Hypothesis Testing and Model Selection

M.J. Bayarri and James O. Berger
Universitat de València and Duke University

www.stat.duke.edu/∼berger/talks/valencia7.ps

Valencia 7, Playa de Las Américas, Tenerife, June 2002

Overall Outline

I. Basics of Bayesian hypothesis testing

II. Basics of Bayesian model selection

- Notation and an example
- Motivation for the Bayesian approach
- Difficulties with the Bayesian approach

III. Methodologies for Bayesian model selection

- Overview
- Training sample methods
- A case study with conventional proper priors

Part I:

Basics of Bayesian hypothesis testing

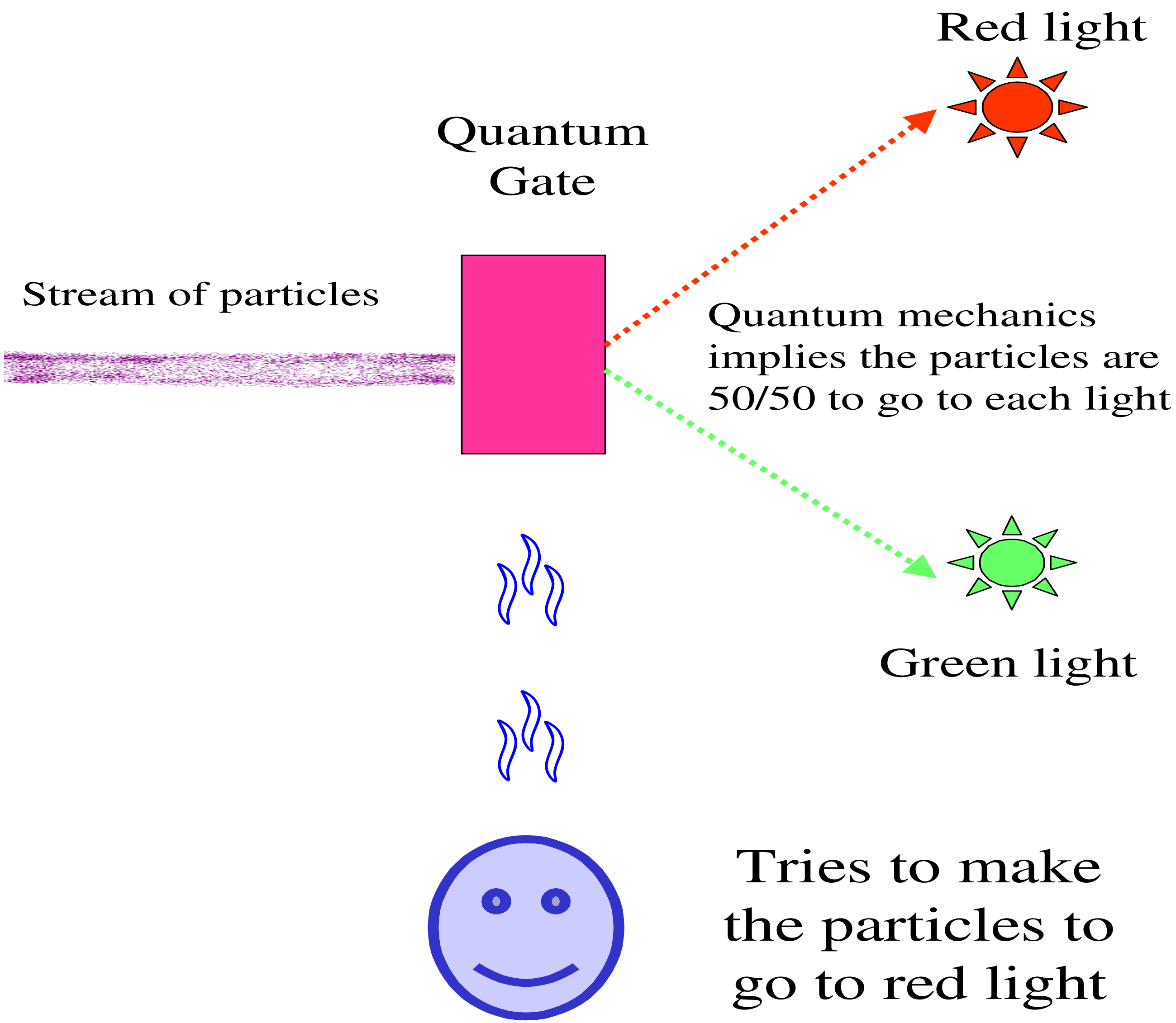
- Formulation of Hypothesis testing (through an example)
- A priori plausibility of hypotheses
- Clash between p-values and Bayesian answers
- Bayesian reporting

Psychokinesis Example

Does a subject possess Psychokinesis?

The experiment:

Schmidt, Jahn and Radin (1987) used electronic and quantum-mechanical random event generators with visual feedback; the subject with alleged psychokinetic ability tries to "influence" the generator.


Red light
Quantum
Gate
Stream of particles
Quantum mechanics
implies the particles are
50/50 to go to each light
Green light
Tries to make
the particles to
go to red light

Data and model:

- Each "particle" is a Bernoulli trial (red $=1$, green $=0$)
  $\theta =$ probability of "1"
  $n = 104,490,000$ trials
  $X = \#$ "successes" ($\#$ of 1's), $\quad X \sim \mathrm{Binomial}(n, \theta)$
  $x = 52,263,470$ is the actual observation

  To test $H_0 : \theta \;=\; \frac{1}{2}$ (subject has no influence)

  versus $H_1 : \theta \;\neq\; \frac{1}{2}$ (subject has influence)

- P-value $= P_{\theta=\frac{1}{2}}(|X - \frac{n}{2}| \geq |x - \frac{n}{2}|) \approx .0003.$
  Is there strong evidence against H_0 (i.e., strong evidence that the subject influences the particles) ?

Bayesian Analysis: (Jefferys, 1990)

Prior distribution:

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

On $H_1 : \theta \neq \frac{1}{2}$, let $\pi(\theta)$ be the prior density for θ.

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

Objective (or default) Bayes: choose

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

$\pi(\theta) = 1$ (on $0 < \theta < 1$)

Posterior probability of hypotheses:

$$
\begin{aligned}
Pr(H_0|x) &= \text{probability that } H_0 \text{ true, given data, } x \\
&= \frac{f(x\,|\,\theta{=}\frac{1}{2})\,Pr(H_0)}{Pr(H_0)\,f(x\,|\,\theta{=}\frac{1}{2}){+}Pr(H_1)\int f(x\,|\,\theta)\pi(\theta)d\theta}
\end{aligned}
$$

For the objective prior,

$$Pr(H_0 \mid x = 52,263,470) \approx 0.92$$

(recall, p-value $\approx$ .0003)

Posterior density on $H_1 : \theta \neq \frac{1}{2}$ is

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

the $Be(\theta \mid 52,263,470 \ , \ 52,226,530)$ density.

Bayes Factor:

An 'objective' alternative to choosing $Pr(H_0) = Pr(H_1) = \frac{1}{2}$ is to report the *Bayes factor*

$$B_{01} = \frac{\text{likelihood of observed data under } \mathrm{H}_0}{\text{`average' likelihood of observed data under } \mathrm{H}_1}$$

$$= \frac{f(x \mid \theta{=}\frac{1}{2})}{\int_0^1 f(x \mid \theta)\pi(\theta)d\theta} \approx 12$$

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

so B_{01} is often thought of as "the odds of H_0 to H_1 provided by the data"

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)

H_0 : Subject has no psychokinetic ability
is a plausible hypothesis

More typical example: Compare Treatments A and B.

$\theta =$ mean treatment difference

$$H_0 : \theta \approx 0 \quad \text{versus} \quad H_1 : \theta \neq 0$$

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 θ, 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 $\approx .0003$, but

the objective posterior probability of the null ≈ 0.92;

(Bayes factor gives 12 to 1 odds in favor of the null).

Was the prior on θ inappropriate? (Few would say it was unfair to give H_0 prior probability of 1/2.)

- 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

An Aside: 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 it can be shown that, if $p < e^{-1}$,

$$B_{01} \geq -e\, p \log(p).$$

- An analogous *lower bound* on conditional Type I frequentist error is $\alpha \geq (1 + [-e\, p \log(p)]^{-1})^{-1}$.

