Tutorial on Bayesian Model and Variable Selection

Ioannis Ntzoufras
Department of Business Administration
University of the Aegean

Contents

1 Introduction to Model Selection

What is Model Selection?

- Evaluation of performance of scientific scenarios and
- Selection of the 'best'.

'Best' Model?

- The 'best' performed model is totally subjective
- Different procedures (or scientists) support different scientific theories, scenarios and models.

Two **MAJOR** principles:

1. *Goodness of Fit*
   How close is theory [model] to reality [data].
2. *Parsimony*
   Simplicity of theory;
   In stats: Economy in parameters.

Available Methods

- Classical Model Selection: Significance Tests and Stepwise Methods:
  (Forward Strategy, Backward Elimination, Stepwise Procedures).
- Bayesian Model Selection
  - Posterior odds and posterior model probabilities.
  - Utility measures.
  - Predictive criteria.
- Model Selection Criteria
  - Akaike Information Criterion (AIC).
  - Bayes Information Criterion (BIC).
  - Other criteria.

Disadvantages of Classical Stepwise Procedures

- Large datasets we observe small p-values even if the hypothesized model is plausible.
- Exact significance level cannot be calculated since stepwise methods are sequential application of simple significance tests (Freedman, 1983).
- The maximum F-to-enter statistic 'is not even remotely like an F-distribution' (Miller, 1984).
- The selection of a single model ignores model uncertainty.
- We can compare only nested models.
- Different models are selected if we use different procedures or start from different models.

Bayesian Model Selection

Bayesian model selection is based on

1. Posterior odds of model m_0 versus model m_1 given by

$$PO_{01} = \frac{f(m_0|\underline{y})}{f(m_1|\underline{y})} = \underbrace{\frac{f(\underline{y}|m_0)}{f(\underline{y}|m_1)}}_{Bayes\ Factor} \times \underbrace{\frac{f(m_0)}{f(m_1)}}_{Prior\ Odds} .$$

2. Posterior probabilities given by

$$f(m|\underline{y}) = \frac{f(\underline{y}|m)f(m)}{\sum\limits_{m_l \in \mathcal{M}} f(\underline{y}|m_l)f(m_l)} = \left(\sum_{m_l \in \mathcal{M}} PO_{m_l,m} \right)^{-1},$$

- $\mathcal{M}$: set of models under consideration,
- $\sum_{m' \in \mathcal{M}} f(m|\underline{y}) = 1$.

$log_{10}(B_{10})$	B_{10}	Evidence against H_0
0.0 to 0.5	1.0 to 3.2	Not worth than a bare mention
0.5 to 1.0	3.2 to 10	Substantial
1.0 to 2.0	10 to 100	Strong
greater than 2	greater than 100	Decisive

Table 1: Bayes Factor Interpretation according to Kass and Raftery (log of 10).

$ln(B_{10})$	B_{10}	Evidence against H_0
0 to 2	1 to 3	Not worth than a bare mention
2 to 5	3 to 12	Positive
5 to 10	12 to 150	Strong
greater than 10	greater than 150	Decisive

Table 2: Bayes Factor Interpretation according to Kass and Raftery (Natural logarithm).

Bayesian Model Averaging

- Adjust predictions (and inference) according to the observed model uncertainty.
- Average over all conditional model specific posterior distributions. weighted by their posterior model probabilities.
- Base predictions on all models under consideration and therefore account for model uncertainty.
- Predictive distribution of a quantity Δ

$$f(\Delta|\underline{y}) = \sum_{m \in \mathcal{M}} f(\Delta|m,\underline{y})f(m|\underline{y})$$

Wasserman (1997) and Hoeting *et al*. (1998) recently provided two well written papers that both review Bayesian model averaging.

Logarithmic scoring rule (LS): Measure of the predictive performance of a model m given by

$$LS_m = -E\{log[f(\Delta|m,\underline{y})]\}$$

and by

$$LS = -E\left\{ log\left[\sum_{m \in \mathcal{M}} f(\Delta|m,\underline{y})f(m|\underline{y}) \right] \right\}$$

for Bayesian model averaging. Lower values of the logarithmic scoring rule indicate better predictive power.

Bayesian model averaging method: better predictive ability since $LS \leq LS_m$, $\forall\ m \in \mathcal{M}$; see Madigan and Raftery (1994), Kass and Raftery (1995) and Raftery *et al*. (1997).

2 Model Selection via Markov Chain Monte Carlo Methods

Problems in Bayesian model selection:

- Integrals involved in $f(m|\underline{y})$ and
- Size of $\mathcal{M}$.

Hence, MCMC methods become an extremely attractive alternative.

