

▪

# Objective Bayesian testing and Model Selection

**M.J. Bayarri and James O. Berger**

Universitat de València, SAMSI and Duke University

*O-Bayes09, Philadelphia, June 5, 2009*

# Outline

## Testing

- Basics of Bayesian hypothesis testing
- $p$ -values are not good measures of evidence

## Model selection/uncertainty

- Basics of Bayesian model selection
- Motivation for the Bayesian approach
- Multiple Testing

## Methodologies

- Conventional priors for Linear Models
- Asymptotics
- Training sample methods

## A simple example of objective testing

**Data and model:**  $\mathbf{X}$  has density  $f(\mathbf{x} \mid \theta)$

$X = \#$  eggs hatched out of  $n$  eggs in a recently polluted area ( so  $f$  is binomial, and  $\theta$  is the true proportion that would hatch).

Suppose  $x = 40$  eggs hatched out of  $n = 100$

**To test:**  $H_0 : \theta = \theta_0$  vs  $H_1 : \theta \neq \theta_0$

**Example:**  $\theta_0$  is the historically known proportion of eggs that hatch in the area. For  $\theta_0 = .5$ , the p-value = 0.05

## The prior distribution

Let  $Pr(H_i)$  = prior probability that  $H_i$  is true,  $i = 0, 1$

On  $H_1 : \theta \neq \theta_0$ , let  $\pi(\theta)$  be the prior density for  $\theta$ .

**Subjective Bayes:** choose the  $Pr(H_i)$  and  $\pi(\theta)$  based on personal beliefs

**Objective (or default) Bayes:** choose

$$Pr(H_0) = Pr(H_1) = \frac{1}{2}$$

$$\pi(\theta) = 1 \quad (\text{on } 0 < \theta < 1)$$

## Posterior probability of hypotheses:

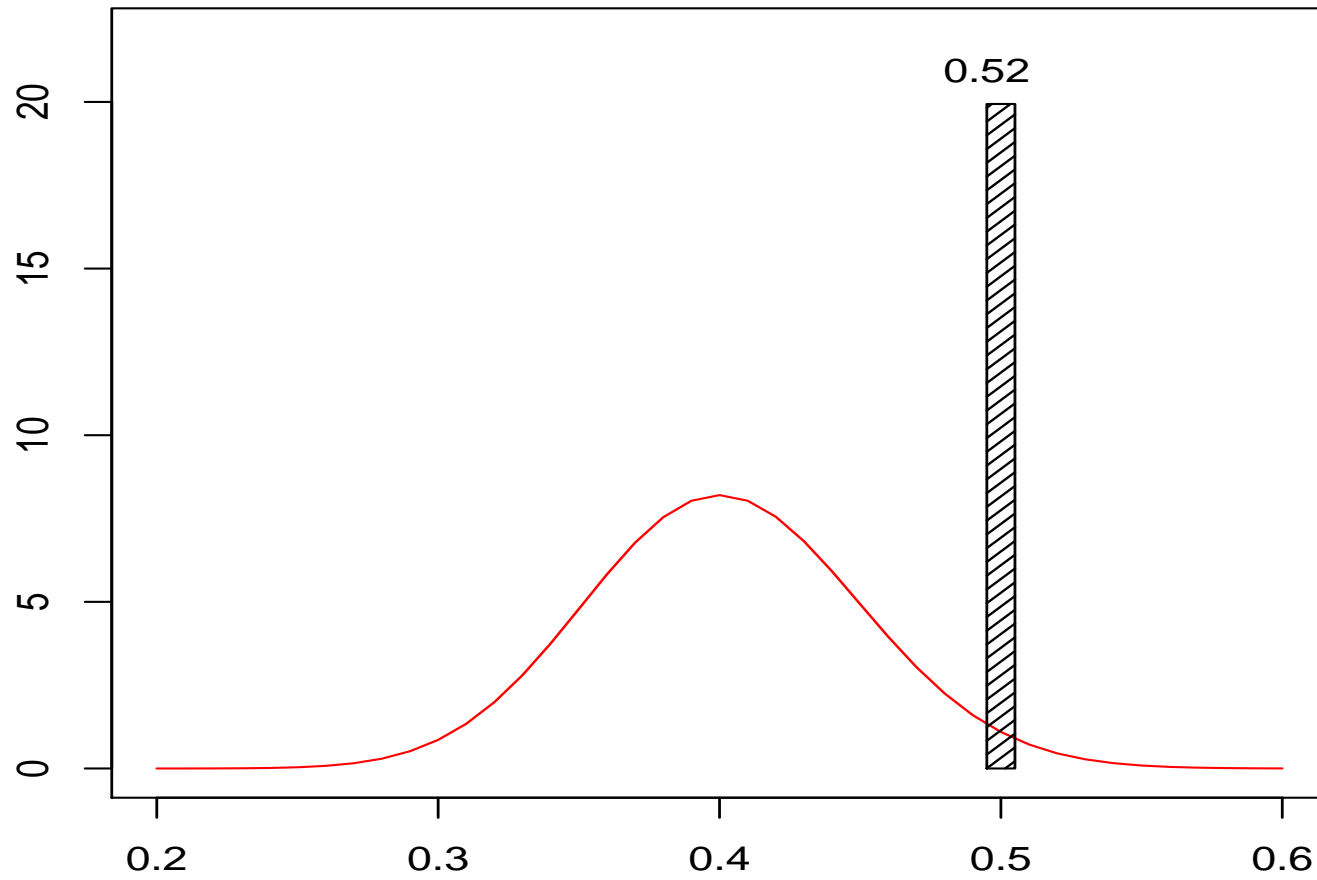
$$\begin{aligned} Pr(H_0 | x) &= \text{probability that } H_0 \text{ true, given data, } x \\ &= \frac{Pr(H_0)f(x|\theta_0)}{Pr(H_0)f(x|\theta_0) + Pr(H_1) \int f(x|\theta)\pi(\theta)d\theta} \end{aligned}$$

For the objective prior,  $Pr(H_0 | x = 40) \approx 0.52$   
(recall, p-value  $\approx .05$ )

Posterior density on  $H_1 : \theta \neq \theta_0$  is

$$\pi(\theta | x, H_1) \propto \pi(\theta)f(x | \theta) \propto 1 \times \theta^x(1 - \theta)^{n-x},$$

the  $Be(\theta | 41, 61)$  density.



## Bayes Factor:

$$\begin{aligned}
 B_{01} &= \frac{f(x|\theta_0)}{\int_{\theta \neq \theta_0} f(x|\theta)\pi(\theta)d\theta} = \frac{f(x|\theta_0)}{\int_{\theta \neq \theta_0} f(x|\theta)d\theta} \\
 &= \frac{\text{likelihood of data under } H_0}{\text{“average” likelihood of data under } H_1}
 \end{aligned}$$

Often used in ‘objective’ analyses since this does not involve prior probabilities of the  $H_i$ .

Note: 
$$\frac{\Pr(H_0|x)}{\Pr(H_1|x)} = \frac{\Pr(H_0)}{\Pr(H_1)} \times B_{01}$$
  
 (posterior odds)                      (prior odds)                      (Bayes factor)

so  $B_{01}$  is often thought of as “the odds of  $H_0$  to  $H_1$  provided by the data”. In the example  $B_{01} = 1.1$

## \*\* How to formulate hypotheses when using Bayesian Testing

**Most Crucial:** The hypotheses must all be “plausible” (i.e., have reasonable prior probability of being true). If a hypothesis has 0 prior probability, its posterior probability is also 0.

**Typical example:** Compare Treatments A and B.

$\theta$  = mean treatment difference

$$H_0 : \theta \approx 0 \quad \text{versus} \quad \theta \neq 0$$

Is  $H_0$  plausible?

**Scenario 1:**

Treatment A = standard chemotherapy

Treatment B = standard chemotherapy + steroids

$\theta \approx 0$  is plausible

**Scenario 2:**

Treatment A = standard chemotherapy

Treatment B = a new radiation therapy

$\theta \approx 0$  is NOT plausible,

better to test  $H_0 : \theta < 0$  vs.  $H_1 : \theta > 0$

**Note:** In Scenario 1, one could test

$H_0 : \theta = 0$  vs.  $H_1 : \theta < 0$  vs.  $H_2 : \theta > 0$ ,

but  $H_0 : \theta = 0$  should be there

## \*\* Approximating a believable null hypothesis by a precise null

A precise null, like  $H_0 : \theta = \theta_0$ , is typically never true *exactly*; rather, it is used as a surrogate for a ‘real null’

$$H_0^\epsilon : |\theta - \theta_0| < \epsilon, \quad \epsilon \text{ small.}$$

**Result** (Berger and Delampady, 1989):

if  $\epsilon < \frac{1}{4} \sigma_{\hat{\theta}}$ , where  $\sigma_{\hat{\theta}}$  is the standard error of the estimate of  $\theta$ , then

$$Pr(H_0^\epsilon | \mathbf{x}) \approx Pr(H_0 | \mathbf{x}).$$

**Note:** Typically,  $\sigma_{\hat{\theta}} \approx \frac{c}{\sqrt{n}}$ , so for large  $n$  the above condition can be violated, and using a precise null may not be appropriate, even if the real null is believable.