Example: Are gamma ray bursts galactic or extra-galactic in origin?

- data in early 90's were 260 observed burst directions
- H_0: data are uniformly directionally distributed (implying extra-galactic origin)
- standard test for uniformity rejected at $p = 0.027$
- $B(p) \geq -e\,p\,\mathsf{log}(p) = -e\,(.027)\,\mathsf{log}(.027) = 1/(3.7)$, so at most odds of 3.7 to 1 against H_0
- $\alpha(p) \geq (1 + [-e\,(.027)\,\mathsf{log}(.027)]^{-1})^{-1} = .21$, so the actual error rate in rejecting H_0 is *at least* .21

Bayesian Reporting in Hypothesis Testing

- The complete posterior distribution is given by
  - $Pr(H_0|x)$, the posterior probability of null hypothesis
  - $\pi(\theta|x, H_1)$, the posterior distribution of θ under H_1
- A useful *summary* of the complete posterior is
  - $Pr(H_0|x)$
  - C, a (say) 95% posterior credible set for θ under H_1
- In the psychokinesis example
  - $Pr(H_0|x) = .92 \rightsquigarrow$ gives the probability of H_0
  - $C = (.50008, .50027) \rightsquigarrow$ shows where θ is if H_1 is true
- For testing precise hypotheses, confidence intervals alone are *not* a satisfactory inferential summary

Part I Conclusions

- $Pr(H_0 \mid x)$ 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.

Part II:

Basics of Bayesian model selection

- Notation and an example
- Motivation for the Bayesian approach
- Difficulties with the Bayesian approach

II.A. 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 $f_1(\mathbf{x}|\boldsymbol{\theta}_1)$

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

$$\ddots$$

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

Assign prior probabilities, $P(M_i)$, to each model (usually, $P(M_i) = \frac{1}{q}$, but having decreasing $P(M_i)$, as model complexity increases, is often reasonable).

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 $\boldsymbol{x}$ 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.

Example: (location-scale)

Suppose $X_1, X_2, \ldots, X_n$ 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 ($\frac{1}{\sigma}\, e^{x-\mu}$, $x \leq \mu$)

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

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.

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

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

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

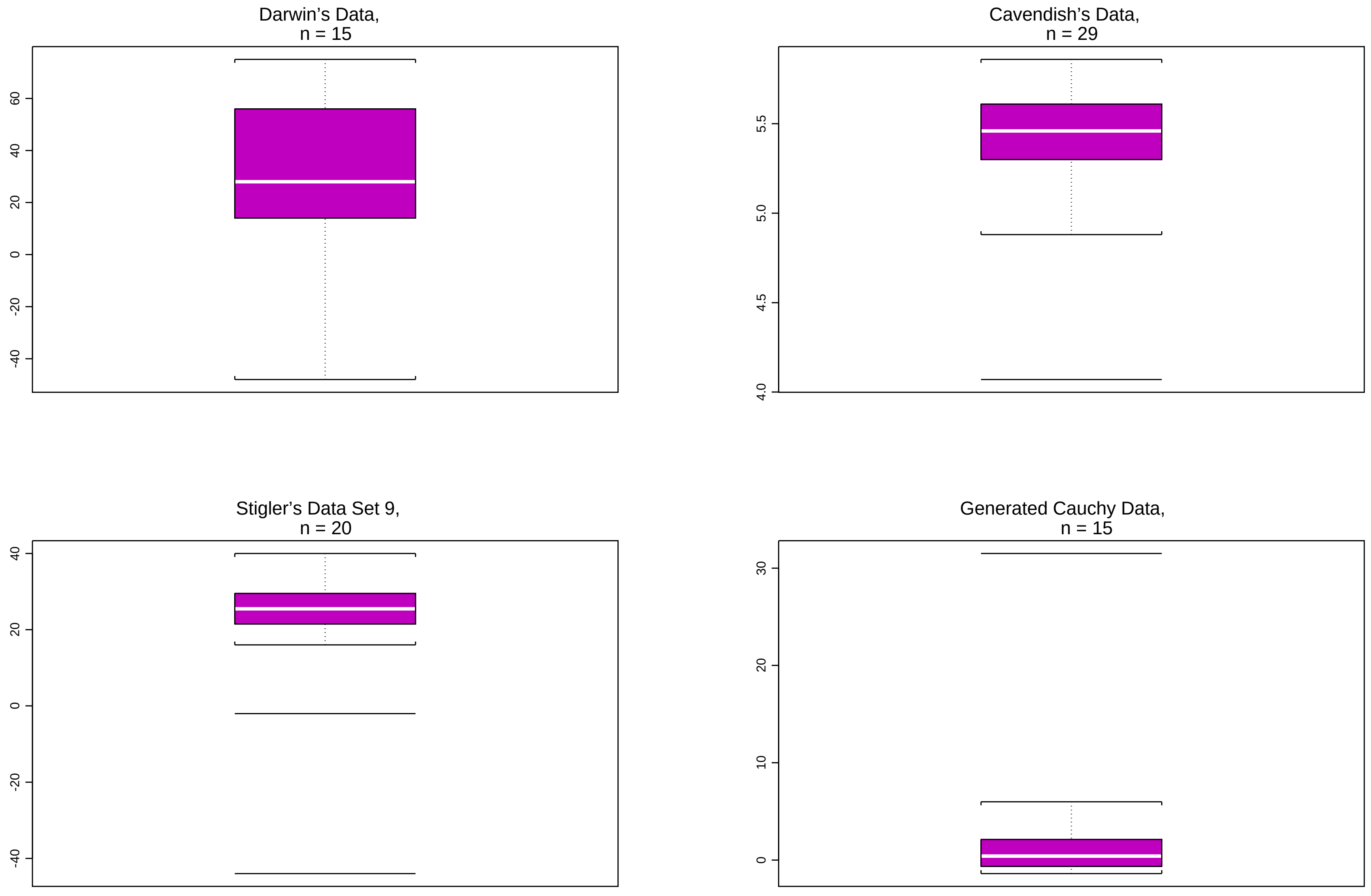
Marginal distributions, $m^N(\mathbf{x}|M)$, for these models can then be calculated in closed form.

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

1. Normal: $m^N(\mathbf{x}|M_N) = \dfrac{\Gamma((n-1)/2)}{(2\,\pi)^{(n-1)/2}\sqrt{n}\,\left(\sum_i (x_i-\bar{x})^2\right)^{(n-1)/2}}$
2. Uniform: $m^N(\mathbf{x}|M_U) = \dfrac{1}{n\,(n-1)(x_{(n)}-x_{(1)})^{\,n-1}}$
3. Cauchy: $m^N(\mathbf{x}|M_C)$ is given in Spiegelhalter (1985).
4. Left Exponential: $m^N(\mathbf{x}|M_L) = \dfrac{(n-2)!}{n^n(x_{(n)}-\bar{x})^{\,n-1}}$
5. Right Exponential: $m^N(\mathbf{x}|M_R) = \dfrac{(n-2)!}{n^n(\bar{x}-x_{(1)})^{\,n-1}}$

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 ($n = 14$)


Darwin's Data,
n = 15
60
40
20
0
-20
-40
Cavendish's Data,
n = 29
5.5
5.0
4.5
4.0
Stigler's Data Set 9,
n = 20
40
20
0
-20
-40
Generated Cauchy Data,
n = 15
30
20
10
0

The objective posterior probabilities of the five models, for each data set,

$$\bar{m}^N(\mathbf{x}|M) = \frac{m^N(\mathbf{x}|M)}{m^N(\mathbf{x}|M_N) + m^N(\mathbf{x}|M_U) + m^N(\mathbf{x}|M_C) + m^N(\mathbf{x}|M_L) + m^N(\mathbf{x}|M_R)},$$

are as follows:

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

II.B. Motivation for the Bayesian approach

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

2. **Prior information can be incorporated**, if desired. Indeed, if the published default posterior probability of H_0 is π^*, and *your* prior probability of H_0 is π_0, then *your* posterior probability of H_0 is

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

Example: In the psychokinesis example, $\pi^* = 0.92$.

