Bayesian Model Choice

Martyn Plummer

International Agency for Research on Cancer
Lyon, France

16 December 2012

The pure approach

For a Bayesian purist, all uncertainty is represented by probability distributions.

- Given a set of candidate models $\mathcal{M}_1 \ldots \mathcal{M}_J$,
- we need to supply prior probabilities $\pi_1 \ldots \pi_J$.

The posterior probabilitity of model $\mathcal{M}_i$ given data Y is

$$p(\mathcal{M}_i \mid Y) = \frac{\pi_i p(Y \mid \mathcal{M}_i)}{\sum_{j=1}^{J} \pi_j p(Y \mid \mathcal{M}_j)}$$

Bayesian Model Averaging

We could select the maximum *a posteriori* model, but:

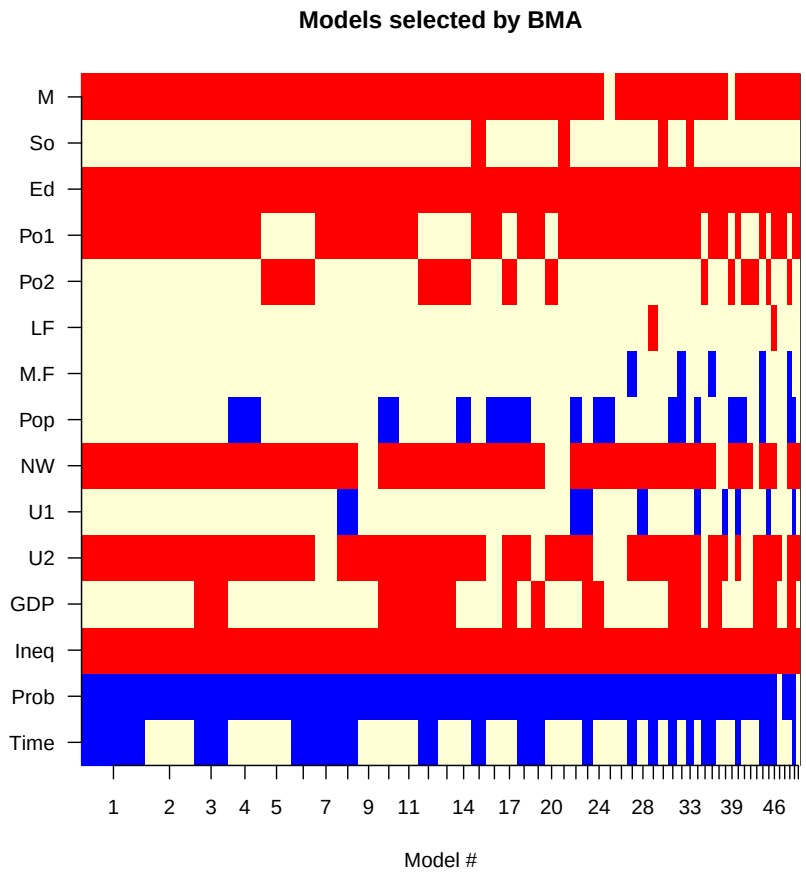
- Further inference would be conditional on the selected model being true, ignoring uncertainty in the choice of model.
- Often $p(\mathcal{M}_i \mid Y) \ll 1$ even for the "best" model.

We may therefore reject model choice in favour of model averaging.

- Keep all candidate models, but down-weight those with small posterior probability.

An example of Bayesian Model Averaging

A variable selection problem: predicting US crime rates



- Variables are on the Y axis
- Models are on the X axis
  - Width $\propto$ posterior probability
  - Sorted in order of posterior probability
- A variable in the model is marked red (positive association) or blue (negative association).

Bayes Factors

Given two candidate models $\mathcal{M}_1$ and $\mathcal{M}_2$

$$\frac{p(\mathcal{M}_1 \mid Y)}{p(\mathcal{M}_2 \mid Y)} = \frac{\pi_1}{\pi_2} \times \frac{p(Y \mid \mathcal{M}_i)}{p(Y \mid \mathcal{M}_2)}$$

$$\text{posterior odds} = \text{prior odds} \times \text{Bayes factor}$$

We can avoid supplying prior probabilities by just using the second factor.

Interpreting Bayes factors

"The Bayes factor is a summary of the evidence provided by the data in favor of one scientific theory by a statistical model, as opposed to another" - Kass and Raftery (1995)

Bayes factor	Interpretation
1 to 3	Not worth more than a bare mention
3 to 20	Positive
20 to 150	Strong
$>$150	Very strong

Models with parameters

- Models typically have parameters that must be estimated.
- For the Bayes factor, parameters are eliminated by integrating them out

$$p(Y \mid \mathcal{M}_i) = \int p(\theta \mid \mathcal{M}_i) p(Y \mid \theta, \mathcal{M}_i) d\theta$$

  to give the marginal likelihood $p(Y \mid \mathcal{M}_i)$.
- Candidate models do not need to share the same parameter space.

Practical problems with Bayes factors

Diffuse reference priors cannot be used for the model parameters:

Lindley-Bartlett paradox

When comparing two nested models with diffuse priors, the Bayes factor may favour the smaller model even when a hypothesis test clearly rejects it.

Philosophical problems with Bayes factors

They require that one of the candidate models is "true".