Descriprion:

- Generate sample $(m^{(t')}, \underline{\beta}^{(t')},\ t' = 1, \ldots, t)$
- Estimate posterior model probabilities by

$$\hat{f}(m) = \frac{1}{t} \sum_{t'=1}^{t} I(m^{(t')} = m) \qquad m \in \mathcal{M}$$

$I(\cdot)$: Indicator function.

- Get samples from $f(\underline{\beta}_{(m)} | m, \underline{y})$ (automatically available).

Why use MCMC in Model selection:

- Automatic after defining the prior distribution,
- Cannot explore the model space otherwise,
- Integrals involved are intractable.
- Bayesian model averaging is straightforward.

- General Model Selection Algorithms
  - Reversible jump (Green, 1995, Bk).
  - Carlin and Chib (1995, JRSS B) Gibbs sampler.
  - Markov chain Monte Carlo model composition [MC^3] (Madigan and York, 1995, I.S.R.).
  - *Metropolised Carlin and Chib Algorithm (Dellaportas et al. , 2002, Stats & Comp.)*
- Variable selection samplers
  - Stochastic Search Variable Selection [SSVS] (George and McCulloch, 1993, JASA).
  - Kuo and Mallick (1998, Sank, B) Gibbs sampler.
  - *Gibbs Variable Selection (Dellaportas et al. , 2000,2002).*

- Fast variable selection algorithms for normal models
  - Clyde *et al.* (1996).
  - Smith and Kohn (1996).
  - Clyde (1998).

2.1 General Model Selection Methods

2.1.1 Reversible Jump

The Procedure

If the current state of the Markov chain is $\left(\underline{\beta}_{(m)}, m\right)$, then

- Generate $\underline{\beta}_{(m)}$ from $f(\underline{\beta}_{(m)} | \underline{y}, m)$ (optional).
- Propose a new model m' with probability $j(m, m')$.
- Generate $\underline{u}$ from proposal $q(\underline{u} | \underline{\beta}_{(m)}, m, m')$.
- Set $(\underline{\beta}'_{(m')}, \underline{u}') = h_{m,m'}(\underline{\beta}_{(m)}, \underline{u})$.
  - $d(m) + d(\underline{u}) = d(m') + d(\underline{u}')$ and
  - $h_{m',m} = h^{-1}_{m,m'}$.

- Accept the proposed move to model m' with probability $\alpha = \min(1, A)$

$$A = \frac{f(\underline{y}|\underline{\beta}'_{(m')}, m') f(\underline{\beta}'_{(m')}|m') f(m') j(m', m) q(\underline{u}'|\underline{\beta}'_{(m')}, m', m)}{f(\underline{y}|\underline{\beta}_{(m)}, m) f(\underline{\beta}_{(m)}|m) f(m) j(m, m') q(\underline{u}|\underline{\beta}_{(m)}, m, m')} \left| \frac{\partial h(\underline{\beta}_{(m)}, \underline{u})}{\partial(\underline{\beta}_{(m)}, \underline{u})} \right|.$$

2.1.2 Carlin and Chib Gibbs Sampler

Characteristic:
Requires realisations of $\{\underline{\beta}_{(m_k)} : m_k \in \mathcal{M}, m\}$.

The Procedure
Suppose that the current state is $\left(\{\underline{\beta}_{(m_k)} : m_k \in \mathcal{M}, m\}\right)$, then

- Generate $\underline{\beta}_{(m)}$ from $f(\underline{\beta}_{(m)}|\underline{y}, m)$.
- Generate $\underline{\beta}_{(m_l)}$ from $f(\underline{\beta}_{(m_l)}|m_l \neq m)$.
  - Pseudo-parameters: $\underline{\beta}_{(m_l)}$ are called,
  - Pseudopriors or linking densities: $f(\underline{\beta}_{(m_l)}|m_l \neq m)$.
  - No need to specify different $f(\underline{\beta}_{(m_l)}|m_l \neq m)$ for different m.

- The model indicator m is generated by

$$f(m|\{\underline{\beta}_{(m_k)} : m_k \in \mathcal{M}\}, \underline{y}) = \frac{A_m}{\sum_{m_k \in \mathcal{M}} A_{m_k}}$$

$$A_m = f(\underline{y}|\underline{\beta}_{(m)}, m) \prod_{m_l \in \mathcal{M}} \left\{ f(\underline{\beta}_{(m_l)}|m) \right\} f(m).$$

Drawback:
Specification and generation from many pseudopriors (at least $|\mathcal{M}| - 1$)

- computationally demanding (time, memory and hard disk limitations)
- procedure is impracticable for large problems.