A skeptic has $\pi_0 = .99$; hence $Pr(H_0|\text{data } x) = 0.999$

A believer has $\pi_0 = .1$; hence $Pr(H_0|\text{data } x) = 0.561$

3. **Consistency:**

   A. If one of the M_i is true, then $\bar{m}_i \to 1$ as $n \to \infty$

   B. If some other model, M^*, is true, $\bar{m}_i \to 1$ for that model closest to M^* in Kullback-Leibler divergence

   (Berk, 1966, Dmochowski, 1995)

4. **Ockham's razor:** Bayes Factors automatically seek parsimony; no adhoc penalties for model complexity are needed. (Jefferys and Berger, 1992)

5. **Frequentist interpretation:** 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.)

   See Sellke, Bayarri and Berger (2001) for discussion and earlier references.

6. **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 ξ is desired, it would be based on

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

(See Geisser, 1993, Draper, 1995, Raftery, Madigan and Hoeting, 1997, Clyde, 1999, ... for discussion.)

7. **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 1: In the location-scale example we choose among 5 non-nested models, with small sample sizes, irregular asymptotics, and no reduced sufficient statistics.

Example 2: Suppose the X_i are i.i.d from the $N(\theta, 1)$ density, where $\theta =$ "mean effect of T1 - mean effect of T2."

Standard Testing Formulation:

$H_0 : \theta = 0$ (no difference in treatments)
$H_1 : \theta \neq 0$ (a difference exists)

A More Revealing Formulation:

$H_0 : \theta = 0$ (no difference)
$H_1 : \theta < 0$ (Treatment 1 is better)
$H_2 : \theta > 0$ (Treatment 2 is better)

8. **Standard Bayesian 'ease of application' motivations.**

   (a) In sequential scenarios, there is no need to 'spend α' for looks at the data; posterior probabilities are not affected by the reason for stopping experimentation.

   (b) The distributions of known censoring variables are not needed for Bayesian analysis.

   (c) The multiple comparisons problem is automatically resolved by Bayesian analysis; as this is related to model selection and introduces an important concept, we give an example.

Multiple comparisons example

- Suppose $x_i \sim N(\mu_i, \sigma^2)$, $i = 1, \ldots, m$, are observed, with σ^2 known, and it is desired to determine which μ_i are nonzero.

- Let p denote the (assumed common) prior probability that μ_i is zero.

- Assume that the nonzero μ_i follow a $N(0, V)$ distribution, with V unknown.

- Assign p the uniform prior on $(0, 1)$ and V the prior density $\pi(V) = \sigma^2/(\sigma^2 + V)^2$.

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

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

- $(p_1, p_2, \ldots, p_m)$ 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 μ_i are zero and which are nonzero. There are 2^m models
- The $P(M_j \mid \mathbf{x})$ are not of much interest.
- The posterior probability that μ_i is nonzero is called the *posterior inclusion probability* for variable *i*, and can also be defined as

$$p_i \equiv \sum_{j\,:\,M_j \ni \text{variable } i} P(M_j \mid \mathbf{x}),$$

i.e., the overall posterior probability that variable *i* is in the model.

Part IIC. Difficulties with the Bayesian approach to model selection:

Difficulty 1: *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 cannot generally be used for model selection or hypothesis testing. (They 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.)

Normal example, known variance:

- *Data:* $X_1, X_2, \ldots, X_n$ are i.i.d $N(\theta, 1)$
- *To test:* $M_1 : \theta = 0$ versus $M_2 : \theta$ arbitrary
- *Marginal densities:*

$$\begin{aligned} m_1(\mathbf{x}) &= \frac{1}{(2\pi)^{n/2}} \exp\{-\frac{1}{2}\sum_{i=1}^{n} x_i^2\}, \\ m_2(\mathbf{x}) &= \int \frac{1}{(2\pi)^{n/2}} \exp\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2\}\pi_2(\theta)d\theta \end{aligned}$$

- *Usual objective prior* (under M_2) is

$$\pi_2^N(\theta) = c \quad \text{(any positive constant)}$$

and then

$$m_2(\mathbf{x}) = \frac{c}{\sqrt{n}(2\pi)^{(n-1)/2}} \exp\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2\}$$

- *Bayes factor* of M_2 to M_1 is

$$B_{21} = \frac{m_2(\mathbf{x})}{m_1(\mathbf{x})} = c\sqrt{\frac{2\pi}{n}} \exp\{\frac{n}{2}\,\bar{x}^2\}$$

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

2. Use of **vague proper priors** is even worse.

In the previous example, if $\pi(\theta) = N(\theta|0, \tau^2)$, then

$$\begin{aligned} B_{21} &= \frac{1}{\sqrt{1+n\tau^2}} \exp\{\frac{\tau^2 n^2 \bar{x}^2}{2(1+n\tau^2)}\} \\ &\approx \frac{1}{\tau\sqrt{n}} \exp\{\frac{n\bar{x}^2}{2}\} \quad \text{for large } \tau. \end{aligned}$$

Thus choosing a vague proper prior, i.e., choosing τ very large, gives the ridiculous answer that M_1 is favored, regardless of the data.

3. **Conjugate priors** can also be bad:

Consider the previous model, but with σ^2 unknown.

"Natural" Conjugate priors:

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

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

- 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 σ^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 $|t| \to \infty$, $B_{21} \to (n+1)^{-(n-1)/2} > 0$

For instance: if $n = 3$, $\quad B_{21} \to \frac{1}{4}$

if $n = 5$, $\quad B_{21} \to \frac{1}{36}$

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

Difficulty 2: *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:

$$
\begin{aligned}
M_1 : \; Y &= \beta_0 + \beta_1 X_1 + \sigma\,\epsilon \\
M_2 : \; Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \sigma\,\epsilon\,, \qquad \epsilon \sim N(0,1)
\end{aligned}
$$

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:

$$
\begin{aligned}
M_1 : \; Y &= \beta_0^{(1)} + \beta_1^{(1)} X_1 + \sigma^{(1)}\,\epsilon \\
M_2 : \; Y &= \beta_0^{(2)} + \beta_1^{(2)} X_1 + \beta_2^{(2)} X_2 + \sigma^{(2)}\,\epsilon\,.
\end{aligned}
$$

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

Difficulty 3: *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.

Difficulty 4: *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

Part III:

Methodologies for Bayesian model selection

- Overview of various approaches
- Training sample methods
- A case study focusing on use of conventional proper priors

IIIA. Overview of various approaches

(Berger and Pericchi, 2001)

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

IIIB. Training Sample Methods of Model Selection

(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, . . .)

- An example: median intrinsic posterior probabilities
- Fractional Bayes factors
- Expected posterior priors

Example: *median intrinsic posterior probabilities*

(modelled after Berger and Pericchi, 1998)

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

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

Standard objective prior:

$Pr(M_1) = Pr(M_2) = \frac{1}{2}$;

Under M_2, $\pi_2(\theta) = 1$

(Formal) Bayes factor:

$$B_{21} = \frac{f(\mathbf{x}|0)}{\int f(\mathbf{x}|\theta)\,(1)d\theta} = \frac{1}{\sqrt{n}}\, e^{-\frac{n}{2}\bar{x}^2}$$

(Formal) posterior probability of M_1:

$$P_1 = \frac{1}{1 + B_{21}^{-1}} = (1 + \sqrt{n}\, e^{-\frac{n}{2}\bar{x}^2})^{-1}$$

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

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

Use on the remaining data, $\mathbf{x}^{(i)} \equiv (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$, to compute the posterior model probabilities

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

Find the median intrinsic posterior probability:

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

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

as the recommended conventional posterior probabilities of M_1 and M_2.

General Algorithm:

- Begin with standard objective priors, π_i^N, for the parameters θ_i in the model M_i ($\pi_i^N(\theta_i) = 1$ is okay).
- Define a "minimal training sample," $\mathbf{x}^* = (x_1^*, \ldots, x_l^*)$, 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 posterior probabilities of all models using the $\pi_i^N(\theta_i|\mathbf{x}^*)$ as priors, for the remaining data.
- Do this for every possible minimal training sample, $\mathbf{x}^*$, and take the median of the results.