## Clash between $p$ -values and Bayes answers

In the example,  $p$ -value  $\approx .05$ , but  
the objective posterior probability of the null  $\approx 0.52$ ;  
( Bayes factor gives  $\approx 1$  to 1 odds in favor of the null).

Was the prior on  $\theta$  inappropriate? (It seems fair to give both hypotheses the same prior probability)

- But it was a neutral, *objective* prior.
- *Any* prior produces Bayes factors orders of magnitude larger than the  $p$ -value.
- Conditional frequentist analysis agrees with the Bayesian answers (and disagrees with the  $p$ -value). For an illustrative *applet* and papers, see [www.stat.duke.edu/~berger](http://www.stat.duke.edu/~berger)

$p$ -values are not good measures of evidence

## AN EXAMPLE

- Experimental drugs  $D_1, D_2, D_3, \dots \rightsquigarrow$  are to be tested (same illness or different illnesses, independents)

- For each, test

$H_1 : D_i$  has negligible effect vs  $H_2 : D_i$  is effective

- Results of the Tests of the Drugs:

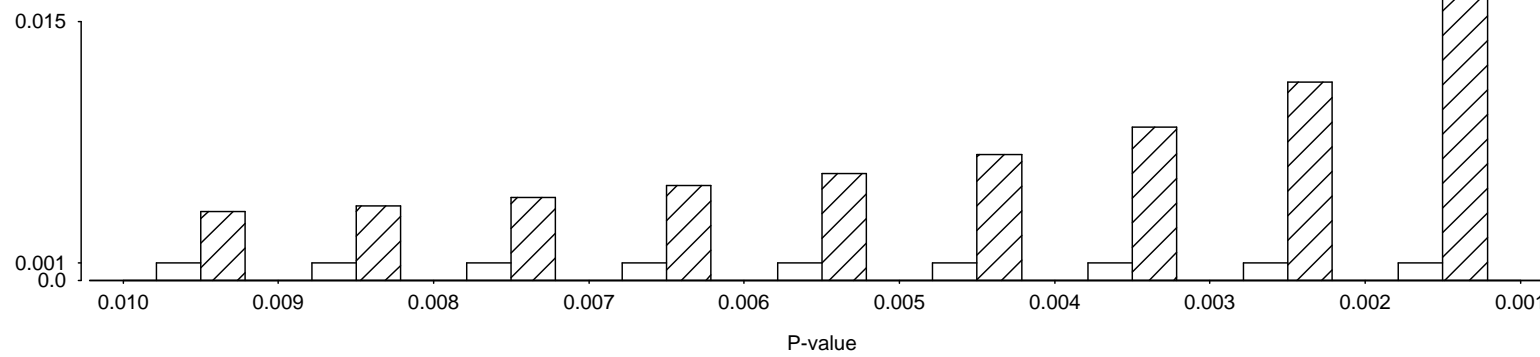
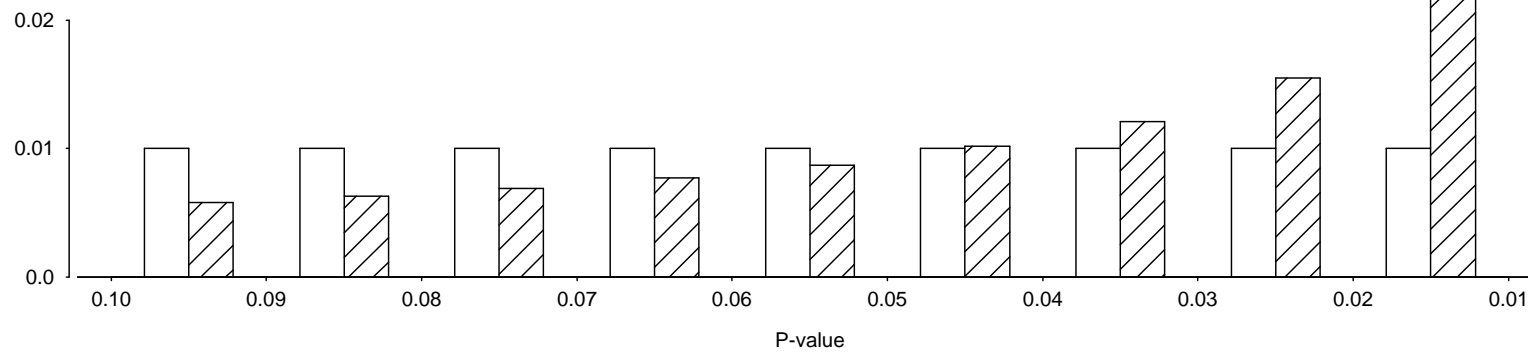
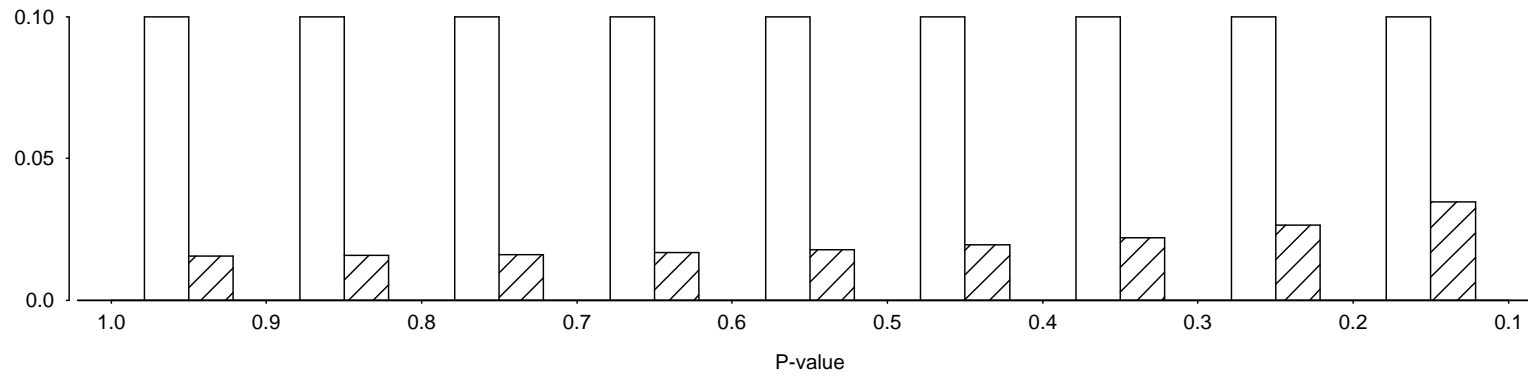
|           |      |             |      |              |             |      |
|-----------|------|-------------|------|--------------|-------------|------|
| TREATMENT | D1   | D2          | D3   | D4           | D5          | D6   |
| P-VALUE   | 0.41 | <b>0.04</b> | 0.32 | 0.94         | <b>0.01</b> | 0.28 |
|           |      |             |      |              |             |      |
| TREATMENT | D7   | D8          | D9   | D10          | D11         | D12  |
| P-VALOR   | 0.11 | <b>0.05</b> | 0.65 | <b>0.009</b> | 0.09        | 0.66 |

- Question: How strongly do we believe that  $D_i$  has a non negligible effect when  $p$ -value  $\approx .05$

## An interesting simulation for normal data

- Generate data sets from  $H_0$  and compute the  $p$ -values
- Generate data sets from  $H_1$ , and compute the  $p$ -values
- Under  $H_0$ , the  $p$ -values  $\sim$  uniform on  $(0, 1)$
- Under  $H_1$ , “random data” might arise from either:
  - (i) Picking a  $\theta \neq 0$  and generating the data
  - (ii) Picking a sequence of  $\theta$ 's and generating the data
  - (iii) Picking a distribution  $\pi(\theta)$  for  $\theta$ ; generating  $\theta$ 's from  $\pi$ ; and then generating data from these  $\theta$ 's

**Example:** taking  $\pi(\theta)$  a  $N(0, 2)$  and  $n = 20$ , yields the following fraction of  $p$ -values in each interval



**Surprise** : No matter *how* one generates “random”  $p$ -values under  $H_1$ , *at most* 23% will fall in the interval  $(.04, .05)$ , so a  $p$ -value near .05 (i.e.,  $|t|$  near 1.96) provides *at most* 3.4 to 1 odds in favor of  $H_1$

(Berger and Sellke, J. Amer.Stat. Assoc., 1987)

**Message** : Knowing that data is “rare” under  $H_0$  is of little use unless one determines whether or not it is also “rare” under  $H_1$ .

**In practice** : For moderate or large sample sizes, data under  $H_1$ , for which the  $p$ -value is between 0.04 and 0.05 is typically as rare or more rare than data under  $H_0$ , so that odds of about 1 to 1 are then reasonable.