2.1.3 Metropolised Carlin and Chib Algorithm

The Procedure
Suppose that the current state is $\left(\underline{\beta}_{(m)}, m\right)$, then

- Generate $\underline{\beta}_{(m)}$ from $f(\underline{\beta}_{(m)}|\underline{y}, m)$.
- Propose a new model m' with probability $j(m, m')$.
- Generate $\underline{\beta}'_{(m')}$ from the proposal $f(\underline{\beta}'_{(m')}|m' \neq m)$.
- Accept the proposed move with probability $\alpha = min(1, A)$

$$A = \frac{f(\underline{y}|\underline{\beta}'_{(m')}, m') f(\underline{\beta}'_{(m')}|m') f(\underline{\beta}_{(m)}|m') f(m') j(m', m)}{f(\underline{y}|\underline{\beta}_{(m)}, m) f(\underline{\beta}_{(m)}|m) f(\underline{\beta}'_{(m')}|m) f(m) j(m, m')}.$$

Important Features

- Requires (only) $\underline{\beta}_{(m)}$ and $\underline{\beta}'_{(m')}$ to calculate α.
- Model m' is proposed with probability $j(m, m')$, independently of the values of any model parameters.

Only need to sample from pseudoprior $f(\underline{\beta}'_{(m')}|m' \neq m)$!!!

RJ and MCC

- MCC is a reversible jump with ...
  - $\left(\underline{\beta}'_{(m')}, \underline{u}'\right) = \left(\underline{u}, \underline{\beta}_{(m)}\right)$,
    $\underline{u}' = \{\underline{\beta}_{(m_l)} : m_l \neq m'\}$, $\underline{u} = \{\underline{\beta}_{(m_l)} : m_l \neq m\}$,
  - proposal densities are replaced by
    $q(\underline{u}|\underline{\beta}_{(m)}, m, m') = \prod_{m_l \in \mathcal{M} \setminus \{m'\}} \left\{ f(\underline{\beta}_{(m_l)}|m') \right\}$ and
    $q(\underline{u}'|\underline{\beta}'_{(m')}, m', m) = \prod_{m_l \in \mathcal{M} \setminus \{m\}} \left\{ f(\underline{\beta}_{(m_l)}|m) \right\}$.
- MCC also coincides to the simpler RJ with:
  - $\left(\underline{\beta}'_{(m')}, \underline{u}'\right) = \left(\underline{u}, \underline{\beta}_{(m)}\right)$, $\underline{u}' = \underline{\beta}_{(m)}$ and $\underline{u} = \underline{\beta}'_{(m')}$,
  - proposal densities are replaced by
    $q(\underline{u}|\underline{\beta}_{(m)}, m, m') = f(\underline{\beta}'_{(m')}|m' \neq m)$ and
    $q(\underline{u}'|\underline{\beta}'_{(m')}, m, m') = f(\underline{\beta}_{(m)}|m \neq m')$

2.1.4 Markov chain Monte Carlo model composition (MC^3)

The Procedure

- Suppose that ...
  - $f(\underline{\beta}_{(m)}|m, \underline{y})$ is available for all models $m \in \mathcal{M}$,
  - $f(\underline{y}|m)$ is also known.
- Consider MCC (or RJ) with
$$\boxed{q(\underline{\beta}_{(m)}|\underline{\beta}'_{(m')}, m, m') = f(\underline{\beta}_{(m)}|m, \underline{y}),}$$
- If $\left(\underline{\beta}_{(m)}, m\right)$ is the current state, then
  - Generate $\underline{\beta}_{(m)}$ from $f(\underline{\beta}_{(m)}|\underline{y}, m)$ (optional).
  - Propose a new model m' with probability $j(m, m')$.

- Generate $\underline{\beta}'_{(m')}$ from the posterior $f(\underline{\beta}'_{(m')}|m, \underline{y})$.
- Accept the proposed model m' with probability

$$\begin{aligned}\alpha &= min\left(1, \frac{f(\underline{y}|m')f(m')j(m', m)}{f(\underline{y}|m)f(m)j(m, m')}\right)\\ &= min\left(1, B_{m'm}\frac{f(m')j(m', m)}{f(m)j(m, m')}\right).\end{aligned}$$

2.2 Variable Selection Algorithms

2.2.1 Stochastic Search Variable Selection

- Originally for Normal models (1993) and then applied in other GLM type models.
- The dimension of the model is constant.
- The model likelihood is given by $f(\underline{y}|\underline{\beta})$ for all models.
- The model indicator m is substituted by $\underline{\gamma}^T = (\gamma_1, \ldots, \gamma_p)$.
- For specified k_j and $\underline{\Sigma}_j$, the indicator variables γ_j are involved in the model through the prior
$$\underline{\beta}_j|\gamma_j \sim \gamma_j N(0, \underline{\Sigma}_j) + (1-\gamma_j)N(0, k_j^{-2}\underline{\Sigma}_j).$$
- Generally SSVS results differ from usual model selection (tend to be close for large k_j).