- The probability model describes the data generating mechanism.
- One of the possible values of the parameters θ describes the true state of nature.
- The prior distribution $p(\theta \mid \mathcal{M})$ gives reasonable prior probability to the true parameter value.

Posterior predictive p-values

Gelman, Meng and Stern (1996)

- Bayesian models describe a data generating process.
- Given the posterior distribution of the parameters, we can simulate a new replicate data set.
- The replicate data set should look like the real data.

Posterior predictive p-values

- Test statistic $T(Y, \theta)$ measures discrepancy between parameters and data
- Replicate data Y^{rep} is conditionally independent of Y given θ
- The p-value

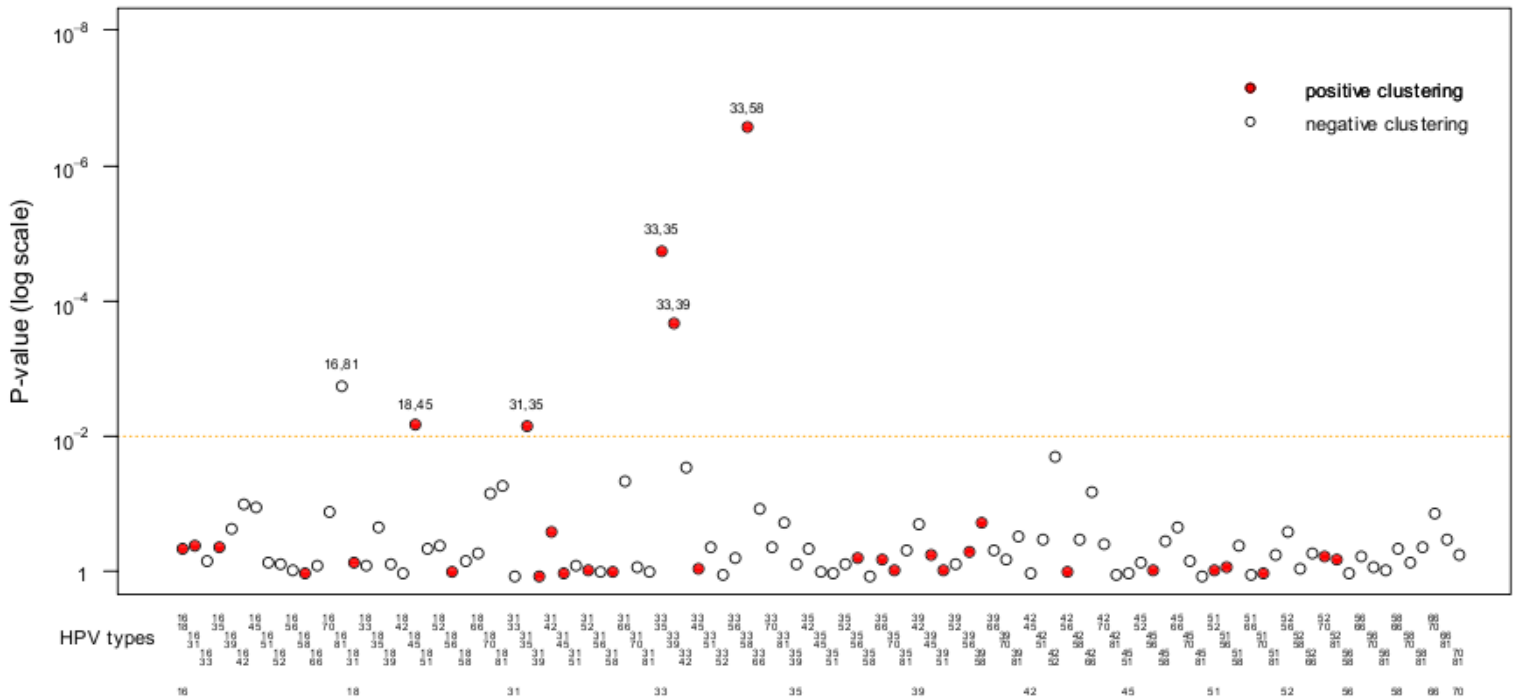
$$P\{T(Y^{rep}, \theta) > T(Y, \theta) \mid Y, \mathcal{M}_i\}$$

can be estimated by iterative simulation by counting the proportion of iterations such that

$$T(Y^{rep}, \theta) > T(Y, \theta)$$

Example of posterior predictive p-values

Using Observed/Expected ratio of multiple HPV infections to find departures from independence.



How many parameters in the model?

Bayesian hierarchical models can vary smoothly between low and high dimensions depending on the strength of the prior.
Consider a set of models indexed by parameter λ

$$\begin{aligned} Y_i \mid \eta_i, \mu &\sim N(\mu + \eta_i, \sigma^2) \\ \eta_i &\sim N(0, \lambda) \\ \mu &\sim N(0, \tau^2) \end{aligned}$$

where σ^2, τ^2 are fixed.

- As $\lambda \downarrow 0$ this tends to a "pooled" model where $Y_1 \ldots Y_n$ have a common mean (μ)
- As $\lambda \uparrow \infty$ we get a "fixed effects" model: learning η_i tells us nothing about η_j for $j \neq i$.