## A partial fix: calibration of $p$ -values

Robust Bayesian theory suggests a way to calibrate  $p$ -values, when no alternative hypothesis is specified.

(Sellke, Bayarri and Berger, 2001).

- A *proper*  $p$ -value satisfies  $H_0 : p(X) \sim \text{Uniform}(0, 1)$ .
- Consider testing this versus  $H_1$ , a reasonable nonparametric alternative for  $p(X)$ .
- Then if  $p < e^{-1}$ ,  $B_{01} \geq -e p \log(p)$ .

so, for a reported  $p$ , compute the lower bound above and interpret as the odds (Bayes factor) of  $H_0$  to  $H_1$

- An analogous *lower bound* on conditional Type I frequentist error is  $\alpha \geq (1 + [-e p \log(p)]^{-1})^{-1}$ .  
so, for a reported  $p$ , compute the lower bound above and interpret as the posterior probability that  $H_0$  is true, or as the (conditional) frequentist Type 1 error probability

**Previous example:** For the hatched eggs example,  
 $p$ -value = 0.05,  $B_{01} = 1.1$ ,  $\Pr(H_0 | x) = 0.52$ .

- $B(p) \geq -e p \log(p) = -e (.05) \log(.05) = 1/(2.46)$ ,  
so at most odds of 2.46 to 1 against  $H_0$
- $\alpha(p) \geq (1 + [-e (.05) \log(.05)]^{-1})^{-1} = .29$ ,  
so the actual error rate in rejecting  $H_0$  is *at least* .29

## Conclusions 1

- $Pr(H_0 | x)$  or  $B$  can be *much larger* than a corresponding p-value.
- Bayes factors are useful for communicating “what the data says,” separate from the  $Pr(H_i)$ .
- When testing, one must think carefully about whether a ‘precise null’ is believable.
- One cannot generally test from confidence intervals.

## Notation for general model selection

**Models** (or hypotheses). Data,  $\mathbf{x}$ , is assumed to have arisen from one of several models:

$M_1$ :  $\mathbf{X}$  has density  $f_1(\mathbf{x} \mid \boldsymbol{\theta}_1)$

$M_2$ :  $\mathbf{X}$  has density  $f_2(\mathbf{x} \mid \boldsymbol{\theta}_2)$

...

$M_q$ :  $\mathbf{X}$  has density  $f_q(\mathbf{x} \mid \boldsymbol{\theta}_q)$

**Assign** prior probabilities,  $P(M_i)$ , to each model

- for small  $q$  usually,  $P(M_i) = \frac{1}{q}$
- having decreasing  $P(M_i)$ , as model complexity increases, is often reasonable
- if  $q$  is huge and multiplicity is an issue, careful quantification of  $P(M_i)$  is required (more later)

**Under model  $M_i$  :**

**Prior** density of  $\boldsymbol{\theta}_i$ :  $\pi_i(\boldsymbol{\theta}_i)$

**Marginal** density of  $\mathbf{X}$ :  $m_i(\mathbf{x}) = \int f_i(\mathbf{x}|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i$   
which measures “how likely is  $\mathbf{x}$  under  $M_i$ ”

**Posterior** density:  $\pi_i(\boldsymbol{\theta}_i|\mathbf{x}) = f_i(\mathbf{x}|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)/m_i(\mathbf{x})$

**Bayes factor** of  $M_j$  to  $M_i$ :  $B_{ji} = m_j(\mathbf{x})/m_i(\mathbf{x})$

**Posterior probability** of  $M_i$ :

$$P(M_i | \mathbf{x}) = \frac{P(M_i)m_i(\mathbf{x})}{\sum_{j=1}^q P(M_j)m_j(\mathbf{x})} = \left[ \sum_{j=1}^q \frac{P(M_j)}{P(M_i)} B_{ji} \right]^{-1}$$

**Particular case** :  $P(M_j) = 1/q$  :

$$P(M_i | \mathbf{x}) = \bar{m}_i(\mathbf{x}) = \frac{m_i(\mathbf{x})}{\sum_{j=1}^q m_j(\mathbf{x})} = \frac{1}{\sum_{j=1}^q B_{ji}}$$

**Reporting** : It is useful to separately report  $\{\bar{m}_i(\mathbf{x})\}$  and  $\{P(M_i)\}$ . Knowing the  $\bar{m}_i$  allows computation of the

$$P(M_i | \mathbf{x}) = \frac{P(M_i) \bar{m}_i(\mathbf{x})}{\sum_{j=1}^q P(M_j) \bar{m}_j(\mathbf{x})}$$

for any prior probabilities.

We mainly focus on OBayes derivation of  $B_{ij}$

## Example: location-scale (Berger & Pericchi)

Suppose  $X_1, X_2, \dots, X_n$  are i.i.d with density

$$f(x_i|\mu, \sigma) = \frac{1}{\sigma} g\left(\frac{x_i - \mu}{\sigma}\right)$$

**Several models** are entertained:

$M_N$ :  $g$  is  $N(0, 1)$

$M_U$ :  $g$  is Uniform  $(0, 1)$

$M_C$ :  $g$  is Cauchy  $(0, 1)$

$M_L$ :  $g$  is Left Exponential  $(\frac{1}{\sigma} e^{(x-\mu)/\sigma}, x \leq \mu)$

$M_R$ :  $g$  is Right Exponential  $(\frac{1}{\sigma} e^{-(x-\mu)/\sigma}, x \geq \mu)$

**Difficulty:** Since these models are not nested, and since there are no low dimensional sufficient statistics for all models, classical model selection is difficult and would typically rely on asymptotics (usually requiring large sample sizes)

**Prior distribution:** choose, for  $i = 1, \dots, 5$ ,

$$P(M_i) = \frac{1}{q} = \frac{1}{5};$$

the objective prior  $\pi_i(\boldsymbol{\theta}_i) = \pi_i(\mu, \sigma) = \frac{1}{\sigma}$  .

**Marginal distributions,**  $m^N(\mathbf{x}|M)$ , for these models can then be calculated in closed form.

Here  $m^N(\mathbf{x} | M) = \int \int \prod_{i=1}^n \left[ \frac{1}{\sigma} g\left(\frac{x_i - \mu}{\sigma}\right) \right] \frac{1}{\sigma} d\mu d\sigma$ . For the five models, the marginals are

$$1. \text{ Normal: } m^N(\mathbf{x} | M_N) = \frac{\Gamma((n-1)/2)}{(2\pi)^{(n-1)/2} \sqrt{n} (\sum_i (x_i - \bar{x})^2)^{(n-1)/2}}$$

$$2. \text{ Uniform: } m^N(\mathbf{x} | M_U) = \frac{1}{n(n-1)(x_{(n)} - x_{(1)})^{n-1}}$$