Formula:

$$
\begin{aligned}
P_i^{\text{med}} &= \text{`intrinsic median' posterior probability of} M_i \\
&= \text{Median } \left\{ \frac{m_i^N(\mathbf{x})}{m_i^N(\mathbf{x}^*) \sum_{j=1}^{m} (m_j^N(\mathbf{x})/m_j^N(\mathbf{x}^*))} \right\},
\end{aligned}
$$

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

Note: When the number of possible training samples is large, one need only sample from them 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 ln (sets) of training samples.

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.

Notation: Model {1,3,4} means

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

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 σ^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
{1,2,3,4}	0.049
{1,2,3}	0.171
{1,2,4}	0.190
{1,3,4}	0.160
{2,3,4}	0.041
{1,2}	0.276
{1,4}	0.108
{3,4}	0.004
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

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

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

- Thus the median probability model is $\{X_1, X_2, X_4\}$.

- 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 prior
$$\pi_j^*(\boldsymbol{\theta}_j) \propto [f_j(\boldsymbol{x}|\boldsymbol{\theta}_j)]^b \cdot \pi_j^N(\boldsymbol{\theta}_j)$$
- Compute Bayes factors using these priors and the "remaining likelihoods":

$$B_{ji}^{FBF} = \frac{\int [f_j(\boldsymbol{x}|\boldsymbol{\theta}_j)]^{(1-b)}\,\pi_j^*(\boldsymbol{\theta}_j)d\boldsymbol{\theta}_j}{\int [f_i(\boldsymbol{x}|\boldsymbol{\theta}_i)]^{(1-b)}\,\pi_i^*(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i}$$

$$= B_{ji}^N \cdot \frac{\int [f_i(\boldsymbol{x}|\boldsymbol{\theta}_i)]^b\,\pi_i^N(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i}{\int [f_j(\boldsymbol{x}|\boldsymbol{\theta}_j)]^b\,\pi_j^N(\boldsymbol{\theta}_j)d\boldsymbol{\theta}_j}$$

- Computationally often comparatively simple.
- Broadly applicable (except for serious concerns in irregular problems).
- Specification of different fractions, b, for different parts of the likelihood can be necessary.

Expected Posterior Priors

(Perez, 1998, Perez and Berger, 2001, 2002, Neal, 2002)

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

Initial marginals: $m_i^N(\boldsymbol{y}) = \int f_i(\boldsymbol{y}|\boldsymbol{\theta}_i)\pi_i^N(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i$.

Training sample posteriors: Consider a training sample, $\boldsymbol{y}^*$, such that the posterior distributions

$$\pi_i^N(\boldsymbol{\theta}_i|\boldsymbol{y}^*) = \frac{f_i(\boldsymbol{y}^*|\boldsymbol{\theta}_i)\pi_i^N(\boldsymbol{\theta}_i)}{m_i^N(\boldsymbol{y}^*)}$$

exist, for $i = 1, \ldots, k$.

Definition: The prior densities

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

where $\boldsymbol{y}^*_{(i)}$ is a minimal random subsample of $\boldsymbol{y}^*$ such that the $\pi_i^N(\boldsymbol{\theta}_i|\boldsymbol{y}^*_{(i)})$ exist, will be called the *expected posterior priors* (or EP priors) for the $\boldsymbol{\theta}_i$, with respect to m^*.

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.

Choosing m^* to be the empirical distribution:
Given observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, let

$$m^*(\boldsymbol{y}^*) = \frac{1}{L} \sum_{l} I_{\{\boldsymbol{y}(l)\}}(\boldsymbol{y}^*),$$

where $\boldsymbol{y}(l) = (y_{l_1}, \ldots, y_{l_m})$ is a subsample of size $0 < m < n$ such that $\pi_i^N(\boldsymbol{\theta}_i|\boldsymbol{y}(l))$ exists for all models M_i, and L is the number of such subsamples of size m.

Computation: Introduce $\boldsymbol{y}^*$ as latent variables, effectively replacing $\pi_i^*(\boldsymbol{\theta}_i) = \int \pi_i^N(\boldsymbol{\theta}_i|\boldsymbol{y}^*_{(i)})m^*(\boldsymbol{y}^*)d\boldsymbol{y}^*$ by

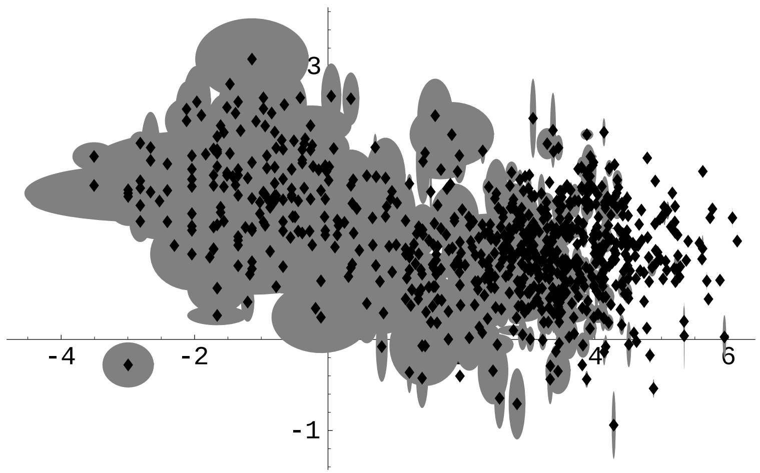
$$\pi_i^N(\boldsymbol{\theta}_i|\boldsymbol{y}^*_{(i)})m^*(\boldsymbol{y}^*) = \frac{\pi_i^N(\boldsymbol{\theta}_i)f_i(\boldsymbol{y}^*_{(i)}|\boldsymbol{\theta}_i)}{m_i^N(\boldsymbol{y}^*_{(i)})}m^*(\boldsymbol{y}^*)\,.$$

BATSE gamma ray burst data set: We analyze 745 measurements taken by the Compton Gamma Ray Observatory between 1991 and 1994. Of interest is the relationship of the duration (T90) and hardness ratio (HR) of the bursts.

Model: Given the *true* gamma ray values $\boldsymbol{y}_i = (y_{i1}, y_{i2})$, the ith vector observation $\boldsymbol{x}_i = (x_{i1}, x_{i2}) = (\mathsf{log(T90)}_i, \mathsf{log(HR)}_i)$ is modeled as

$$\boldsymbol{x}_i \sim N_2 \left(\boldsymbol{y}_i, \begin{pmatrix} \sigma_{i1}^2 & 0 \\ 0 & \sigma_{i2}^2 \end{pmatrix} \right),$$

with the σ_{ij}^2 being known error variances.


3
-4
-2
4
6
-1

Prior modeling:

- The true gamma ray values, $\boldsymbol{y}_i = (y_{i1}, y_{i2})$, are modeled by a mixture of k bivariate normal distributions,

$$\boldsymbol{y}_i \sim \sum_{j=1}^{k} w_j N_2(\boldsymbol{\mu}_j, \Sigma_j), \quad i = 1, \ldots, 745,$$

  where the mixture weights w_j sum to one.

- $\pi(w_1, \ldots, w_k) \propto 1$ is the prior density for the weights.

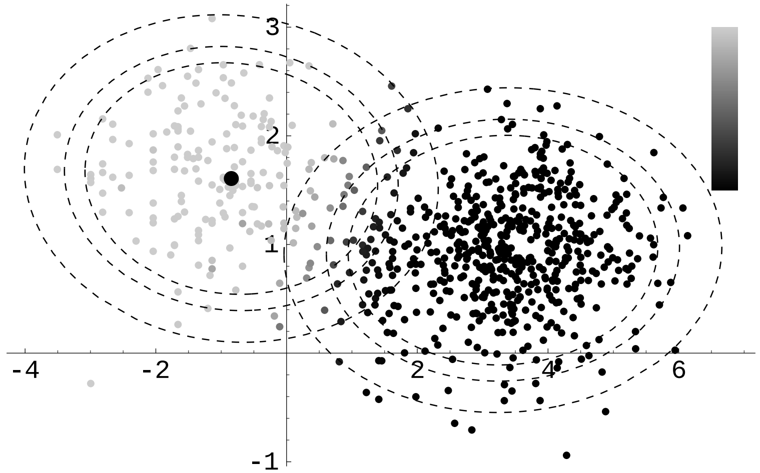
- Expected posterior priors were used for the component parameters $(\boldsymbol{\mu}_j, \Sigma_j)$, based on the initial noninformative component prior distributions $\pi^N(\boldsymbol{\mu}_j, \Sigma_j) = |\Sigma_j|^{-1/2}$. (Also proposed by Moreno and Liseo, 1998.)