The Procedure

Suppose that the current state is $(\underline{\beta}, \underline{\gamma})$, then, for $j = 1, \ldots, p$,

- Generate $\underline{\beta}_j$ from
$$f(\underline{\beta}_j|\underline{\beta}_{\setminus j}, \underline{\gamma}, \underline{y}) \propto f(\underline{y}|\underline{\beta}, \underline{\gamma})f(\underline{\beta}_j|\gamma_j)$$
$\underline{\beta}_j$: vector of parameters of j term.
- Generate $\gamma_j \sim Bernoulli\left(\frac{O_j}{1+O_j}\right)$ with
$$O_j = \frac{f(\gamma_j = 1|\underline{\beta}, \underline{\gamma}_{\setminus j}, \underline{y})}{f(\gamma_j = 0|\underline{\beta}, \underline{\gamma}_{\setminus j}, \underline{y})} = \underbrace{\frac{f(\underline{\beta}|\gamma_j = 1, \underline{\gamma}_{\setminus j})}{f(\underline{\beta}|\gamma_j = 0, \underline{\gamma}_{\setminus j})}}_{Prior\ Ratio}\underbrace{\frac{f(\gamma_j = 1, \underline{\gamma}_{\setminus j})}{f(\gamma_j = 0, \underline{\gamma}_{\setminus j})}}_{Prior\ Odds}$$
$\underline{\gamma}_{\setminus j}$: all components of $\underline{\gamma}$ except γ_j.

Variable selection step does not (directly) depend on the model likelihood!

2.2.2 Kuo and Mallick Sampler

Characteristics:

- Originally for Normal models (1993) but can be applied in other GLM type models.
- Likelihood is given by $f(\underline{y}|\underline{\beta}, \underline{\gamma})$.
- Model indicator m is substituted by $\underline{\gamma}$.
- Indicator variables γ_j are involved in the model by substituting $\underline{\beta}_j$ by $\gamma_j\underline{\beta}_j$ in the linear predictor.
- Prior is given by $f(\underline{\beta})$ for all models.
- Generally KM results differ than other model selection methods due to the fact that the underlying priors are automatically defined by $f(\underline{\beta})$.

The Procedure

If the current state is $(\underline{\beta}, \underline{\gamma})$, then, for $j = 1, \ldots, p$,

- Generate $\underline{\beta}_j$ from
  - $f(\underline{y}|\underline{\beta}, \underline{\gamma})f(\underline{\beta}_j|\underline{\beta}_{\setminus j})$ if $\gamma_j = 1$
  - $f(\underline{\beta}_j|\underline{\beta}_{\setminus j})$ if $\gamma_j = 0$
- Generate $\gamma_j \sim Bernoulli\left(\frac{O_j}{1+O_j}\right)$ with

$$O_j = \underbrace{\frac{f(\underline{y}|\underline{\beta}, \gamma_j = 1, \underline{\gamma}_{\setminus j})}{f(\underline{y}|\underline{\beta}, \gamma_j = 0, \underline{\gamma}_{\setminus j})}}_{Likelihood\ Ratio}\underbrace{\frac{f(\gamma_j = 1, \underline{\gamma}_{\setminus j})}{f(\gamma_j = 0, \underline{\gamma}_{\setminus j})}}_{Prior\ Odds}.$$

Advantage: extremely straightforward.
Disadvantage: There is no flexibility to improve efficiency.

2.2.3 Gibbs Variable Selection

Characteristics

- Natural hybrid of SSVS and the Kuo and Mallick (1998) sampler.
- Same likelihood as in Kuo and Mallick Sampler.
- Specify prior as $f(\underline{\beta}|\underline{\gamma})f(\underline{\gamma})$.
  Consider the partition of $\underline{\beta} = \left\{\underline{\beta}_{(\gamma)}, \underline{\beta}_{(\backslash\gamma)}\right\}$ into
  - $\underline{\beta}_{(\gamma)}$: parameters in model ($\gamma_j = 1$)
  - $\underline{\beta}_{(\backslash\gamma)}$: parameters not in model ($\gamma_j = 0$)

  then $f(\underline{\beta}|\underline{\gamma})$ may be partitioned into
  - **Prior**: $f(\underline{\beta}_{(\gamma)}|\underline{\gamma})$ and **Pseudoprior**: $f(\underline{\beta}_{(\backslash\gamma)}|\underline{\beta}_{(\gamma)}, \underline{\gamma})$.