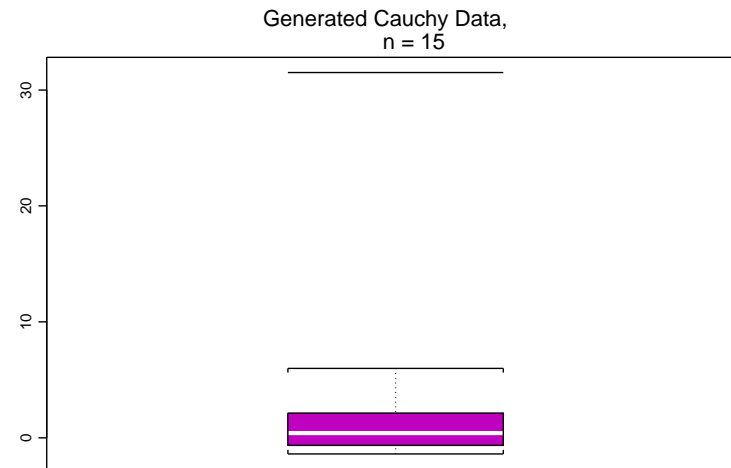
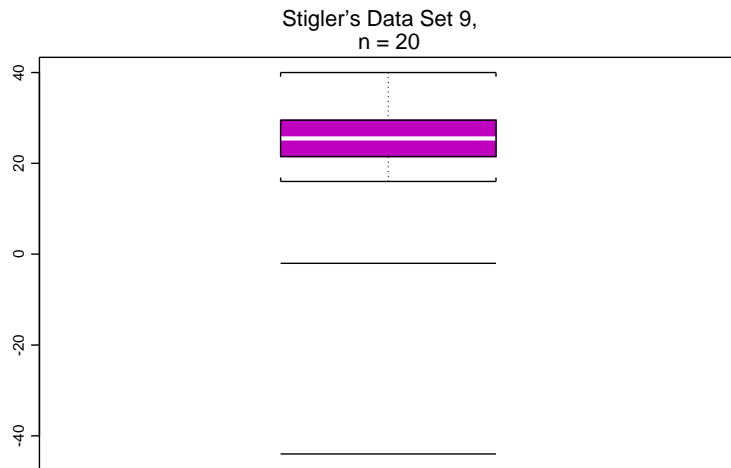
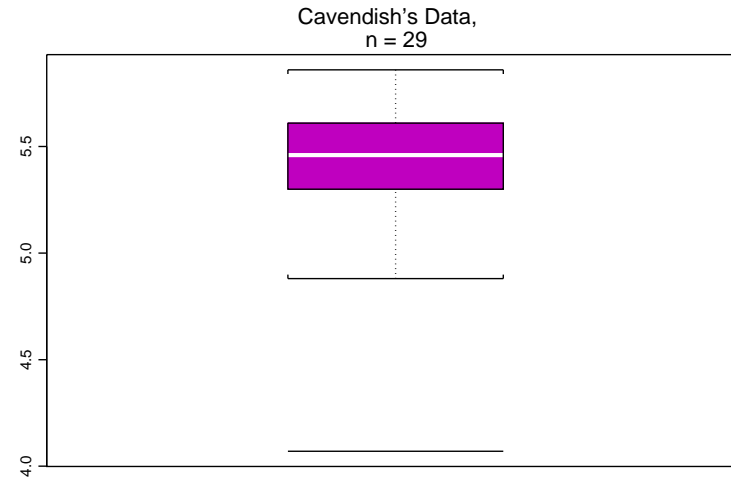
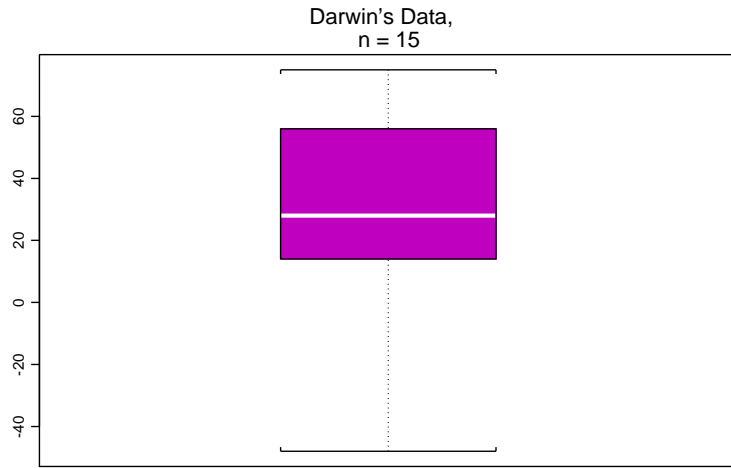
3. Cauchy:  $m^N(\mathbf{x} | M_C)$  is given in Spiegelhalter (1985).

$$4. \text{ Left Exponential: } m^N(\mathbf{x} | M_L) = \frac{(n-2)!}{n^n (x_{(n)} - \bar{x})^{n-1}}$$

$$5. \text{ Right Exponential: } m^N(\mathbf{x} | M_R) = \frac{(n-2)!}{n^n (\bar{x} - x_{(1)})^{n-1}}$$

## Consider four classic data sets:

- Darwin's data ( $n = 15$ )
- Cavendish's data ( $n = 29$ )
- Stigler's Data Set 9 ( $n = 20$ )
- A randomly generated Cauchy sample ( $n = 14$ )



The objective posterior probabilities of the five models, for each data set,  $\bar{m}^N(\boldsymbol{x})$ , are as follows:

| DATA<br>SET | MODELS              |                     |        |                     |                     |
|-------------|---------------------|---------------------|--------|---------------------|---------------------|
|             | Normal              | Uniform             | Cauchy | L. Exp.             | R. Exp.             |
| Darwin      | .390                | .056                | .430   | .124                | .0001               |
| Cavendish   | .986                | .010                | .004   | $4 \times 10^{-8}$  | .0006               |
| Stigler 9   | $7 \times 10^{-8}$  | $4 \times 10^{-5}$  | .994   | .006                | $2 \times 10^{-13}$ |
| Cauchy      | $5 \times 10^{-13}$ | $9 \times 10^{-12}$ | .9999  | $7 \times 10^{-18}$ | $1 \times 10^{-4}$  |

## Motivation for the Bayesian approach

**Ease of interpretation:**  $Pr(H_0 | \mathbf{x})$  reflects real expected error rates (and  $p$ -values do not).

**Prior information can be incorporated**, if desired. Indeed, if the published default posterior probability of  $H_0$  is  $\bar{m}_0$ , and *your* prior probability of  $H_0$  is  $\pi_0$ , then *your* posterior probability of  $H_0$  is

$$Pr(H_0 | \text{data } x) = \left[ 1 + \left( \frac{1}{\pi_0} - 1 \right) \left( \frac{1}{\bar{m}_0} - 1 \right) \right]^{-1}$$

*Example:* In the hatched eggs example,  $\bar{m}_0 = 0.52$ .

“ecologist” has  $\pi_0 = .1 \rightsquigarrow Pr(H_0 | \text{data } x) = 0.11$

“politician” has  $\pi_0 = .9 \rightsquigarrow Pr(H_0 | \text{data } x) = 0.91$

## Consistency:

- A. If one of the  $M_i$  is true, then  $\bar{m}_i \rightarrow 1$  as  $n \rightarrow \infty$
- B. If some other model,  $M^*$ , is true,  $\bar{m}_i \rightarrow 1$  for that model closest to  $M^*$  in Kullback-Leibler divergence

(Berk, 1966, Dmochowski, 1995)

**Ockham's razor:** attributed to thirteen-century Franciscan monk William of Ockham (Occam in latin)

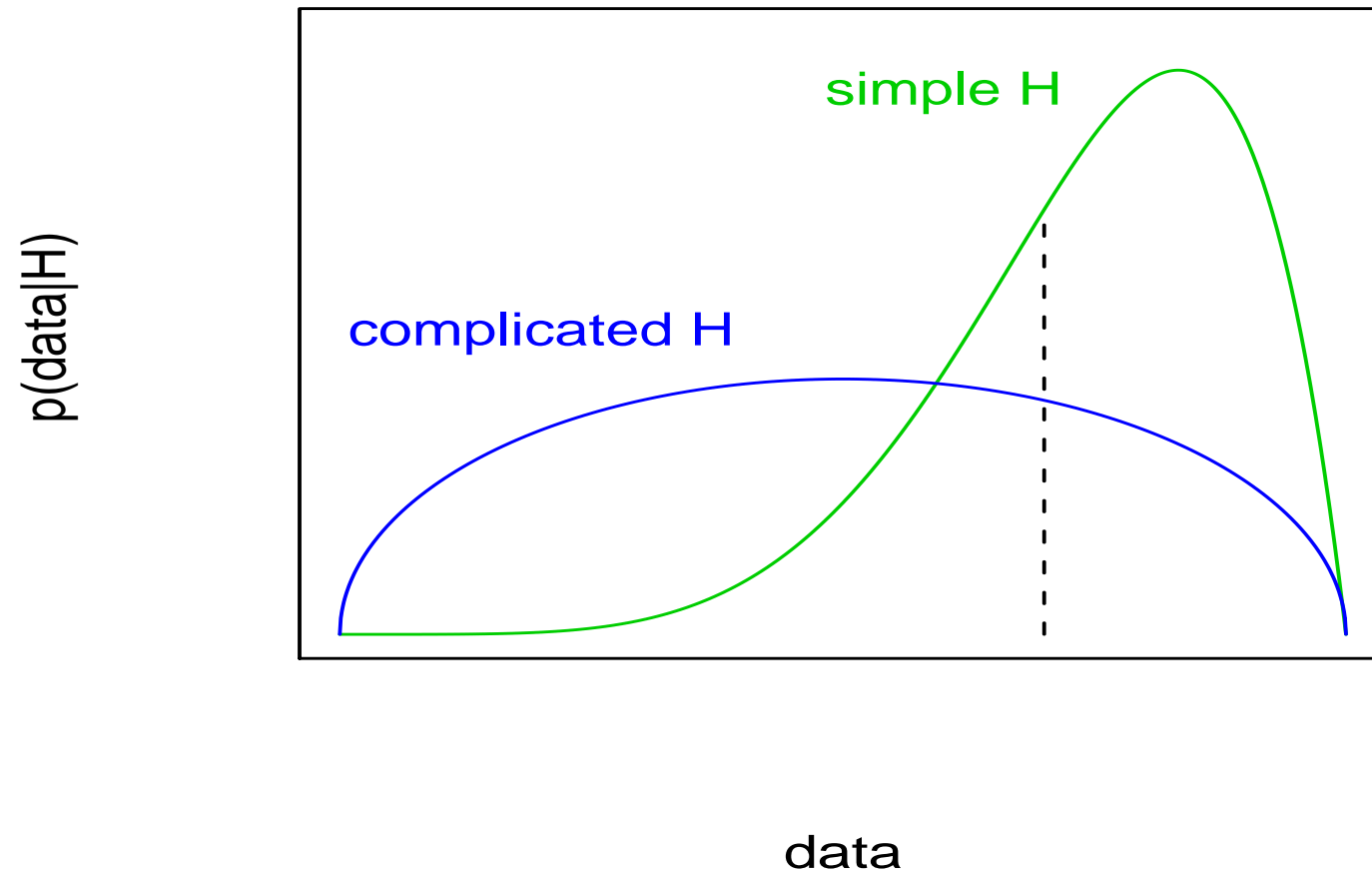
*'Pluralitas non est ponenda sine necessitate'*

*'frustra fit per plura quod potest fieri per pauciora'*

it is vain to do with more what can be done with fewer

... preferring the simpler of two hypothesis to the more complex when both agree with data is an old principle in science

## predictive probabilities prefer simple models



Bayes Factors automatically seek parsimony; no adhoc penalties for model complexity are needed. (Jefferys and Berger, 1992)

## Generality of application:

The approach is viable for *any* models

- regular or irregular,
- nested or not,
- large or small sample sizes,
- two or multiple models.