The hat matrix

For $Y \in \mathbb{R}^n, \theta \in \mathbb{R}^p$ and an $n \times p$ matrix of covariates X

$$Y \sim N(X\boldsymbol{\theta}, \sigma^2 I)$$

The fitted value is the expected value Y evaluated at the maximum likelihood estimate $\widehat{\theta}$

$$E\left(Y \mid \widehat{\boldsymbol{\theta}}\right) = HY$$

where

$$H = X(X^T X)^{-1} X^T$$

H is called the "hat matrix".

Properties of the hat matrix

The diagonal elements of the hat matrix $h_i = H_{ii}$ satisfy

$$0 \leq h_i \leq 1$$

$$\sum_{i=1}^{n} h_i = p$$

They are measures of "influence". The larger the value of h_i, the more $\widehat{\theta}$ will change if we drop observation Y_i.

The hat matrix for Bayesian models

If we introduce a prior distribution into the model

$$\theta \sim N(\theta_0, \Psi)$$

then we can still define a Bayesian hat matrix

$$H = X(\sigma^2\Psi + X^T X)^{-1} X^t$$

that satisfies

$$E_{y|\theta}(Y \mid \overline{\theta}) = E_y(Y) + H\ [Y - E_y(Y)]$$

We define the *effective number of parameters* for a linear model as

$$p_D = \sum_{i=1}^{n} h_i$$

The deviance information criterion

Spiegelhalter, Best, Carlin and van der Linde (2002) introduced the deviance information criterion

$$\text{DIC} = \overline{D} + p_D$$

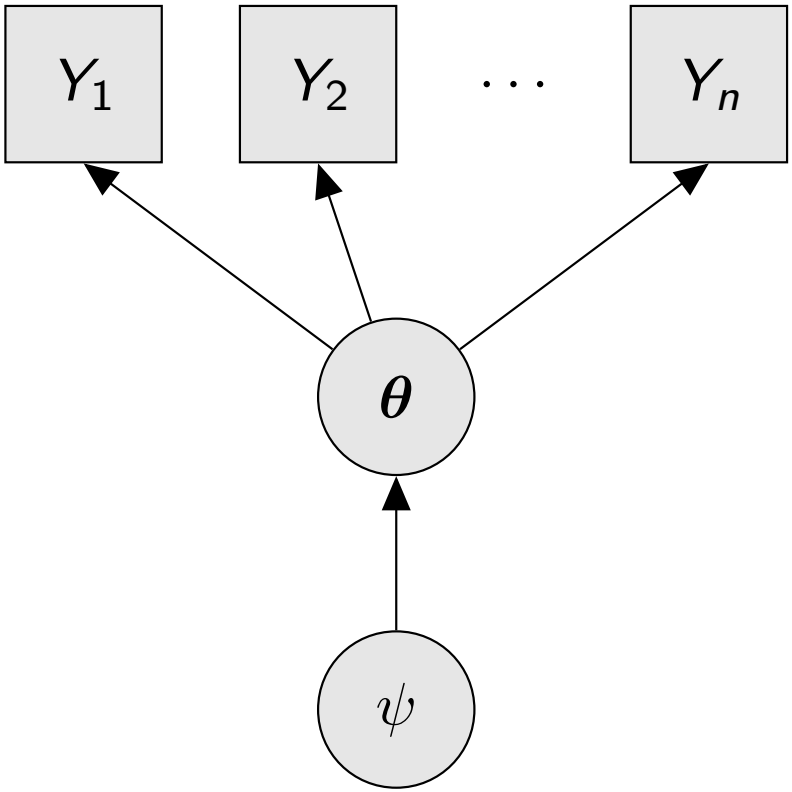
which combines

A measure of fit $\overline{D}$, the expected deviance.

A measure of complexity p_D, the effective number of parameters

Choose the model with the smallest DIC

Notation for DIC



$\mathbf{Y} = (Y_1 \ldots Y_n)$ is a vector of observations

$\boldsymbol{\theta}$ is a vector of parameters common to all models (the focus)

Models differ in the prior structure $p(\boldsymbol{\theta} \mid \psi)$

Spiegelhalter et al definition of p_D

The effective number of parameters

The effective number of parameters in a model was defined by Spiegelhalter *et al* (2002) as

$$p_D = \overline{D} - D\left(\overline{\boldsymbol{\theta}}\right)$$

In general p_D depends on Y.

Spiegelhalter et al definition of p_D

The effective number of parameters

The effective number of parameters in a model was defined by Spiegelhalter *et al* (2002) as

$$p_D = \overline{D} - D\left(\overline{\boldsymbol{\theta}}\right)$$

where

$$\begin{aligned} \overline{\boldsymbol{\theta}} &= \mathrm{E}\left(\boldsymbol{\theta} \mid \mathbf{Y}\right) \\ \overline{D} &= \mathrm{E}\left(D(\boldsymbol{\theta}) \mid \mathbf{Y}\right) \end{aligned}$$

In general p_D depends on Y.