The Procedure
If the current state is $(\underline{\beta}, \underline{\gamma})$, then

- Generate parameters $\underline{\beta}_{(\gamma)}$ from

$$f(\underline{\beta}_{(\gamma)}|\underline{\beta}_{(\backslash\gamma)}, \underline{\gamma}, \underline{y}) \propto f(\underline{y}|\underline{\beta}, \underline{\gamma}) f(\underline{\beta}_{(\gamma)}|\underline{\gamma}) f(\underline{\beta}_{(\backslash\gamma)}|\underline{\beta}_{(\gamma)}, \underline{\gamma})$$

- Generate pseudo-parameters $\underline{\beta}_{(\backslash\gamma)}$ from $f(\underline{\beta}_{(\backslash\gamma)}|\underline{\beta}_{(\gamma)}, \underline{\gamma})$
- Generate $\gamma_j \sim Bernoulli\left(\frac{O_j}{1+O_j}\right)$ with

$$O_j = \underbrace{\frac{f(\underline{y}|\underline{\beta}, \gamma_j = 1, \underline{\gamma}_{\backslash j})}{f(\underline{y}|\underline{\beta}, \gamma_j = 0, \underline{\gamma}_{\backslash j})}}_{Likelihood\ Ratio} \underbrace{\frac{f(\underline{\beta}|\gamma_j = 1, \underline{\gamma}_{\backslash j})}{f(\underline{\beta}|\gamma_j = 0, \underline{\gamma}_{\backslash j})}}_{Prior/Pseudoprior\ Ratio} \underbrace{\frac{f(\gamma_j = 1, \underline{\gamma}_{\backslash j})}{f(\gamma_j = 0, \underline{\gamma}_{\backslash j})}}_{Prior\ Odds}$$

Simpler Approach

- Assume prior: $f(\underline{\beta}_j|\underline{\gamma}_j) = \gamma_j N(\mathbf{0}, \underline{\Sigma}_j) + (1-\gamma_j)N(\underline{\bar{\mu}}_j, \underline{S}_j)$,
  $\underline{\bar{\mu}}_j$ and $\underline{S}_j$: are pseudoprior parameters (tuned to achieve optimal convergence).
- The full conditional posterior distribution is now given by

$$f(\underline{\beta}_j|\underline{\beta}_{\backslash j}, \underline{\gamma}, \underline{y}) \propto \begin{cases} f(\underline{y}|\underline{\beta}, \underline{\gamma})N(0, \underline{\Sigma}_j) & \gamma_j = 1 \\ N(\underline{\bar{\mu}}_j, \underline{S}_j) & \gamma_j = 0 \end{cases}$$

  This approach is ...
  - Simple to apply
  - Efficient when covariates are not highly correlated.
  - Easy to specify pseudopriors
    Get $\underline{\bar{\mu}}_j$ and $\underline{S}_j$: from a pilot run of the full model; see Dellaportas and Forster (1999).

2.3 Proposal Distributions

- Proposal Distributions for Model Parameters
  - Independent distributions for each term j: $N(\underline{\bar{\mu}}_j, \underline{S}_j)$.
  - SSVS type proposal: $N(\mathbf{0}_{d_j}, \underline{\Sigma}_j/k_j^2)$.
  - Maximum likelihood based: $N\left(\underline{\hat{\beta}}_{(m)}, \underline{\hat{\Sigma}}_{(m)}\right)$.
  - Alternative easy-to-use choice: $N(\underline{\hat{\beta}}_{(m)}, \underline{\Sigma}_{(m)}/k^2)$.
  - Using conditional maximised likelihood.
  - Giudici and Roberts (1998) automatic choice.
  - Brooks, Giudici and Roberts (2001): Optimal Proposals
  - Green and Mira (2001): Delayed rejection algorithm.

- Proposal Distributions on Model Space
  - Common proposal: Uniform distribution.
  - 'Local' and 'Global' proposals.
  - $j(m, m) = 0$ better than $j(m, m) > 0$ (Liu 1996a,b).
  - Set $j(m, m')$ using Laplace or BIC approximations.
  - Use an MC^3 when size of $\mathcal{M}$ is large.

2.3.1 Proposal Distributions for Model Parameters

- Independent distributions for each term j: $N(\underline{\bar{\mu}}_j, \underline{S}_j)$.
  Get pseudoparameters from pilot run of the full model.
- SSVS type proposal: $N(\mathbf{0}_{d_j}, \underline{\Sigma}_j/k_j^2)$,
  with $\underline{\Sigma}_j$ the prior covariance matrix.
- Maximum likelihood based: $N\left(\underline{\hat{\beta}}_{(m)}, \underline{\hat{\Sigma}}_{(m)}\right)$;
  where $\underline{\hat{\beta}}_{(m)}$ and $\underline{\hat{\Sigma}}_{(m)}$ are the MLE of model m.
- Alternative easy-to-use choice: $N(\underline{\hat{\beta}}_{(m)}, \underline{\Sigma}_{(m)}/k^2)$.