*Example:* In the location-scale example we choose among 5 non-nested models, with small sample sizes, irregular asymptotics, and no reduced sufficient statistics.

**\*\* Frequentist interpretation** (Berger, Brown and Wolpert):

- In many important situations involving testing of  $M_1$  versus  $M_2$ ,  $B_{12}/(1 + B_{12})$  and  $1/(1 + B_{12})$  are ‘optimal’ conditional frequentist error probabilities of Type I and Type II, respectively.
- Indeed,  $Pr(H_0|\text{data } \boldsymbol{x})$  is the *frequentist type I error probability*, conditional on observing data of the same “strength of evidence” as the actual data  $\boldsymbol{x}$ .
- (Classical unconditional error probabilities make the mistake of reporting the error averaged over data of very different strengths.)

## Accounting for model uncertainty:

- Selecting a single model and using it for inference ignores model uncertainty, resulting in
  - inferior inferences,
  - considerable overstatements of accuracy.
- The Bayesian approach incorporates this uncertainty by *model averaging*; if, say, inference concerning  $\xi$  is desired, it would be based on

$$\pi(\xi | \mathbf{x}) = \sum_{i=1}^q P(M_i | \mathbf{x}) \pi(\xi | \mathbf{x}, M_i).$$

Note:  $\xi$  needs to have the *same* meaning across models. Most frequent use is for prediction.

(see Geisser 93, Draper 95, Raftery et al. 97, Clyde 99 ... )

**Example:** In the location-scale example:

- **Posterior inferences** about the location  $\mu$  is based on the averaged posterior:

$$\begin{aligned}\pi(\mu | \mathbf{x}) &= w_N \pi_N(\mu | \mathbf{x}) + w_U \pi_U(\mu | \mathbf{x}) + w_C \pi_C(\mu | \mathbf{x}) + \\ &w_L \pi_L(\mu | \mathbf{x}) + w_R \pi_R(\mu | \mathbf{x})\end{aligned}$$

where  $w_i = Pr(M_i | \mathbf{x})$  was given before and  $\pi_i(\mu | \mathbf{x})$  is the usual objective posterior under model  $M_i$

- **Prediction** of next data  $X_{n+1}$  is similarly based on

$$\begin{aligned}m(x_{n+1} | \mathbf{x}) &= w_N m_N(x_{n+1} | \mathbf{x}) + w_U m_U(x_{n+1} | \mathbf{x}) \\ &+ w_C m_C(x_{n+1} | \mathbf{x}) + w_L m_L(x_{n+1} | \mathbf{x}) + w_R m_R(x_{n+1} | \mathbf{x})\end{aligned}$$

where  $m_i(x_{n+1} | \mathbf{x}) = \int f_i(x_{n+1} | \boldsymbol{\theta}_i) \pi_i(\boldsymbol{\theta}_i | \mathbf{x}) d\boldsymbol{\theta}_i$  is posterior predictive under  $M_i$

## Handling multiplicities (Scott & Berger)

- Suppose  $x_i \sim N(\mu_i, \sigma^2)$ ,  $i = 1, \dots, m$ , are observed, with  $\sigma^2$  known, and it is desired to determine which  $\mu_i$  are nonzero.
- Let  $p$  denote the (assumed common) prior probability that  $\mu_i$  is zero.
- Assume that the nonzero  $\mu_i$  follow a  $N(0, V)$  distribution, with  $V$  unknown.
- Assign  $p$  the uniform prior on  $(0, 1)$  and  $V$  the (proper) prior density  $\pi(V) = \sigma^2 / (\sigma^2 + V)^2$ .

- Then the posterior probability that  $\mu_i \neq 0$  is

$$p_i = 1 - \frac{\int_0^1 \int_0^1 p \prod_{j \neq i} \left( p + (1-p)\sqrt{1-w} e^{wx_j^2/(2\sigma^2)} \right) dp dw}{\int_0^1 \int_0^1 \prod_{j=1}^m \left( p + (1-p)\sqrt{1-w} e^{wx_j^2/(2\sigma^2)} \right) dp dw}.$$

- $(p_1, p_2, \dots, p_m)$  can be computed numerically if  $m$  is moderate. For large  $m$ , it is most efficient to do the computation via importance sampling, with a common importance sample for all  $p_i$ .

*Example:* Consider the following ten ‘signal’ observations:

-8.48, -5.43, -4.81, -2.64, -2.40, 3.32, 4.07, 4.81, 5.81, 6.24

Generate  $n = 10, 50, 500,$  and  $5000$   $N(0, 1)$  ‘noise’ observations.

| $n$  | Central seven 'signal' observations |      |      |      |     |     |      | #noise     |
|------|-------------------------------------|------|------|------|-----|-----|------|------------|
|      | -5.4                                | -4.8 | -2.6 | -2.4 | 3.3 | 4.1 | 4.81 | $p_i > .6$ |
| 10   | 1                                   | 1    | .94  | .89  | .99 | 1   | 1    | 1          |
| 50   | 1                                   | 1    | .71  | .59  | .94 | 1   | 1    | 0          |
| 500  | 1                                   | 1    | .26  | .17  | .67 | .96 | 1    | 2          |
| 5000 | 1.0                                 | .98  | .03  | .02  | .16 | .67 | .98  | 1          |

Table 1: The posterior probabilities of being nonzero for the central 'signal' means (the others always had  $p_i = 1$ ).

**Note:** The penalty for multiple comparisons is automatic.

## Comments:

- A model here would consist of the specification of which  $\mu_i$  are zero and which are nonzero. There are  $2^m$  models
- The  $P(M_j | \mathbf{x})$  are not of much interest.
- The posterior probability that  $\mu_i$  is nonzero is precisely the *posterior inclusion probability* for variable  $i$ , and can also be defined as

$$p_i \equiv \sum_{j: M_j \text{ has } \mu_i \neq 0} P(M_j | \mathbf{x}),$$

i.e., the overall posterior probability that variable  $i$  (that is,  $\mu_i$ ) is in the model.

## Further insights on prior distributions:

- Prior for model selection has two components:
  - Model specific prior  $\pi_i(\boldsymbol{\theta}_i)$
  - Prior on model space  $Pr(M_i)$

Both are important and control different aspects that non-Bayes methods have to control by ad-hoc methods

- Model specific prior
  - endorses Bayes model selection methods with an automatic **Occam's razor** effect: if two models are compatible with the data, choose the simpler one.
  - No need for ad-hoc penalties for model dimension

- Priors on model space control for **multiplicity**
  - No need for adjustments of the  $\alpha$  level, and/or the type of error to be ‘controlled’ and/or the way to ‘control’ it
  - Not all the priors adjust for multiplicity:
    - \* Fixed priors will not work: the % of true nulls needs to be learned from the data
    - \* Not all type of ‘learning’ works. An important particular case is the empirical Bayes prior.

These are very new, interesting and intriguing results (Berger and Scott 2009)

## Some Difficulties

### Inadequacy of common priors

such as objective priors, vague proper priors, and conjugate priors

1. **Objective priors** are usually improper ( $\int \pi(\theta) d\theta = \infty$ ) and improper priors cannot generally be used for model selection or hypothesis testing.
2. **'Vague' proper priors** (arbitrarily large variances or arbitrarily 'chopped' improper priors) are **worse**
3. **Conjugate priors** can also be bad.

## An important exception:

Objective (improper) priors can be used when the differing models have the same type of parameters, in particular when they have the same *invariance structure*, as in the location-scale example; Berger, Pericchi and Varshavsky, 1998, show the *right-Haar* prior can then be used.

