting Y marginally
  - **Bayes Factors:**
  - integrate out θ's and ψ
  - $\log p(Y^{\text{rep}})$ estimated by $\log p(y)$,
  - no penalty



An example

- Suppose the three levels of our model concerned classes within schools within a country
- Then if we were interested in predicting results of future *classes* in those actual schools, then Θ is the focus and deviance-based methods such as DIC are appropriate;
- If we were interested in predicting results of future *schools* in that country, then Ψ is the focus and marginal-likelihood methods such as AIC are appropriate;
- If we were interested in predicting results for a new *country*, then no parameters are in focus and Bayes factors are appropriate to compare models.
- This suggests that Bayes factors may in many circumstances be inappropriate measures by which to compare models