DIC and the Akaike Information Criterion

DIC can also be written:

$$D\left(\overline{\boldsymbol{\theta}}\right) + 2p_D$$

In this form it resembles the classical Akaike Information Criterion (Akaike 1974)

$$D(\widehat{\boldsymbol{\theta}}) + 2p$$

where $\widehat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$ and p is the number of parameters

For non-hierarchical models with a non-informative prior on $\boldsymbol{\theta}$,

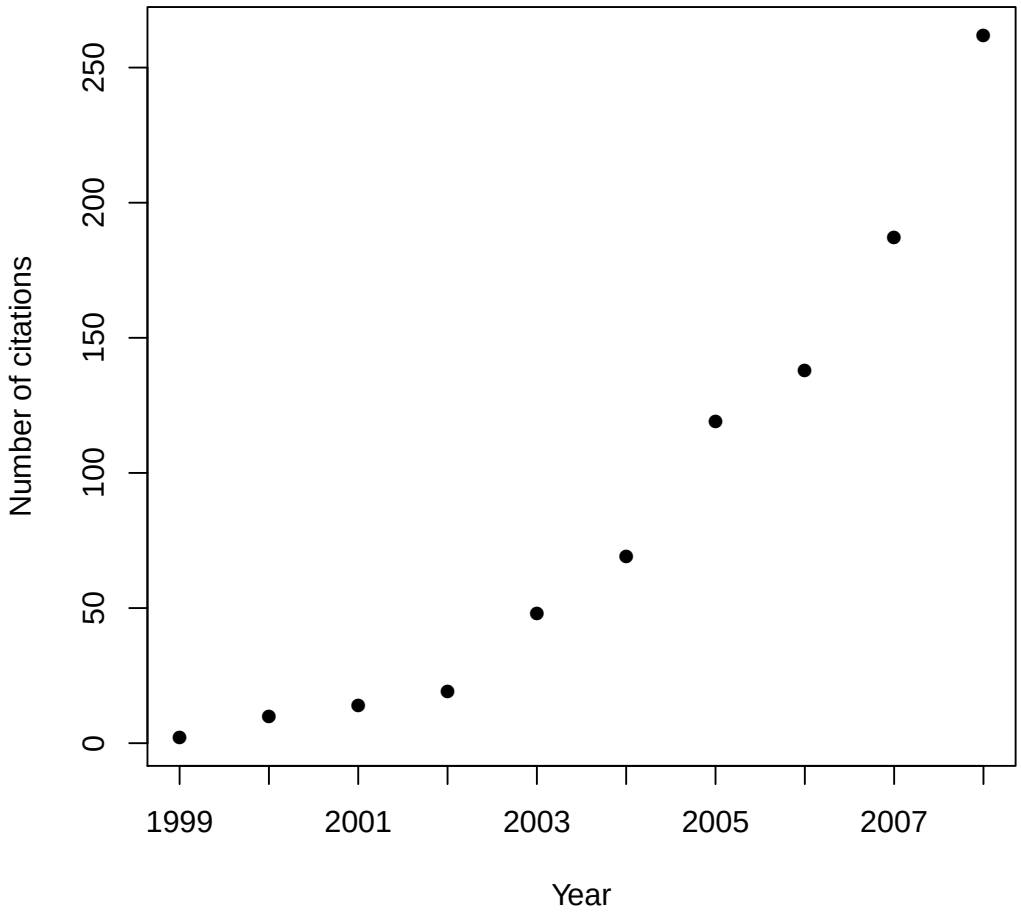
$$\text{DIC} = \text{AIC}$$

Advantages of DIC

- Easy to calculate using Markov Chain Monte Carlo (MCMC) simulation.
- Implemented in WinBUGS/OpenBUGS.
- Widely used and cited
  - Described in text books on Bayesian data analysis
  - 543 citations on ISI database.

Citations of Spiegelhalter *et al* (2002)

from the ISI citation database



Early publications cite the technical report Spiegelhalter, Best and Carlin (1998).

Limitations of DIC

DIC also inherits some of the limitations of AIC.

- Restricted to nested models
  - Difference between AIC of two models is $O_p(n^{1/2})$ in general but $O_p(1)$ for nested models (Ripley 1996)
- Not consistent.
  - Given a set of nested models, DIC will tend to choose a model that is too large as $n \uparrow \infty$.

Problems with p_D

- p_D depends on the parameterization of θ
  - Because it depends on the plug-in estimate $\overline{\boldsymbol{\theta}}$ which changes if we change coordinates
- It cannot be calculated for categorical parameters.
- p_D may be negative
  - Spiegelhalter and colleagues claim this occurs in an ill-fitting model (prior-data conflict).
  - This is an example of Gelman's folk theorem of statistical computing: "When you have computational problems, often there's a problem with your model"

Too many DICs

Many people have suggested alternative penalties that are not p_D but have a similar *prima facie* plausibility

- Plummer (2002) Alternative definition of p_D
- Gelman et al (2004) Use half the posterior variance of the deviance instead of p_D (R2WinBUGS)
- Celeux et al (2006) Eight variations of DIC for mixture models