- Using conditional maximised likelihood:

$$q(\underline{\beta}_j | \underline{\beta}_{(\gamma_j=0,\underline{\gamma}_{\setminus j})}, \gamma_j = 1, \gamma_j = 0, \underline{\gamma}_{\setminus j}) =$$

$$N\left(\left(\underline{X}_j^T \underline{\hat{H}} \underline{X}_j\right)^{-1} \underline{X}_j^T \underline{\hat{H}} \underline{\eta}_j^*, \left(\underline{X}_j^T \underline{\hat{H}} \underline{X}_j\right)^{-1} \right),$$

where

  - $\underline{\hat{H}}$ is the weight matrix used in observed information matrix of the 'saturated' model and
  - $\underline{\eta}_j^*$ is a vector with elements given by

$$\{\underline{\eta}_j^*\}_i = g(y_i) - \sum_{l \in \mathcal{V} \setminus \{j\}} \gamma_l \underline{x}_{il} \underline{\beta}_l.$$

Alternatively, for simplicity, we may substitute the covariance matrix by $\underline{\Sigma}_j / k^2$.

- Giudici and Roberts (1998) automatic choice.
  Scale parameter varies according to proposed values maximizing the acceptance probability when proposed parameters are zero.
- Brooks, Giudici and Roberts (2001) proposals by maximising acceptance ratio.

2.3.2 Proposal Distributions on Model Space

- Common proposal: Uniform distribution.
- 'Local' and 'Global' proposals.
  - Global proposals result in low acceptance rates
  - Local proposals are preferred (in structured $\mathcal{M}$).
  - Generally, RJ with local proposals perform well.
    May exhibit difficulties in some ill-posed problems.
    In such cases combination may be optimal.
- $j(m,m) = 0$ is more efficient than $j(m,m) > 0$ (Liu 1996a,b).
- Set $j(m,m')$ using Laplace or BIC approximations.
- When size of $\mathcal{M}$ is large:
  Use an MC^3 based on approximations to get rough estimates of posterior weights.

3 Prior Specification

3.1 Jeffreys-Lindley Paradox

Consider two models m_0 and m_1;

- $d(m)$ dimension of model m,
- $d(m_0) < d(m_1)$; model m_0 is simpler.

1. If sample size $n \to \infty$: $B_{10} \to 0$
   Bayes factor supports simpler models in contradiction to significance tests (Lindley, 1957, Bk).
2. If prior variance of additional parameters $\to \infty$: $B_{10} \to 0$ (Bartlett, 1957, Bk).

(1) and/or (2) are referred in literature as

- 'Lindley's paradox' → for any case where Bayesian and significance tests result in contradictive evidence (Shafer, 1982, JASA).
- 'Bartlett' paradox → Kass and Raftery (1995, JASA)
- 'Jeffreys' paradox → Lindley (1980, *An.Stat.*), Berger and Delampady (1987, *St.Science*)
- 'Jeffreys-Lindley's paradox' → Robert (1993, *St.Sinica*).
- 'Bartlett - Lindley' paradox → Chipman *et al.* (2000, Tec.Rep.).
- For detailed discussion → Shafer (1982, JASA).

We focus on Variable Selection Problems for GLM.

Let us consider a GLM with $n \times 1$ vector of linear predictors given by

$$\underline{\eta} = \underline{X}_{(m)} \underline{\beta}_{(m)}$$

- $\underline{X}_{(m)}$ = design matrix of model m
- $\underline{\beta}_{(m)}$ = vector of parameters involved in the linear predictors.

3.2 Prior Distributions for the parameters of the linear predictor

$$f\left(\underline{\beta}_{(m)}|m\right) \sim N\left(\underline{\mu}_{\beta_{(m)}}, \underline{\Sigma}_{(m)}\right)$$

Low Information Prior Distributions proposed in literature:

- $\underline{\mu}_{\beta_m} = \mathbf{0}$: prior centered against alternative hypothesis.
- $\underline{\Sigma}_{(m)} = c^2 \underline{V}_{(m)}$ or $\underline{\Sigma}_{(m)} = c^2 \underline{V}_{(m)} \sigma^2$ in regression.

The choice of $\underline{\Sigma}_{(m)}$ remains difficult. Two types of prior distributions

- Block Diagonal Covariance Matrix (independent priors)
- Non-diagonal Covariance Matrix

Normal Independent priors, $\underline{V}_{(m)} = Diagonal(v_i^2)$:

- George and McCullogh (1993, JASA) in SSVS
- Geweke (1996, *B.Stat.*): Independent truncated normal distributions in regression.