Computation: A reversible jump MCMC scheme was used, with the $\mathbf{y}^*$ introduced as latent variables. Identifiability problems were addressed by ordering $\boldsymbol{\mu}$ on the first coordinate.

Results:

- The posterior probability that there are only $k = 2$ components in the mixture is essentially one.
- Posterior estimates of the component parameters were found.
- Individual burst classification probabilities were found.

Component 1		
$\hat{w}_1 = 0.2440$		
$\hat{\mu}_{T90}$= -0.8451	$\hat{\mu}_{HR}$= 1.6071	
$\hat{\sigma}_{T90}$= 1.0431	$\hat{\sigma}_{HR}$= 0.4966	$\hat{\rho}$ =-0.03098
Component 2		
$\hat{w}_2 = 0.7560$		
$\hat{\mu}_{T90}$ = 3.3103	$\hat{\mu}_{HR}$ = 0.9478	
$\hat{\sigma}_{T90}$ =1.1036	$\hat{\sigma}_{HR}$ = 0.4924	$\hat{\rho}$ =0.01358



3
2
1
-1
-4
-2
2
4
6

An aside: using EPPriors subjectively

- Subjectively elicit a predictive $m(\mathbf{x}^*)$, where $\mathbf{x}^*$ is a minimal training sample (Kadane, 1980).
- 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 π^* 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!

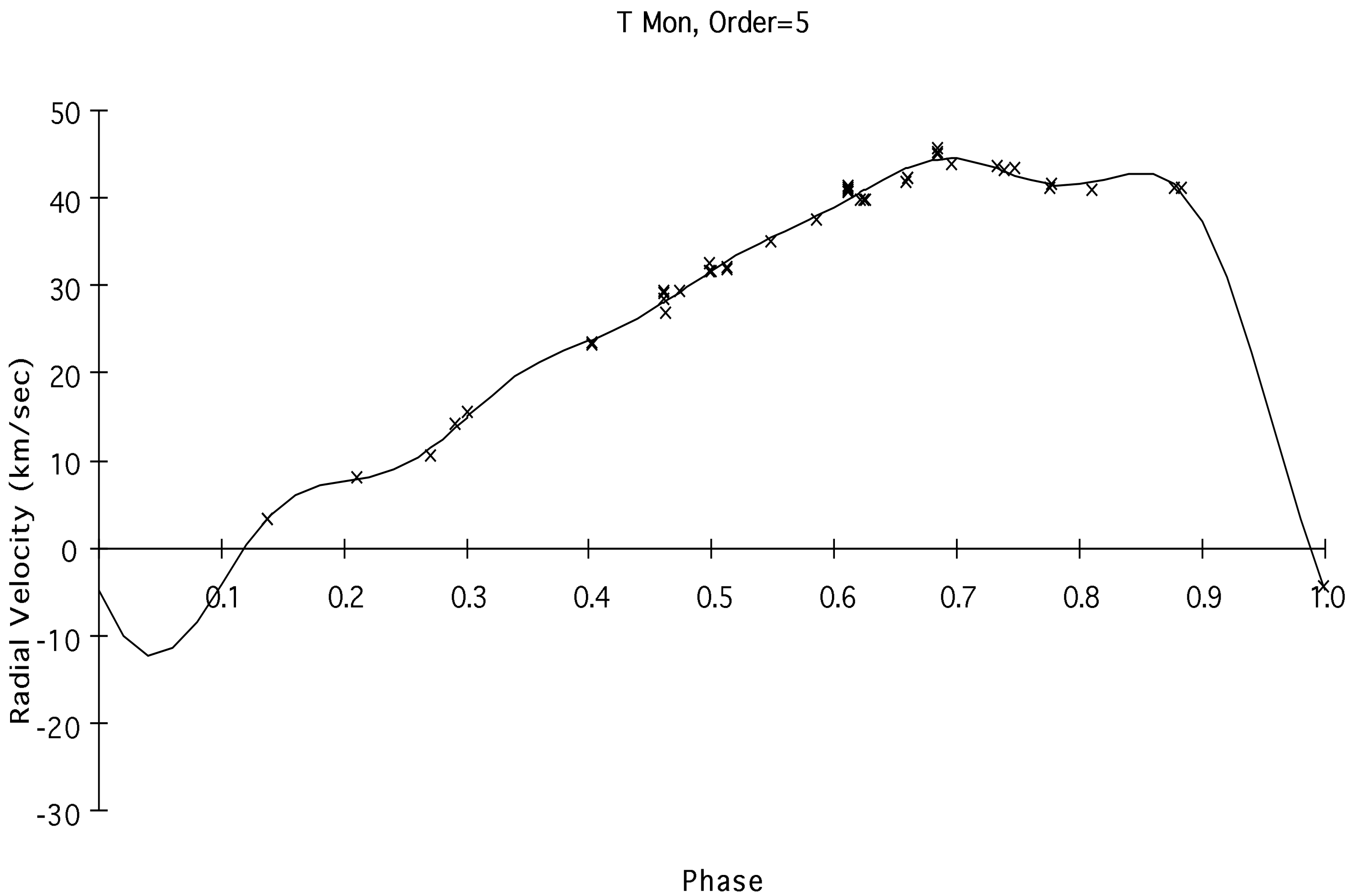
IIIC. A case study involving the conventional proper prior approach: *Bayesian Model Selection and Analysis for Cepheid Star Oscillations*

(Berger, Jefferys, Müller, and Barnes, 2001)

- The astronomical problem
- Bayesian model selection
- The data and likelihood function
- Choice of prior distributions
- Computation and results

The Astronomical Problem

- A Cepheid star pulsates, regularly varying its luminosity (light output) and size.
- From the Doppler shift as the star surface pulsates, one can compute surface velocities at certain phases of the star's period, thus learning how the radius of the star changes.
- From the luminosity and 'color' of the star, one can learn about the angular radius of the star (the angle from Earth to opposite edges of the star).
- Combining, allows estimation of s, a star's distance.


T Mon, Order=5
Radial Velocity (km/sec)
50
40
30
20
10
0
-10
-20
-30
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Phase

Curve Fitting

To determine the overall change in radius of the star over the star's period, the surface velocity must be estimated at phases other than those actually observed, leading to a curve fitting problem (also for luminosity). Difficulties:

- Observations have measurement error.
- Phases at which observations are made are unequally spaced.
- 100 possible models (curve fits) are entertained.
- Resulting models have from 10 to 50 parameters.

Other Statistical Difficulties

- All uncertainties (measurement errors, model uncertainty and estimated curve inaccuracy) need to be accounted for.
- There are significant nonlinear features of the model.
- Prior information, such as understanding of the Lutz-Kelker bias, needs to be incorporated into the analysis.

Bayesian Analysis with Uncertain Models

- The data, $\boldsymbol{Y}$, arises from one of the models $M_1, \ldots, M_q$.
- Under M_i, the density of $\boldsymbol{Y}$ is $f_i(\boldsymbol{y} \mid \boldsymbol{\theta}_i)$.
- $\boldsymbol{\theta}_i$ is an unknown vector of parameters under M_i.
- Specify prior probabilities of models (usually $1/q$).
- Prior distributions $\pi_i(\boldsymbol{\theta}_i)$ for the $\boldsymbol{\theta}_i$ are specified.
- Bayes theorem then gives the model posterior probabilities (and posterior distributions of the other unknowns).

Model Averaging

- Suppose Models 5, 6, and 7 have posterior probabilities 0.34, 0.56 and 0.10, respectively.
- Model uncertainty is then handled by *Bayesian model averaging:* if Models 5, 6, and 7 provided distance estimates of 750, 790, and 800 parsecs, the 'model-averaged' distance estimate would be $(0.34)\,750 + (0.56)\,790 + (0.10)\,800 = 777.4$ parsecs.
- The 'model-averaged' variance can also be computed; it can be several times that of a single model.