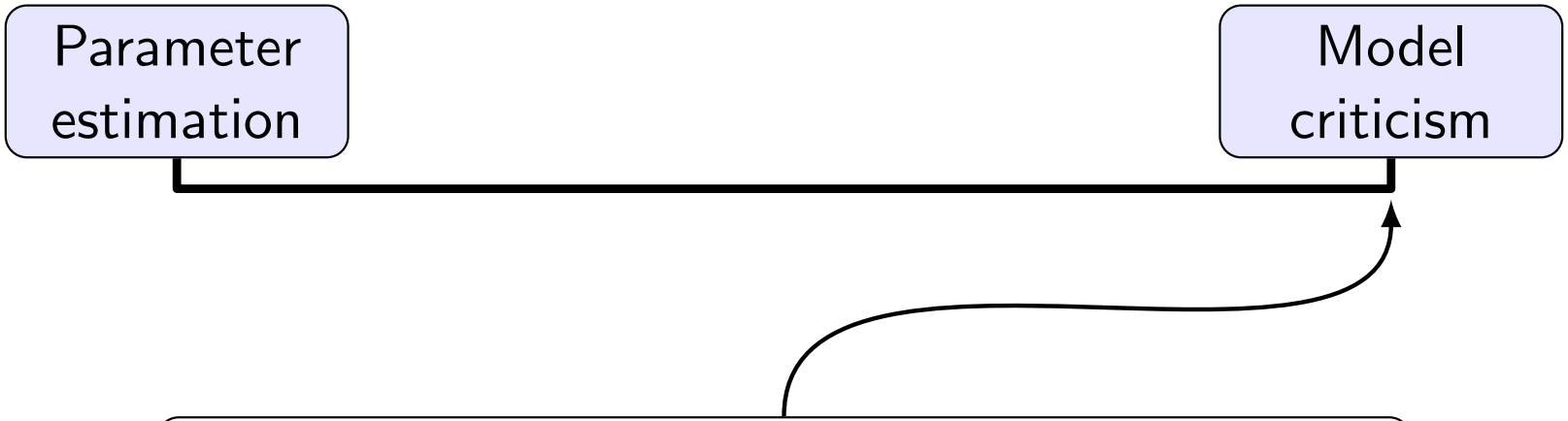
Criticism of DIC

> *The main conclusion of our paper is thus that DIC lacks a natural generalisation outside exponential families or, alternatively, that it happened to work within exponential families while lacking a true theoretical foundation.*

– Celeux, Forbes, Robert, and Titterington (2006)

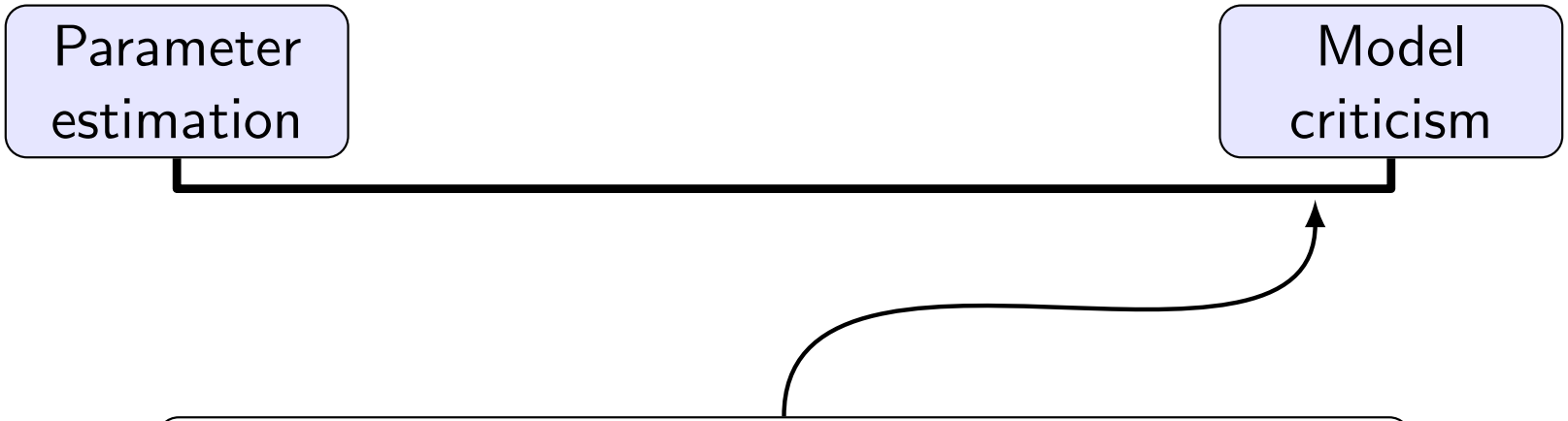
You cannot use the data twice

- The same data cannot be used for both parameter estimation and model criticism.
- Bayesian approaches can be classified by how much of the data they use for each purpose.