Non-diagonal Covariance Matrix

- REGRESSION: $\underline{\Sigma}_{(m)} = c^2 \underline{V}_{(m)} \sigma^2$
  * $\underline{V}_{(m)}^{-1} = \underline{X}_{(m)}^T \underline{X}_{(m)} \rightarrow$ Zellner's g-priors (Zellner, 1980).
  * $c^2 \in [10, 100]$ proposed by Smith and Kohn (1996, *J.Econ.*).
  * $c^2 = n \rightarrow$ Unit Information priors (Kass and Wasserman, 1995, JASA).
  * Fernandez *et al.* (2001, *J.Econ.*) used various values for c^2; proposed $c^2 = \max\{d(m)^2, n\}$.

- Contingency tables: $\underline{\Sigma}_{(m)} = c^2 \underline{V}_{(m)}$
  * Albert (1996, *Can.J.St.*): based on prior beliefs on odds ratios.
  * Dellaportas and Forster (1999,Bk) based on Knuinman and Speed (1988, *Bc*); $\underline{V}_{(m)}^{-1} = \underline{X}_{(m)}^T \underline{X}_{(m)}$, $c^2 = 2 \times \#cells$.
  * Ntzoufras *et al*. (2000, JSCS): combination of the above for SSVS.
- GLM $\rightarrow$ Raftery (1996, Bk):
  * diagonal covariance matrix and mean zero for covariates based on sample variances.
  * Nonzero mean and correlation of intercept with the rest of parameters.
  * $c^2 = 2.85^2$ based on mathematical arguments.
- Ntzoufras *et al*. (2001): Constructed 'equivalent' priors across GLM with different link function based on Taylor expansion.

- Unit Information Prior $\underline{\Sigma}_{(m)} = n(-\underline{H}_{(m)})^{-1}$ (Kass and Wasserman, 1995, JASA); $\underline{H}_{(m)}$ is the Hessian matrix.
- Kuo and Mallick (1998, Sankya): Define prior only on full model.
- Using Imaginary data to construct an informative prior: Chen *et al*. (1999, JRSSB).
- George and Foster (2000, Bk): Empirical Bayes Approach.
- Expected Posterior Prior Distributions (Perez and Berger, 2000)

3.3 Prior Distributions on Model Space

- Usual naive prior: Uniform prior on model space $\mathcal{M}$ $p(m) = 1/|\mathcal{M}|$. Informative in terms of dimension (Chipman *et al.* , 2000, Tec.Rep.).
- Alternative: Use prior on dimension (Chipman *et al.* , 2000, Tec.Rep.).
- Use Beta prior on common inclusion probability (George and McCullogh, 1997, *St.Sin.*, Kohn *et al.* , 2001 *St.Comp.*).
- Elicit imaginary data: Chen *et al*. (1999, JRSSB)
- Use Empirical Bayes Approach (George and Foster, 2000, Bk).
- Prior distribution based on Dilution of models (George, 1999, *B.Stat*).

3.4 What Prior in BUGS

- Standardize variables or use STZ constraints
- Use unit information priors (may incorporate data)
- Empirical approach: Estimate posterior variance and set

  $$prior\ variance = c^2 \times posterior\ variance.$$

  For $c^2 = 1$ (approx) posterior Bayes factor.
  For $c^2 = n$ (approx) unit information prior (BIC)

- Use $\underline{\Sigma}_m = (\underline{X}_m^T \underline{X}_m)^{-2} \sigma^2$ for Normal models
- For logistic regression models and/or poisson log-linear models may use priors of Dellaportas *et al.* (2000,2002).
- Generally use a range of prior distribution base inference.

4 Bayesian Model and Variable Selection Using Bugs

- Carlin and Chib Method
- Variable Selection Methods (SSVS, KM, GVS)

4.1 Carlin and Chib Method Using BUGS

- BUGS Examples vol.2, page 47, example 13: Pines dataset.
- Data originally used by Williams (1959, Regression Analysis) and re-analyzed by Carlin and Chib (1995, JRSS,B).
  - 42 speciments of radiata pine.
  - y_i: maximum comprehensive strength.
  - x_i: density.
  - z_i: density adjusted for resin content.

- Two competing models:
  - Model 1: $y_i \sim Normal(\alpha + \beta x_i, \tau_1)$
  - Model 2: $y_i \sim Normal(\gamma + \delta z_i, \tau_2)$
- Data originally used by Williams (1959, Regression Analysis) and re-analyzed by Carlin and Chib (1995, JRSS,B).
  - 42 speciments of radiata pine.
  - y_i: maximum comprehensive strength.
  - x_i: density.
  - z_i: density adjusted for resin content.