BUT note that improper (or proper but ‘vague’) priors can not be used for parameters not occurring in all models

*Example:* in the hatching eggs example,  $\theta$  was not present in  $H_0$  and got a proper uniform prior under  $H_1$

## Normal example:

- *Data:*  $X_1, X_2, \dots, X_n$  are i.i.d  $N(\theta, \sigma^2)$
- *To test:*  $M_1 : \theta = 0$  versus  $M_2 : \theta$  arbitrary
- Assume  $\sigma^2$  known and (under  $M_2$ )

**objective** prior:  $\pi_2^N(\theta) = c$

$$\text{Bayes factor: } B_{21} = c \sqrt{\frac{2\pi}{n}} \exp\left\{\frac{n}{2} \bar{x}^2\right\}$$

*But  $c$  was arbitrary, so  $B_{21}$  is then arbitrary!*

**vague proper** prior:  $\pi(\theta) = N(\theta | 0, \tau^2)$ ,

$$\text{Bayes factor: } B_{21} \approx \frac{1}{\tau\sqrt{n}} \exp\left\{\frac{n\bar{x}^2}{2}\right\} \quad \text{for arbitrarily large } \tau.$$

Thus choosing a vague proper prior, i.e., *choosing  $\tau$  very large, gives the ridiculous answer that  $M_1$  is favored!*

**Moral:** NEVER USE ‘ARBITRARY’ VAGUE PROPER PRIORS FOR MODEL SELECTION ; improper non-informative priors can give reasonable results.

**conjugate** priors. Assume now  $\sigma^2$  unknown and (under  $M_2$ ) the

*“Natural” conjugate priors:*

$$\pi_1(\sigma^2) = 1/\sigma^2,$$

$$\pi_2(\theta | \sigma^2) = N(\theta | 0, \sigma^2), \quad \pi_2(\sigma^2) = 1/\sigma^2.$$

- Note that the improper  $\pi_i(\sigma^2) = 1/\sigma^2$  can be used because they are both invariant scale parameters.
- Using “vague” inverse gamma priors for the  $\sigma^2$  would result in the same problems shown below.

Bayes factor:  $B_{21} = \sqrt{n+1} \left( \frac{1+t^2/(n+1)}{1+t^2} \right)^{n/2}$   
 where  $t = \sqrt{n}\bar{x}/s$  is the usual t-statistic.

Silly behavior: as  $|t| \rightarrow \infty$ ,  $B_{12} \rightarrow (n+1)^{-(n-1)/2} > 0$

$$\begin{aligned} \text{For instance: if } n = 3, & \quad B_{12} \rightarrow \frac{1}{4} \\ & \quad \text{if } n = 5, \quad B_{12} \rightarrow \frac{1}{36} \end{aligned}$$

**Note:** Commonly used *g-priors* always exhibit this behavior when  $\sigma^2$  is unknown, but it is only a serious problem for small  $n$  (relative to number of parameters).

This unappealing property is becoming known as “**information inconsistency**” and it is associated with thin tails

## \*\* Meaning of parameters

Parameters in different models typically have different meanings, so that separate (proper) priors must be constructed for each model

The difficulty applies to both objective priors for model selection *and* subjective priors; indeed, when parameters do not have a meaning independent of the model, construction of suitable subjective priors can be very difficult.

## Example:

### WRONG WAY:

$$M_1 : Y = \beta_0 + \beta_1 X_1 + \sigma \epsilon$$

$$M_2 : Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \sigma \epsilon, \quad \epsilon \sim N(0, 1)$$

*Single Prior:*  $\pi(\beta_0) \pi(\beta_1) \pi(\beta_2) \pi(\sigma)$ , independent of the model.

(This may be okay if  $\mathbf{X}'\mathbf{X}$  is orthogonal.)

### RIGHT WAY:

$$M_1 : Y = \beta_0^{(1)} + \beta_1^{(1)} X_1 + \sigma^{(1)} \epsilon$$

$$M_2 : Y = \beta_0^{(2)} + \beta_1^{(2)} X_1 + \beta_2^{(2)} X_2 + \sigma^{(2)} \epsilon.$$

*Distinct Priors:*  $\pi_1(\beta_0^{(1)}, \beta_1^{(1)}, \sigma^{(1)})$  and  $\pi_2(\beta_0^{(2)}, \beta_1^{(2)}, \sigma^{(2)})$ .

## SPECIFIC EXAMPLE

Predict fuel consumption  $Y$  from weight  $X_1$  and engine size  $X_2$ .

- $\beta_1$  has very different meaning under  $M_1$  and  $M_2$ ; regressing  $Y$  on  $X_1$  alone produces larger  $\beta_1$  than regressing on both (due to high correlation).
- $\sigma_1^2$  will typically be larger than  $\sigma_2^2$

## Computation

Computation of the needed marginal distributions (or their ratios) can be very challenging in high dimensions.

The total number of models under consideration can be enormous (e.g., in variable selection), so good search strategies are needed.

We do not address computation here. See Carlin and Chib (1995), Green (1995), Kass and Raftery (1995), Verdinelli and Wasserman (1995), Raftery, Madigan and Hoeting (1997), Clyde (1999), Chib and Jeliazkov (2001), Dellaportas, Forster and Ntzoufras (2001), Godsill (2001), and Han and Carlin (2001), Berger and Molina (2002), among others.

## Evaluating model selection methodologies.

- There are no clear guidelines for evaluating objective model selection procedures.
- Some criteria, such as consistency, would seem obvious, but are violated by many proposed Bayesian procedures.
- Success in examples, real and ‘hard artificial’ is helpful.
- *Our favorite guiding principle:* Objective model selection procedures should correspond, in some way, to actual Bayes factors, posterior probabilities or posterior expected losses, arising from reasonable priors (*intrinsic priors*: Berger and Pericchi, 1996, 2001, Berger and Mortera, 1999, Moreno, Bertolino, and Racugno, 2000, Sun and Kim, 2000)

## Conclusions 2

- Bayesian model selection has very nice properties. It is generally applicable, is consistent, handles multiplicities, sequential testing, incorporates model uncertainty . . . , etc.
- Bayesian model selection acts as an automatic Occam's razor: **NO AD-HOC PENALTIES** needed.
- 'standard' prior can not be used in general; improper priors can be used sometimes, **'vague proper priors' are specially dangerous.**
- attention to different meaning of 'same name' parameters
- 'objective', default proposals should approximately correspond to actual Bayesian analyses

## Some Solutions: Methodologies for BMS

1. Use **subjectively elicited priors**.
2. **Orthogonalize parameters**, directly or indirectly
3. Use **predictively matched** priors
4. Use a **default strategy** such as
  - Conventional proper priors
  - Asymptotics, such as BIC
  - Intrinsic Bayes factors and related approaches
  - Fractional Bayes factors and extensions
  - Expected posterior priors

**Subjectively** elicit priors.

- Removes difficulties with objective and “vague proper” priors
- It’s perfectly Bayesian, so no need to evaluate it
- The rest of difficulties remain
- It is virtually always unfeasible in practice

**Orthogonalize** parameters, so that common orthogonal parameters can sometimes be argued to have same ‘meaning’ across models, thus often justifying use of a common objective (or subjective) prior across models.

Objective (improper) priors are often used in common, orthogonal parameters.

## Predictively matched priors

**Idea:** If  $\mathbf{X}^*$  are observables of interest, the pair  $\{M_i, \pi_i\}$  would yield the density:

$$m_i(\mathbf{x}^*) = \int f_i(\mathbf{x}^* | \theta_i) \pi_i(\theta_i) d\theta_i$$

$\{M_i, \pi_i\}$  is “predictively matched” to  $\{M_j, \pi_j\}$  if

$$m_i(\mathbf{x}^*) \approx m_j(\mathbf{x}^*)$$

**Note:**  $\mathbf{X}^*$  is some small set of possible observations. Often, it is a ‘minimal training sample’

**Example:** Use of right Haar measures as priors for models having the same invariance structure (Berger, Pericchi and Varshawsky, 1998)

**Lemma:** If  $x$  and  $x'$  are i.i.d  $\frac{1}{\sigma}g\left(\frac{x-\mu}{\sigma}\right)$  then

$$\begin{aligned} m(x, x') &= \int \int \left[ \frac{1}{\sigma^2} g\left(\frac{x-\mu}{\sigma}\right) g\left(\frac{x'-\mu}{\sigma}\right) \right] \frac{1}{\sigma} d\mu d\sigma \\ &= \frac{1}{2|x-x'|} \quad (\text{for all } g). \end{aligned}$$

Therefore,  $\pi^N(\mu, \sigma) = 1/\sigma$  *simultaneously* provides a type of predictive match for *all* location-scale families.

Predictive matching is an important concept: objective priors should be 'balanced' accross models.

## Conventional (partly proper) priors

Jeffreys (1961, generalized by Zellner & Siow) proposed to deal with the indeterminacy of improper priors by:

- Using noninformative (improper) priors only for common (orthogonal) parameters

**Note:** arbitrary constants cancel

- Using default **proper** priors (but **NOT** vague proper priors) for parameters that occur in one model but not the other

### Example: Jeffreys (1939)

- $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} N(\theta, \sigma^2)$
- To test:  $M_1 : \theta = 0$        $M_2 : \theta \neq 0$
- use  $\pi_1(\sigma^2) = \frac{1}{\sigma^2}$  under  $M_1$

$$\pi_2(\theta, \sigma^2) = \frac{1}{\sigma^2} \frac{1}{\sigma \pi \left(1 + \frac{\theta^2}{\sigma^2}\right)} \quad \text{under } M_2$$

that is,  $\pi_2(\sigma^2) = 1/\sigma^2$  and  $\pi_2(\theta | \sigma^2) = \text{Cau}(\theta | 0, \sigma^2)$

**Note:** because  $\sigma^2$  is a scale parameter under each model, use of the “conventional” prior  $1/\sigma^2$  for  $\sigma^2$  is reasonable. But  $\theta$ , which occurs only in  $M_2$  must have a proper prior.