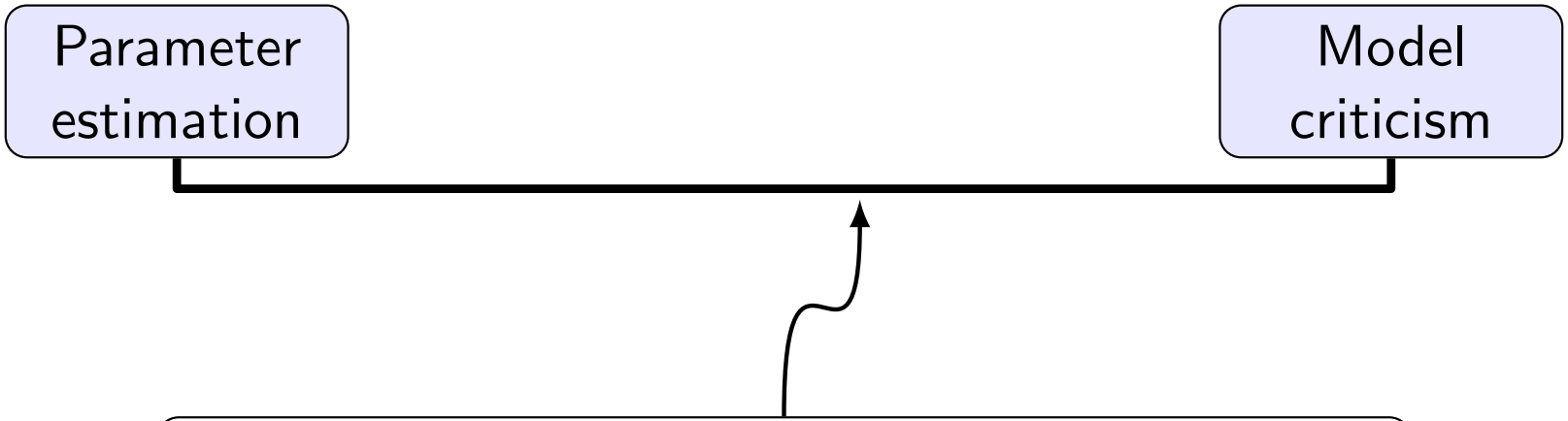
Bayes factors

- Model choice based on marginal likelihood without any attempt to estimate model parameters.
- Incompatible with improper or diffuse reference priors.



Intrinsic and partial Bayes factors

- Designed for use with reference priors.
  - Sacrifice a minimal part of the data to get a proper posterior.
  - Use this as a prior with the rest of the data to calculate a Bayes factor

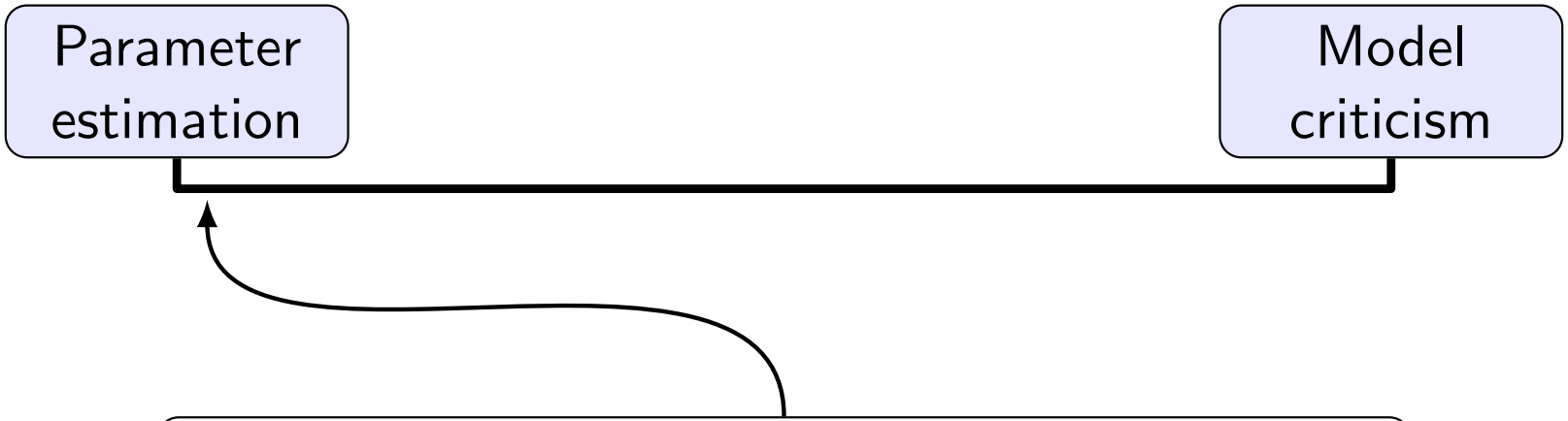


Two-phase studies

- Data collection in two phases
  - Initial data collection used for parameter estimation (hypothesis generating)
  - New data collection for model criticism (hypothesis testing)
- Used by genome-wide association studies to find and then test candidate genes.

k-fold cross-validation

- Randomly split the data $k \approx 10$ groups
  - Use data from $k-1$ groups for parameter estimation.
  - Use remaining group for model criticism.
- Repeat k times and pool results



leave-one-out (loo) cross-validation

- Drop each observation in turn.
  - Use other $n-1$ observations for parameter estimation.
  - Use dropped observation for model criticism.
- More computationally expensive than k-fold cross-validation.

Posterior predictive inference

- Use all data for parameter estimation
- No data left for model criticism ...
  - Simulate replicate data sets from the model!
  - Look for discrepancies between simulated and real data.

Loss functions

We adopt a utilitarian approach to model choice

- A model is considered useful if it gives good out-of-sample predictions.
- Ideally we have two data sets
  - $\mathbf{Z}$, a set of training data
  - $\mathbf{Y}$, a set of validation data
- We measure the utility of a model with a loss function

$$L(\mathbf{Y}, \mathbf{Z})$$

that measures the ability to make good predictions of $\mathbf{Y}$ from $\mathbf{Z}$

Exact replicate loss functions

- In practice, we have only one data set $\mathbf{Y}$.
- It is tempting to use $L(\mathbf{Y}, \mathbf{Y})$ as a utility
  - Call this the exact replicate loss function
  - It uses the data twice (*e.g* the posterior Bayes factor, Aitkin 1991)
- We expect $L(\mathbf{Y}, \mathbf{Y})$ to be conservative.
- If we can quantify *how conservative* $L(\mathbf{Y}, \mathbf{Y})$ is then we can pay a rational price for using the data twice

Linear loss functions

Definition

A loss function is linear if it breaks down into a sum of contributions from each element of the test data $\mathbf{Y}$.

$$L(\mathbf{Y}, \mathbf{Z}) = \sum_i L(Y_i, \mathbf{Z})$$

Loss functions based on the deviance are linear if the elements of $\mathbf{Y}$ are conditionally independent given $\boldsymbol{\theta}$.