The Data and Statistical Model

- Data:
  - m observed radial velocities U_i, $i = 1, \ldots, m$.
  - n vectors of photometry data consisting of luminosity V_i, $i = 1, \ldots, n$, and color index C_i, $i = 1, \ldots, n$.
- Specified standard deviations σ_{U_i}, σ_{V_i}, and σ_{C_i}; unknown adjustment factors are inserted, leading to variances $\sigma^2_{U_i}/\tau_u$, $\sigma^2_{V_i}/\tau_v$, $\sigma^2_{C_i}/\tau_c$.
- The statistical model for measurement error:

$$U_i \sim N(u_i, \sigma^2_{U_i}/\tau_u),$$

$$V_i \sim N(v_i, \sigma^2_{V_i}/\tau_v),$$
$$C_i \sim N(c_i, \sigma^2_{C_i}/\tau_c),$$

where u_i, v_i, and c_i denote the true unknown mean velocity, luminosity, and color index.

- Define $\boldsymbol{G}_u$ and $\boldsymbol{G}_v$ to be the known diagonal matrices of the variances $\sigma^2_{U_i}$ and $\sigma^2_{V_i}$.

Curve Fitting I: Fourier Analysis

- Model the periodic velocities, u, as trigonometric polynomials. For the velocity u at phase ϕ,

$$u = u_0 + \sum_{j=1}^{M} [\theta_{1j} \cos(j\phi) + \theta_{2j} \sin(j\phi)],$$

  where u_0 is the mean velocity and M is the (unknown) order of the trigonometric polynomial.

- There is a similar polynomial model for luminosity, v, having unknown order N.

Statistically, these trigonometric polynomial models can be written as the linear models

$$\boldsymbol{U} = u_0\mathbf{1} + \boldsymbol{X}_u\boldsymbol{\theta}_u + \boldsymbol{\varepsilon}_u \text{ and } \boldsymbol{V} = v_0\mathbf{1} + \boldsymbol{X}_v\boldsymbol{\theta}_v + \boldsymbol{\varepsilon}_v,$$

- u_0 and v_0 are the (unknown) mean velocity and luminosity and $\mathbf{1}$ is the column vector of ones.
- $\boldsymbol{X}_u$ and $\boldsymbol{X}_v$ are matrices of the trigonometric covariates (e.g., terms like $\sin(j\phi)$);
- $\boldsymbol{\theta}_u$ and $\boldsymbol{\theta}_v$ are the unknown Fourier coefficients;
- $\boldsymbol{\varepsilon}_u$ and $\boldsymbol{\varepsilon}_v$ are independently multivariate normal errors: $\boldsymbol{N}(\mathbf{0}, \boldsymbol{G}_u/\tau_u)$ and $\boldsymbol{N}(\mathbf{0}, \boldsymbol{G}_v/\tau_v)$.

Color need not be modeled separately because it is related to luminosity (v) and velocity (or change in radius) by

$$c_i = a[0.1v_i - b + 0.5 \log(\phi_0 + \Delta r_i/s)],$$

where a and b are known constants, ϕ_0 and s are the angular size and distance of the star, and Δr, the change in radius corresponding to phase ϕ, is given by

$$\Delta r = -g \sum_{j=1}^{M} \frac{1}{j} [\theta_{1j} \sin(j(\phi - \Delta\phi)) - \theta_{2j} \cos(j(\phi - \Delta\phi))],$$

with ‘phase shift’ $\Delta\phi$ and g a known constant.

Objective Priors with Uncertain Models

- For estimation within one model, objective prior distributions are readily available (Jeffreys-rule priors, maximum entropy priors, reference priors).
  Example: If θ is an unknown normal mean, use $\pi(\theta) = 1$.

- These are typically improper (integrate to infinity); this is not a problem for estimation, but for testing and model selection is usually inappropriate.

- Various guidelines for choice of priors in model selection have been given: here are three we need.

Common Model Parameters

- When all models have 'common' parameters, they can be assigned the usual improper objective prior.
  Example: All Cepheid radial velocity models have common mean level u_0; it is okay to assign the usual objective (improper) prior $\pi(u_0) = 1$.
- This is generally true if models have the same 'location parameters' (as above) or 'scale parameters' (as in a normal variance).
- For discussion of other types of 'common parameters', see Berger and Pericchi (2001).

Priors for General Linear Models:

$$\boldsymbol{Y} = \theta_0 \mathbf{1} + \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

- $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ are observations (e.g., radial velocities in the Cepheid problem);
- $\boldsymbol{X}$ is the matrix of covariates (e.g., terms like $\sin(j\phi)$ from trigonometric polynomials);
- $\boldsymbol{\theta}$ is an unknown vector (e.g., Fourier coefficients);
- $\mathbf{1}$ is a vector of ones and θ_0 the unknown mean level;
- $\boldsymbol{\varepsilon}$ is $\boldsymbol{N}(\mathbf{0}, \sigma^2 \boldsymbol{G})$, where $\boldsymbol{G}$ (e.g., the diagonal matrix of measurement variances) is known.

- The recommended prior (from Zellner and Siow, 1980)
  - for θ_0 is $\pi(\theta_0) = 1$;
  - for $\boldsymbol{\theta}$, given σ^2, is Cauchy$(\mathbf{0}, n\sigma^2(\boldsymbol{X}'\boldsymbol{G}^{-1}\boldsymbol{X})^{-1})$; for computational convenience, this can be written in two stages as

$$\pi(\boldsymbol{\theta} \mid \sigma^2, \tau) \text{ is } \boldsymbol{N}(\mathbf{0}, \tau n\sigma^2(\boldsymbol{X}'\boldsymbol{G}^{-1}\boldsymbol{X})^{-1});$$

$$\pi(\tau) = \frac{1}{\sqrt{2\pi}\,\tau^{3/2}} \exp(-\frac{1}{2\tau}).$$

Choice of Cepheid Prior Distributions

- The orders of the trigonometric polynomials, (M, N), are given a uniform distribution up to some cut-off (e.g., $(10, 10)$).

- τ_u, τ_v, τ_c, which adjust the measurement standard errors, are given the standard objective priors for 'scale parameters,' namely the Jeffreys-rule priors $\pi(\tau_u) = \frac{1}{\tau_u}$, $\pi(\tau_v) = \frac{1}{\tau_v}$, and $\pi(\tau_c) = \frac{1}{\tau_c}$.

- The mean velocity and luminosity, u_0 and v_0, are 'location parameters' and so can be assigned the standard objective priors $\pi(u_0) = 1$ and $\pi(v_0) = 1$.

- The angular diameter ϕ_0 and the unknown phase shift $\Delta\phi$ are also assigned the objective priors $\pi(\Delta\phi) = 1$ and $\pi(\phi_0) = 1$. It is unclear if these are 'optimal' objective priors but the choice was found to have neglible impact on the answers.
- The Fourier coefficients, $\boldsymbol{\theta}_u$ and $\boldsymbol{\theta}_v$, occur in linear models, so the Zellner-Siow priors can be utilized.

- The prior for distance s of the star should account for
  - *Lutz-Kelker bias*: a uniform spatial distribution of Cepheid stars would yield a prior proportional to s^2.
  - The distribution of Cepheids is flattened wrt the galactic plane; we use an exponential distribution.

So, we use $\pi(s) \propto s^2 \exp(-|s \sin\beta|/z_0)$,

- β being the known galactic latitude of the star (its angle above the galactic plane),
- z_0 being the 'scale height,' assigned a uniform prior over the range $z_0 = 97 \pm 7$ parsecs.