Jeffreys argued that a sensible 'objective' prior under  $M_2$ :

- is centered at zero (the simpler model,  $M_1$ );
- has scale  $\sigma$  (the scale of 1 observation)
- is symmetric around zero ( $M_1$ )
- has no moments (to avoid information inconsistency)

## Predictive matching:

Jeffrey's priors are predictively matched for one observation  $y$

$$m_1(y) = m_2(y) = \frac{1}{|y|}$$

## Linear Models (Zellner Siow priors)

- Assume the 'full' model:

$$\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_e\boldsymbol{\beta}_e + \boldsymbol{\epsilon} \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

(full rank matrices, 'orthogonal' parameters)

- To test:  $H_1 : \boldsymbol{\beta}_e = \mathbf{0}$  versus  $H_2 : \boldsymbol{\beta}_e \neq \mathbf{0}$
- ZS (1980, 1984) prior  $\pi_i$  under  $M_i$ ,  $i = 1, 2$ :

$$\pi_1(\boldsymbol{\beta}_1, \sigma) = \sigma^{-1},$$

$$\pi_2(\boldsymbol{\beta}_1, \boldsymbol{\beta}_e, \sigma) = \sigma^{-1} \text{Ca}_{k_e}(\boldsymbol{\beta}_e \mid \mathbf{0}, n\sigma^2(\mathbf{X}_e^t\mathbf{X}_e)^{-1}),$$

- i.e., in the spirit of Jeffreys, 'common'  $(\boldsymbol{\beta}_1, \sigma)$  parameters have *same* improper prior, non-common  $\boldsymbol{\beta}_e$  (conditional on the common parameters) has a (proper) Cauchy prior

- Bayes factors are not closed form but have very easy expressions as one dimensional integrals in terms of the residual sum of squares:

$$B_{21} = \int \left( 1 + t n \frac{SSE_2}{SSE_1} \right)^{-(n-q_1)/2} (1 + t n)^{(n-q)/2} IGa\left(t \mid \frac{1}{2}, \frac{1}{2}\right) dt$$

which is easy to evaluate either numerically or by MC.

Alternatively, analytical (Laplace) approximations can be used for huge model spaces

- The **same** expression is valid also in the non-orthogonal case, for matrices not of full rank, and for 'null' models given by linear restrictions (ANOVA models)(Bayarri & García-Donato)

## Example: Hald regression data

*Possible regressors:*  $X_1, X_2, X_3, X_4$

*Full Model:* is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

*Models under consideration:* Subsets of regressors (all models contain the intercept  $\beta_0$ ).

*Notation:* Model  $\{1,3,4\}$  means

$$Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

*Answers:*

| model         | posterior probability |
|---------------|-----------------------|
| $\{1,2,3,4\}$ | 0.0107                |
| $\{1,2,3\}$   | 0.1018                |
| $\{1,2,4\}$   | 0.1030                |
| $\{1,3,4\}$   | 0.0816                |
| $\{2,3,4\}$   | 0.0181                |
| $\{1,2\}$     | 0.5180                |
| $\{1,4\}$     | 0.1634                |
| $\{3,4\}$     | 0.0034                |
| others        | $< 0.0003$            |

## Using the posterior probabilities:

- Optimal use is through model averaging.
- The most common use is through selecting the largest posterior probability model.
- But the best single model is typically the *median probability model*, defined (Barbieri and Berger, 2002) as that model consisting of precisely the variables whose **posterior inclusion probability** (i.e. overall posterior probability of being included in a model) is at least  $1/2$ .

*Note:* If computation is done by a model-jumping MCMC, the median probability model consists of those coordinates that were present in over half the iterations.

- For the Hald data, the posterior inclusion probabilities are

$$p_1 = \sum_{j: M_j \ni \text{variable 1}} P(M_j | \mathbf{y}) = 0.978,$$

$$p_2 = 0.752, \quad p_3 = 0.216, \quad p_4 = 0.380.$$

- Thus the median probability model is  $\{X_1, X_2\}$ , which, in this case, is also the highest probability model

Inclusion probabilities are an important concept in Bayesian model selection. Appears naturally in variable selection, multiple testing, search in huge model spaces .... their full potential is still under investigation.

# Extensions of Jeffreys-Zellner-Siow methodology

## 1. Using other model specific priors

- Using conjugate Normal ( $g$  priors) gives close-form expressions, but produces *information inconsistency*, that is for a finite  $n$ ,  $B_{01}$  would not go to 0 as the evidence in favor of  $M_1$  goes to  $\infty$   
 $\rightsquigarrow$  thin tails not good
- Use instead mixtures of Normal  $g$ -priors (like ZS and others)  $\rightsquigarrow$  consistency, but not close-form
- Rescue for model selection priors used in *Robust* estimation contexts (-: see Forte et al. Poster :-)

## 2. Extending the methodology (Jeffreys' ideas) outside linear models (-: see García-Donato's talk in Jeffreys' session :-)

Intrinsic priors (Moreno's talk) are very good priors which are generally applicable (non only for linear models)

## Asymptotic Approaches

Use Asymptotics expansions, leading to crude (sometimes *very* crude) approximations to Bayes factors, as BIC (*Bayesian Information Criterion*). They are based on Laplace approximation:

$$\begin{aligned}
 B_{21}^L &= \frac{\int f_2(\mathbf{x}|\boldsymbol{\theta}_2)\pi_2(\boldsymbol{\theta}_2)d\boldsymbol{\theta}_2}{\int f_1(\mathbf{x}|\boldsymbol{\theta}_1)\pi_1(\boldsymbol{\theta}_1)d\boldsymbol{\theta}_1} \\
 &\approx \frac{f_2(\mathbf{x}|\hat{\boldsymbol{\theta}}_2)|\hat{\mathbf{I}}_2|^{-1/2}}{f_1(\mathbf{x}|\hat{\boldsymbol{\theta}}_1)|\hat{\mathbf{I}}_1|^{-1/2}} \cdot \frac{(2\pi)^{q_2/2}\pi_2(\hat{\boldsymbol{\theta}}_2)}{(2\pi)^{q_1/2}\pi_1(\hat{\boldsymbol{\theta}}_1)},
 \end{aligned}$$

where  $\hat{\boldsymbol{\theta}}_1$  and  $\hat{\boldsymbol{\theta}}_2$  are the m.l.e.'s for  $\boldsymbol{\theta}_1$  and  $\boldsymbol{\theta}_2$  (which have dimensions  $q_1$  and  $q_2$ ) and  $\hat{\mathbf{I}}_1$  and  $\hat{\mathbf{I}}_2$  are observed information matrices.

**Idea:** Replace  $\pi_2(\hat{\boldsymbol{\theta}}_2)/\pi_1(\hat{\boldsymbol{\theta}}_1)$  by an “appropriate” constant.

Example: BIC (Bayes Information Criterion)  $B_{21} \approx \frac{f_2(\mathbf{x}|\hat{\boldsymbol{\theta}}_2)}{f_1(\mathbf{x}|\hat{\boldsymbol{\theta}}_1)} \cdot n^{\frac{1}{2}(q_2-q_1)}$

where  $q_2 = \dim(\boldsymbol{\theta}_2)$  and  $q_1 = \dim(\boldsymbol{\theta}_1)$

### Some Problems:

- Requires **fixed**  $q_i$  and  $n$  **MUCH** larger than  $q_i$
- Improved versions keep the crucial role of the prior (Berger et al. 2001 GBIC, Berger et al\* 2009, PBIC). Furthermore, **PBIC** has very simple close-form expressions :-)
- It is not very clear what  $n$  in the formula should be. Use instead a “good” definition of **effective sample size** (-: see BBP talk at JSM about **TESS** :-)

# Training Sample Methods of Model Selection

- Huge literature

(Good, 1950, Smith and Spiegelhalter, 1980, de Voss, 1993, Gelfand and Dey, 1994, O'Hagan, 1995, 1997, Varshavsky, 1995, Berger and Pericchi, 1996, . . . , 2002, De Santis and Spezzaferri, 1996, 1997, Dmochowski, 1996, Sansó, Pericchi and Moreno, 1996, Bertolino and Racugno, 1997, Iwaki, 1997, Gelfand and Ghosh, 1998, Lingham and Sivaganesan, 1997, 1999, Moreno, Bertolino and Racugno, 1998, 1999, Perez, 1998, Ghosh and Samanta, 1999, Key, Pericchi and Smith, 1999, Nadal, 1999, Schluter, Deely and Nicholson, 1999, Beattie, Fong, and Lin, 2001, Berger and Perez, 2002, Neal, 2002, . . . )

- Uses (imaginary or part of real) data to train the improper prior into a proper prior, and the rest to compute Bayes factors
  - An example: median intrinsic posterior probabilities
  - Fractional Bayes factors
  - Expected posterior priors

## Example: median intrinsic posterior probabilities

(modeled after Berger and Pericchi, 1998)

*Data:*  $X_1, X_2, \dots, X_n$  are  $N(\theta, 1)$

*Models:*  $M_1 : \theta = 0, \quad M_2 : \theta \neq 0$

*Standard objective prior:*

$$Pr(M_1) = Pr(M_2) = \frac{1}{2}; \quad \text{Under } M_2, \pi_2(\theta) = 1$$

*(Formal) Bayes factor:*  $B_{21} = \frac{1}{\sqrt{n}} e^{-\frac{n}{2}\bar{x}^2}$

*(Formal) posterior probability of  $M_1$ :*  $P_1 = (1 + \sqrt{n} e^{-\frac{n}{2}\bar{x}^2})^{-1}$

*Note:* This is not legitimate, since  $\pi_2(\theta)$  was improper and  $M_1$  and  $M_2$  do not both have an unknown location parameter.