Optimism

Definition

The optimism of $L(Y_i, \mathbf{Y})$

$$p_{opt_i} = \mathrm{E}\left\{L(Y_i, \mathbf{Y}_{-i}) - L(Y_i, \mathbf{Y}) \mid \mathbf{Y}_{-i}\right\}$$

Optimism

Definition

The optimism of $L(Y_i, \mathbf{Y})$

$$p_{opt_i} = \mathrm{E}\{L(Y_i, \mathbf{Y}_{-i}) - L(Y_i, \mathbf{Y}) \mid \mathbf{Y}_{-i}\}$$

where

$$\mathbf{Y}_{-i} = (Y_1, \ldots Y_{i-1}, Y_{i+1}, \ldots Y_n)$$

is the data set with observation i removed.

Penalized loss functions

Definition

The penalized loss for observation i

$$L(Y_i, \mathbf{Y}) + p_{opt_i}$$

has the same expectation (given $\mathbf{Y}_{-i}$) as the cross-validation loss

$$L(Y_i, \mathbf{Y}_{-i})$$

Definition

Sum over the penalized losses to get the total penalized loss

$$L(\mathbf{Y}, \mathbf{Y}) + p_{opt}$$

The plug-in deviance

- The plug-in deviance

$$L^p(\mathbf{Y}, \mathbf{Z}) = -2 \log \left[p \left\{ \mathbf{Y} \mid \bar{\boldsymbol{\theta}}(\mathbf{Z}) \right\} \right]$$

  is a linear loss-function based on the deviance.
- It depends only on the posterior expectation of $\boldsymbol{\theta}$

$$\bar{\boldsymbol{\theta}}(\mathbf{Z}) = \mathrm{E}(\boldsymbol{\theta} \mid \mathbf{Z})$$

The penalized plug-in deviance

- For a linear model (with known variance)

$$L^p(\mathbf{Y}, \mathbf{Y}) + p_{opt} = \overline{D} + \sum_i p_{D_i}/(1 - p_{D_i})$$

  where p_{D_i} is the contribution of observation i to p_D.
- This formula is asymptotically correct for generalized linear mixed models (with canonical link).

Plug-in deviance and DIC

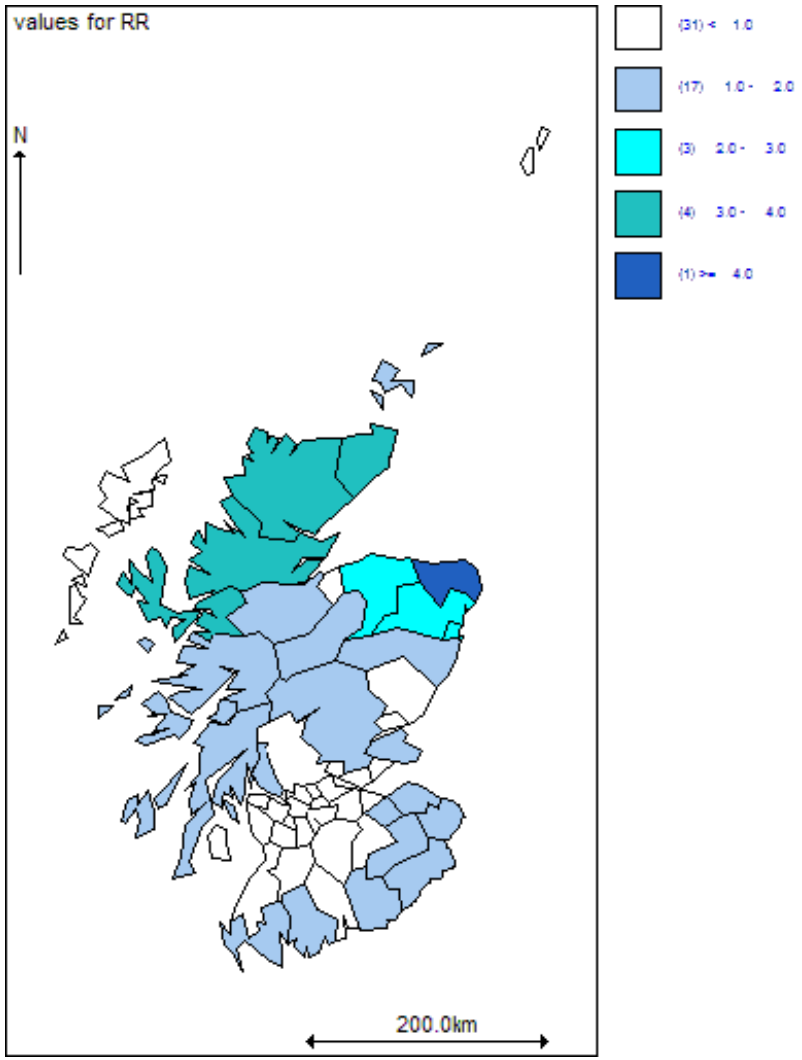
When there are no influential observations $p_{D_i} \ll 1$ and

$$\sum_i p_{D_i}/(1 - p_{D_i}) \approx \sum_i p_{D_i} = p_D$$

and DIC is an approximation to the penalized plug-in deviance.