Computation

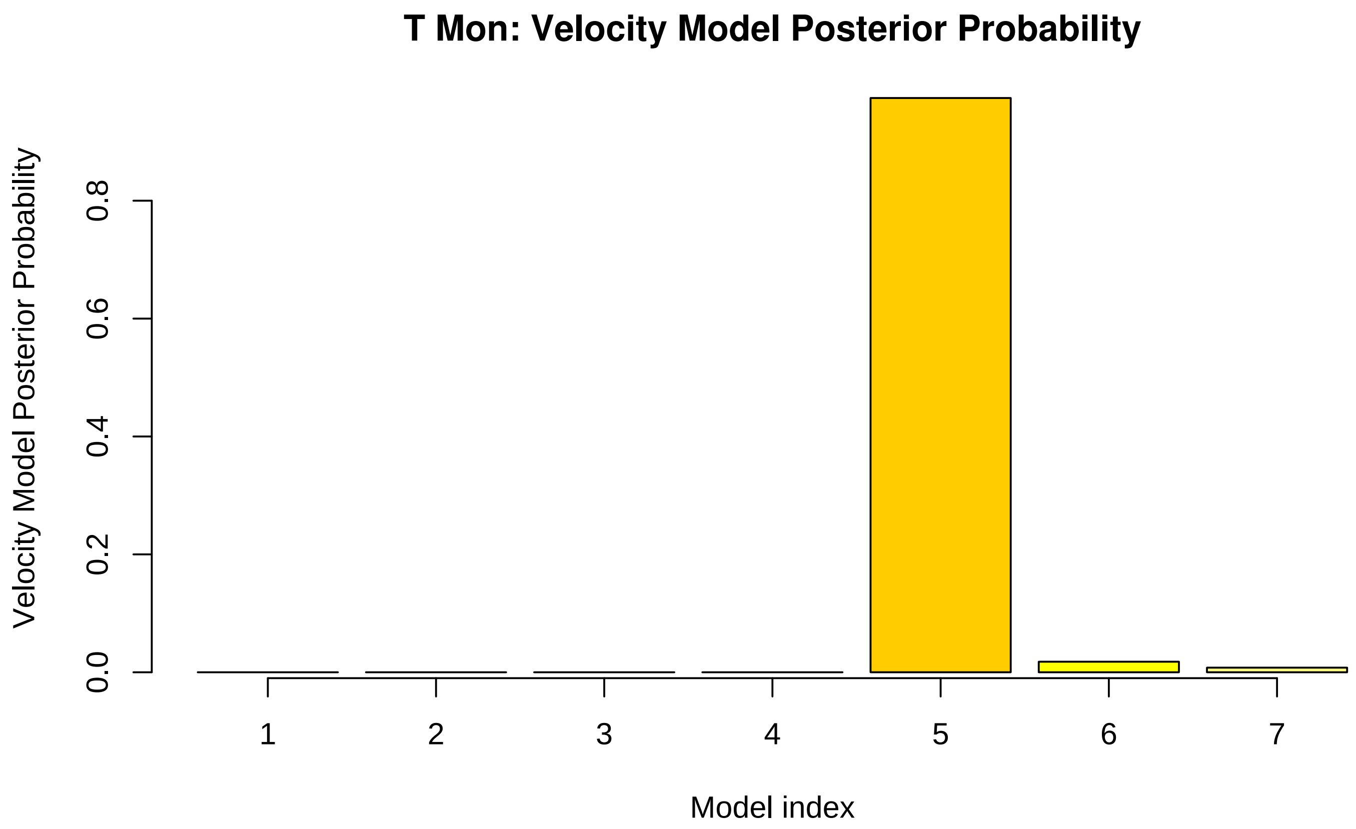
A reversible-jump MCMC algorithm of the type reviewed in Dellaportas, Forster and Ntzoufras (2000) is used to move between models and generate posterior distributions and estimates.

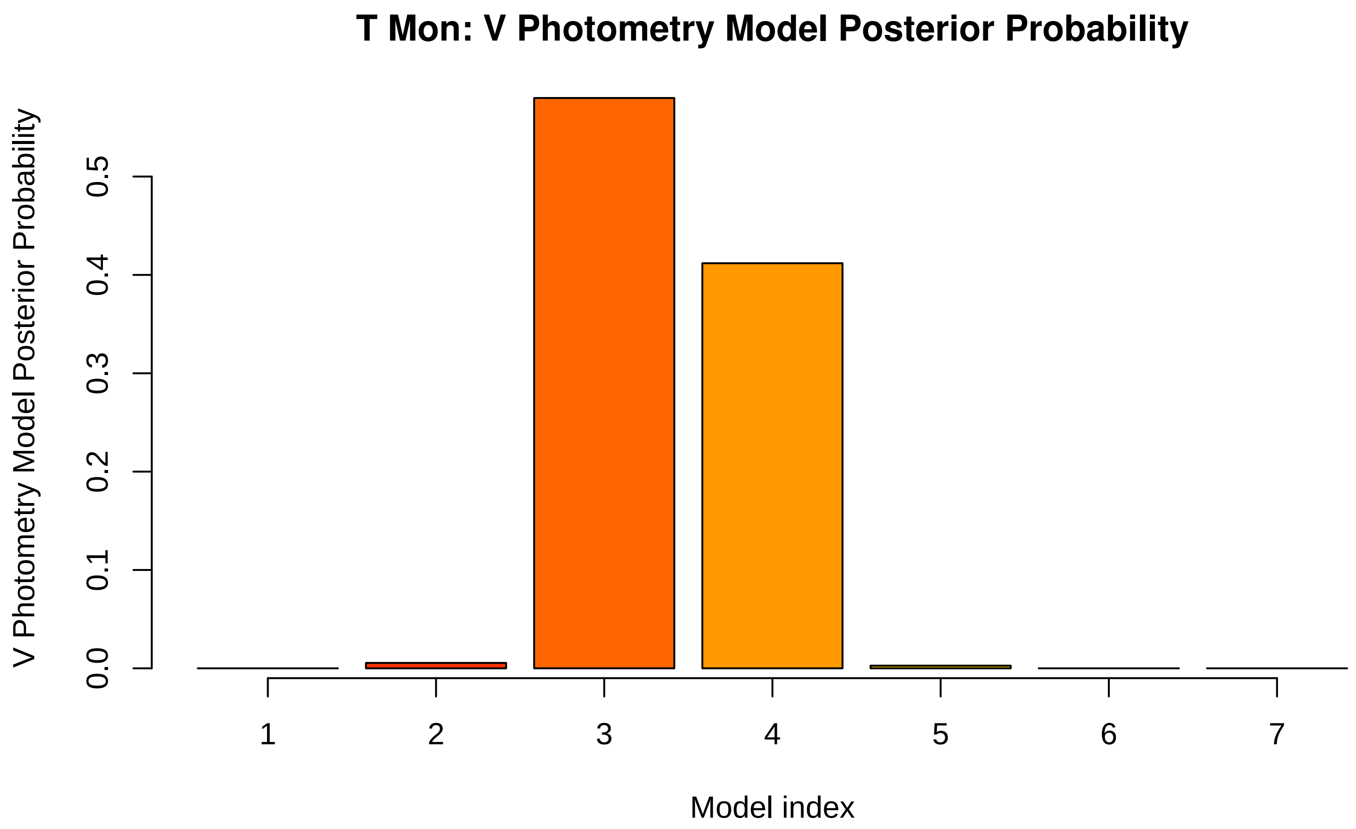
- The full conditional distributions for the variance and precision parameters and hyperparameters are standard gamma and inverse-gamma distributions and are sampled with Gibbs sampling.

- For $\Delta\phi$, ϕ_0 and s, we employ a random-walk Metropolis algorithm using, as the proposal distribution, a multivariate normal distribution centered on the current values and with a covariance matrix found from linearizing the problem for these three parameters.

- The Fourier coefficients $\boldsymbol{\theta}_u$ and $\boldsymbol{\theta}_v$, as well as u_0 and v_0, are also sampled via Metropolis. The natural proposal distributions are found by combining the normal likelihoods with the normal part of the Zellner-Siow priors, leading to conjugate normal posterior distributions.