## Obtaining a proper prior by use of a training sample:

Choose one observation, say  $x_i$ , and compute

$$\pi_2(\theta \mid x_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\theta - x_i)^2} \text{ (proper).}$$

Use on the remaining data,  $\mathbf{x}^{(i)} \equiv (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$ , to compute the posterior model probabilities (or the Bayes factors for more than two models)

$$P_1(x_i) = 1 - P_2(x_i) = [1 + \sqrt{n} e^{\frac{n}{2}\bar{x}^2} e^{-\frac{1}{2}x_i^2}]^{-1}.$$

**Intrinsic methods:** Since the resulting probabilities (and Bayes factors) depend on the specific  $x_i$  chosen, do it for all possibilities and ‘average’

## Find the median intrinsic posterior probability:

- Find  $P_1(x_i)$  for all training samples  $\{x_i, i = 1, \dots, n\}$ ;
- Use the median of  $P_1(x_i)$  (and  $P_2(x_i)$ ) over all training samples,

$$P_1^{\text{med}} = 1 - P_2^{\text{med}} = [1 + \sqrt{n} e^{\frac{n}{2}\bar{x}^2} e^{-\frac{1}{2}\text{med}\{x_i^2\}}]^{-1},$$

as the recommended conventional posterior probabilities of  $M_1$  and  $M_2$ .

**Note:** one need only sample from the possible training samples and take the median posterior probability over those sampled. Indeed, if  $n$  is the sample size of the data, it usually suffices to draw just  $l n$  (sets) of training samples.

## General Algorithm:

- Begin with standard objective priors,  $\pi_i^N$ , for the parameters  $\theta_i$  in the model  $M_i$  ( $\pi_i^N(\theta_i) = 1$  is okay).
- Define a “minimal training sample,”  $\mathbf{x}^* = (x_1^*, \dots, x_l^*)$ , as any subset of the data which is as small as possible such that the posterior distributions,  $\pi_i^N(\theta_i | \mathbf{x}^*)$ , are all proper. (Usually,  $l = \#$  parameters in largest model.)
- Compute the required Bayes factors with the remaining data, using the  $\pi_i^N(\theta_i | \mathbf{x}^*)$  as priors.
- Do this for every possible minimal training sample,  $\mathbf{x}^*$ , and take the median of the results.

## Example: Hald regression data

*Possible regressors:*  $X_1, X_2, X_3, X_4$  (plus an intercept)

*Initial objective priors:*  $\pi_i(\boldsymbol{\beta}, \sigma) = 1/\sigma$

*Minimal training samples:*  $\mathbf{y}^*$  consists of any subset of six distinct observations (since there are a maximum of six unknown parameters, including  $\sigma^2$ ) and their covariates

*Formula for  $m_i^N$ :*

$$m_i^N(\mathbf{y}) = \frac{\pi^{k_i/2} \Gamma((n - k_i)/2)}{\sqrt{\det(X_{(i)}^t X_{(i)})} R_i^{(n-k_i)/2}},$$

where, for model  $M_i$ ,  $k_i$  is the number of regressors plus one,  $X_{(i)}$  is the design matrix, and  $R_i$  is the residual sum of squares.

*Answers:*

| model     | posterior probability | Zellner-Siow |
|-----------|-----------------------|--------------|
| {1,2,3,4} | 0.049                 | 0.0107       |
| {1,2,3}   | 0.171                 | 0.1018       |
| {1,2,4}   | 0.190                 | 0.1030       |
| {1,3,4}   | 0.160                 | 0.0816       |
| {2,3,4}   | 0.041                 | 0.0181       |
| {1,2}     | 0.276                 | 0.5180       |
| {1,4}     | 0.108                 | 0.1634       |
| {3,4}     | 0.004                 | 0.0034       |
| others    | < 0.0003              | < 0.0003     |

- For the Hald data, the posterior inclusion probabilities are

$$p_1 = \sum_{j: M_j \ni \text{variable 1}} P(M_j | \mathbf{y}) = 0.955,$$

$$p_2 = 0.727, \quad p_3 = 0.425, \quad p_4 = 0.553.$$

- Thus the median probability model is  $\{X_1, X_2, X_4\}$ , which is *not* the most probable model.
- Computation of predictive risks shows that the median probability model has the lowest risk of all single models, considerably below that of the maximum probability model.

## Fractional Bayes Factor (O'Hagan, 1995, 1997)

**Idea:** Instead of using a fraction of the data as a training sample, use a fraction of the likelihood

### Algorithm:

- Choose some “fraction”  $0 < b < 1$
- For model  $M_j$ , choose the (proper) prior
$$\pi_j^*(\boldsymbol{\theta}_j) \propto [f_j(\boldsymbol{x} \mid \boldsymbol{\theta}_j)]^b \cdot \pi_j^N(\boldsymbol{\theta}_j)$$
- Compute marginal likelihoods (and Bayes factors) using these priors and the “remaining likelihoods”:

$$m_j(\boldsymbol{x}) = \int [f_j(\boldsymbol{x} \mid \boldsymbol{\theta}_j)]^{(1-b)} \pi_j^*(\boldsymbol{\theta}_j) d\boldsymbol{\theta}_j$$

- broadly applicable (but serious concerns in irregular problems).

## Expected Posterior Priors (Perez and Berger, 2001, 2002, Neal, 2002)

**Initial priors:**  $\pi_i^N(\boldsymbol{\theta}_i)$ , typically improper.

**Initial marginals:** corresponding  $m_i^N(\mathbf{y})$

**Training sample posteriors:** Training sample,  $\mathbf{y}^*$ , such that posterior distributions  $\pi_i^N(\boldsymbol{\theta}_i | \mathbf{y}^*)$  exist  $\forall i$ . The ‘prior’ densities:

$$\pi_i^*(\boldsymbol{\theta}_i) = \int \pi_i^N(\boldsymbol{\theta}_i | \mathbf{y}^*) m^*(\mathbf{y}^*) d\mathbf{y}^*,$$

is the *expected posterior priors* (or EP priors) for the  $\boldsymbol{\theta}_i$ , with respect to  $m^*$  ( $\mathbf{y}^*$  can have different dimensions for the different models)

*Note:* The EP priors,  $\pi_i^*(\boldsymbol{\theta}_i)$ , will not be proper unless  $m^*$  itself is proper, but are always properly ‘calibrated’ across models.

*An attractive particular choice* of  $m^*$  is the empirical distribution

## An aside: using EPPriors subjectively

- Subjectively elicit a predictive  $m(\mathbf{x}^*)$ , where  $\mathbf{x}^*$  is a minimal training sample
- Goal: for each  $M_i$ , find  $\pi_i^*(\theta_i)$  so that  $m_i^* = \int f_i(\mathbf{x}^* | \theta_i) \pi_i(\theta_i) d\theta_i$  is close to  $m(\mathbf{x}^*)$ .
- Define  $\pi_i^{(l)}(\theta_i) = \int \pi_i^{(l-1)}(\theta_i | \mathbf{x}^*) m(\mathbf{x}^*) d(\mathbf{x}^*)$ . Then  $\pi_i^{(l)}$  converges to a  $\pi^*$  such that the resulting  $m_i^*(\mathbf{x}^*)$  is closest to  $m^*(\mathbf{x}^*)$  in Kullback-Leibler divergence. (Shyamalkumar, 1996).
- The first step is the ‘biggest’ and is the EPPrior with respect to  $m(\mathbf{x}^*)$ ; this is also easy to compute with!

## Concluding Comments

- Fully subjective Bayes MS is virtually never applicable.
- Methodologies for O'Bayes estimation and model selection are **very different**. MS mainly based in:
  - Choosing a 'good' prior: 'conventional' (partly proper), predictively matched, intrinsic, robust, DB priors . . .
  - Rely on asymptotics
  - Use common objective priors, but train them or calibrate the resulting Bayes factors with the aid of some data (real or imaginary)
- O'Bayes MS show considerable promise and very satisfactory results, but much more research is needed

Many thanks for your attention!