A necessary condition is $p_D \ll n$.

Example: Lip Cancer in Scotland



- A classic problem in disease mapping
- We want an accurate representation of spatial variation in a rare disease (lip cancer)
- But we want to ignore random fluctuations due to small disease counts

Model for the lip cancer data

We use a generalized linear mixed model (GLMM) with Poisson family and log link

$$\log\{\mathrm{E}(Y_i)\} = \alpha_0 + \gamma_i + \delta_i + \log(E_i)$$

where

- $\gamma_1 \ldots \gamma_n$ are unstructured random effects
- $\delta_1 \ldots \delta_n$ have a conditional autoregressive prior
- $E_1 \ldots E_n$ are expected numbers of cases based on population structure

Which effects to include for an optimal disease map?

DIC is a poor approximation for the lip cancer data

- The effective number of parameters (p_D) is close to the number of independent observations ($n = 56$).
- p_D is a poor approximation to the correct penalty for $\overline{D}$.

Model	p_D	Correct penalty
Pooled	1.0	1.1
Exchangeable	43.5	570.5
Spatial	31.0	163.9
Exchangeable + spatial	31.6	166.4

The expected deviance

- The expected deviance

$$L^e(\mathbf{Y}, \mathbf{Z}) = -2 \int d\boldsymbol{\theta}\; p(\boldsymbol{\theta} \mid \mathbf{Z}) \log \{p(\mathbf{Y} \mid \boldsymbol{\theta})\}$$

  is a linear loss function based on the deviance.
- In its exact replicate form

$$L^e(\mathbf{Y}, \mathbf{Y}) = \overline{D}$$

The penalized expected deviance

- For a linear model (with known variance)

$$L^e(\mathbf{Y}, \mathbf{Y}) + p_{opt} = \overline{D} + 2\sum_i p_{D_i}/(1 - p_{D_i})$$

- Similar to the plug-in deviance but with a penalty twice the size.

Exponential family models

In exponential family models, the penalized expected deviance is

$$\overline{D} + 2\varphi^{-1} \sum_{i=1}^{n} \text{Cov}\left(\theta_i, \mu_i \mid \mathbf{Y}_{-i}\right)$$

where

- θ_i is the canonical parameter
- μ_i is the mean value parameter
- φ is the scale parameter

Estimation of p_{opt} in general models

If $J(\theta_2, \theta_2)$ is the undirected information divergence between the predictive density of $\mathbf{Y}$ at $\theta = \theta_1$ and the density at $\theta = \theta_2$, then

$$p_{opt_i} = \int d\boldsymbol{\theta} \int d\boldsymbol{\theta}' \; p\left(\boldsymbol{\theta} \mid \mathbf{Y}_{-i}\right) p\left(\boldsymbol{\theta}' \mid \mathbf{Y}_{-i}\right) J_i(\boldsymbol{\theta}, \boldsymbol{\theta}')$$

Estimation of p_{opt} in general models

If $J(\theta_2, \theta_2)$ is the undirected information divergence between the predictive density of $\mathbf{Y}$ at $\theta = \theta_1$ and the density at $\theta = \theta_2$, then

$$p_{opt_i} = \int d\boldsymbol{\theta} \int d\boldsymbol{\theta}' \; p\left(\boldsymbol{\theta} \mid \mathbf{Y}_{-i}\right) p\left(\boldsymbol{\theta}' \mid \mathbf{Y}_{-i}\right) J_i(\boldsymbol{\theta}, \boldsymbol{\theta}')$$

Approximation of p_{opt}

In the absence of influential observations

$$p_{opt_i} \approx \int d\boldsymbol{\theta} \int d\boldsymbol{\theta}' \; p\left(\boldsymbol{\theta} \mid \mathbf{Y}\right) p\left(\boldsymbol{\theta}' \mid \mathbf{Y}\right) J_i(\boldsymbol{\theta}, \boldsymbol{\theta}')$$

which may be estimated using two parallel chains.
In this case $p_{opt} \approx 2p_D^*$ where p_D^* is the "effective number of parameters" proposed by Plummer (2002) and the penalized expected deviance is:

$$\overline{D} + 2p_D^*$$

Necessary conditions for DIC

- DIC can be justified as an approximation to the penalized plug-in deviance (in regular models).
- But there are necessary conditions attached
  - Conditional independence of $Y_1 \ldots Y_n$ given θ
  - No influential observations ($p_D \ll n$)

  The latter may not be satisfied by models with individual-level random effects.

Comments and questions

- In regular models p_D^* is a useful summary of the model dimension and also provides influence diagnostics.
- Is it useful in singular models?
- Is the condition $p_D \ll n$ also necessary for WAIC.
- We still do not have a good threshold for a "significant" difference in XIC.
- WAIC should be easy to implement in JAGS (Just Another Gibbs Sampler) `http://mcmc-jags.sourceforge.net`.