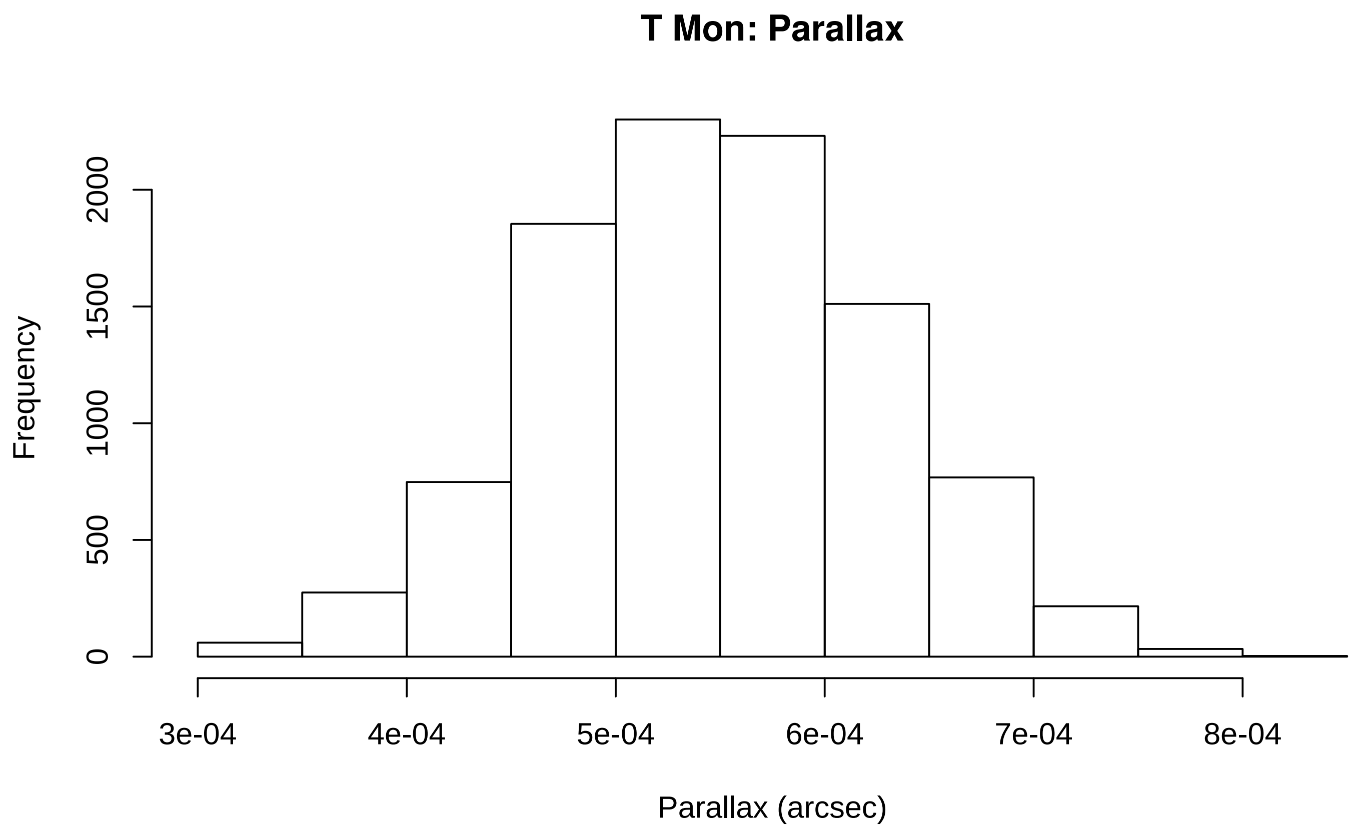
- Proposal for moves between models:
  - A ‘burn-in’ portion of the MCMC with uniform model proposals yielded initial posterior model probabilities, which were then used as the proposal for subsequent model moves.
  - Simultaneously, new values were proposed for the Fourier coefficients.
- The proposal distributions listed above lead to a well-mixed Markov chain, so that only 10,000 iterations of the MCMC computation needed to be performed.

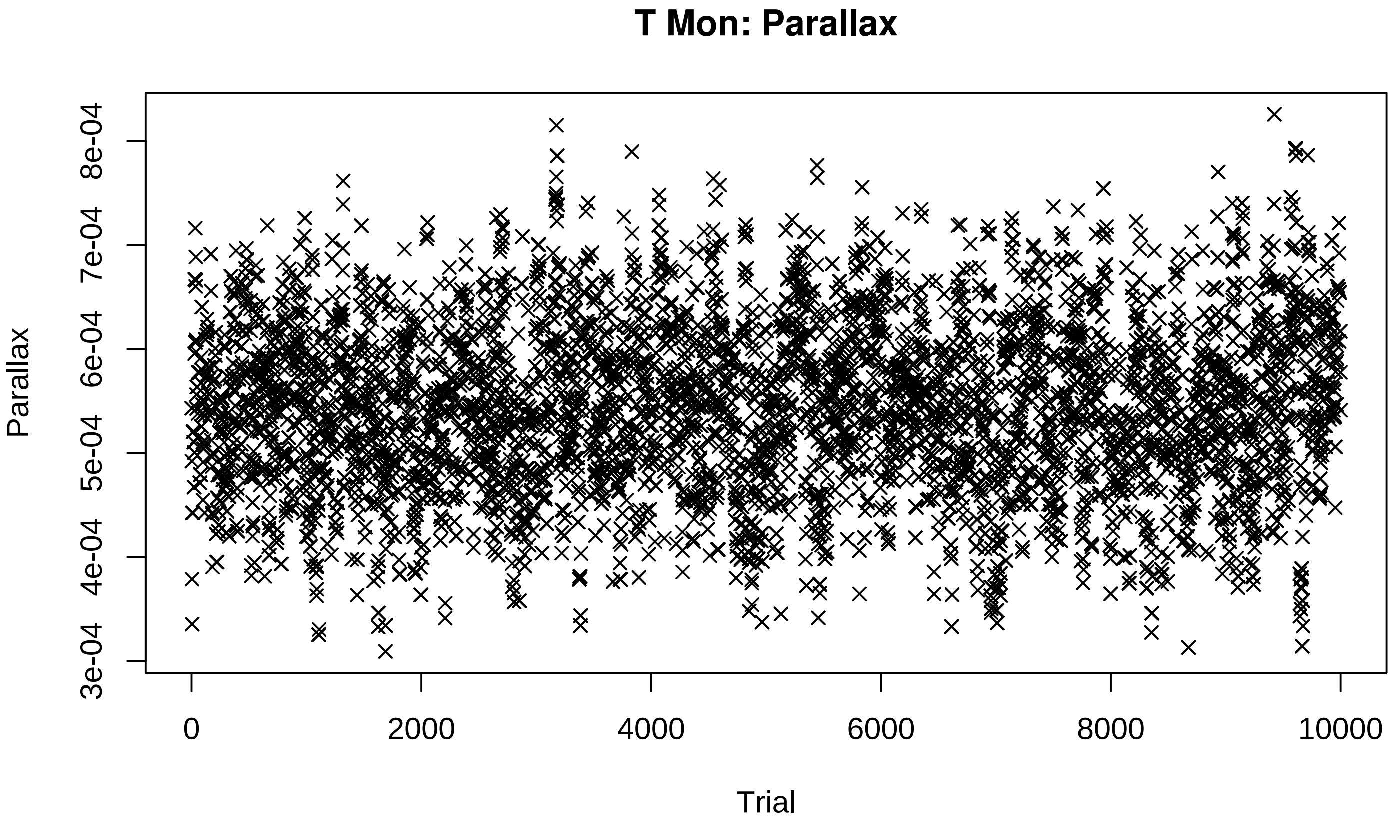

T Mon: Velocity Model Posterior Probability
Velocity Model Posterior Probability
0.0
0.2
0.4
0.6
0.8
1
2
3
4
5
6
7
Model index


T Mon: V Photometry Model Posterior Probability
V Photometry Model Posterior Probability
0.0
0.1
0.2
0.3
0.4
0.5
1
2
3
4
5
6
7
Model index

Results for other parameters

- Estimates, standard errors, etc., for any of the unknowns or parameters in the model are also available from the MCMC computation.
- The MCMC computational strategy discussed above will automatically perform 'model-averaging' over these models, when computing posterior quantities of interest.
- The simulation history of a parameter can be graphed to check the mixing of the Markov chain (and thus assess the validity of the computation).


T Mon: Parallax
Frequency
0
500
1000
1500
2000
3e-04
4e-04
5e-04
6e-04
7e-04
8e-04
Parallax (arcsec)


T Mon: Parallax
Parallax
3e-04
4e-04
5e-04
6e-04
7e-04
8e-04
0
2000
4000
6000
8000
10000
Trial