Alternative we could have written
$\mu_i = I(m=1)(\alpha + \beta x_i) + [1 - I(m=1)](\gamma + \delta z_i)$

	Model 1	Model 2
Model Structure	$Y_i \sim N(0,\tau_1)$	$Y_i \sim N(0,\tau_2)$
	$\mu_i = \alpha + \beta x_i$	$\mu_i = \gamma + \delta z_i$
Prior	$f(\alpha,\beta,\tau_1 \mid m=1)$	$f(\gamma,\delta,\tau_2 \mid m=2)$
Pseudoprior	$f(\alpha,\beta,\tau_1 \mid m=2)$	$f(\gamma,\delta,\tau_2 \mid m=1)$

- MODEL 1
  - $f(\alpha|m) = N(\mu_\alpha[m], \tau_\alpha[m])$
  - $f(\beta|m) = N(\mu_\beta[m], \tau_\beta[m])$
  - $f(\tau_1|m) = \Gamma(r1[m], l1[m])$
  - for $m=1$: Prior
  - for $m=2$: Pseudo-Prior
- MODEL 2
  - $f(\gamma|m) = N(\mu_\gamma[m], \tau_\gamma[m])$
  - $f(\delta|m) = N(\mu_\delta[m], \tau_\delta[m])$
  - $f(\tau_2|m) = \Gamma(r2[m], l2[m])$
  - for $m=2$: Prior
  - for $m=1$: Pseudo-Prior

Procedure:

1. Pilot Run 1: Run MCMC for Model 1
2. Estimate parameters of Model 1
3. Pilot Run 2: Run MCMC for Model 2
4. Estimate parameters of Model 2
5. Run CC algorithm with pseudoparameters specified by 2 and 4

Comment 1: The Effect of Lindley's Paradox is not direct since the two models have the same dimension.

Comment 2: We may change prior model probabilities to achieve mobility across models and estimate posterior or Bayes factors more accurately.

BUGS CODE

```
model
{
#       standardise values
        for (i in 1:n){
                xs[i]<-(x[i]-mean(x[]))/sd(x[]);
                ys[i]<-(y[i]-mean(y[]))/sd(y[]);
                zs[i]<-(z[i]-mean(z[]))/sd(z[]);
        }
#       model likelihoods
        for (i in 1:n){
                ys[i]~dnorm( mu[i,m], tau[m]  );
                mu[i,1]<-alpha +beta *xs[i];
                mu[i,2]<-gamma +delta*zs[i];
        }
```



```
#       priors
        m~dcat(p[]);  # categorical 1/2
        p[1]<-0.5;    # "non-informative"
        p[2]<-1-p[1];
        mdl<-m-1
#
#        alternative prior (bernoulli)
#        m<-mdl+1
#        mdl~dbern(p)
#        p<-0.5
#
```



```
#       priors for model parameters
        alpha~dnorm(mu.alpha[m], tau.alpha[m]);
        beta ~dnorm(mu.beta[m] , tau.beta[m] );
        gamma~dnorm(mu.gamma[m], tau.gamma[m]);
        delta~dnorm(mu.delta[m], tau.delta[m]);
        tau[1]~dgamma( r1[m], l1[m] );
        tau[2]~dgamma( r2[m], l2[m] );
#
```



```
#       prior parameters
        mu.alpha[1]<-0.0;
        mu.beta[1] <-0.0;
        mu.gamma[2]<-0.0;
        mu.delta[2]<-0.0;
        tau.alpha[1]<-1.0E-06
        tau.beta[1] <-1.0E-04;
        tau.gamma[2]<-1.0E-06;
        tau.delta[2]<-1.0E-04;
        r1[1]<-0.0001;
        l1[1]<-0.0001;
        r2[2]<-0.0001;
        l2[2]<-0.0001;
```



```
#       pseudoparameters
        mu.alpha[2]<-???;
        mu.beta[2] <-???;
        mu.gamma[1]<-???;
        mu.delta[1]<-???;
        tau.alpha[2]<-???;
        tau.beta[2] <-???;
        tau.gamma[1]<-???;
        tau.delta[1]<-???;
        r1[2]<-???;
        l1[2]<-???;
        r2[1]<-???;
        l2[1]<-???;
}
```

Use as initial pseudopriors $N(0,1)$ and $\Gamma(1,1)$

PILOT RUN RESULTS

Model 1: `p[1]<-1.0`

```
node    mean     sd         MC error  2.5%    median   97.5%  start sample
mdl    0.0      0.0        3.162E-12 0.0     0.0      0.0    1001 1000
alpha  7.834E-4 0.06047    0.00168  -0.1231 -9.51E-4  0.1172 1001 1000
beta   0.9275   0.05988    0.002036  0.8124  0.9279   1.047  1001 1000
tau[1] 6.92     1.551      0.05284   4.267   6.823   10.3    1001 1000
----------------------------------------------------------------------
gamma -0.005343 1.001      0.03553  -1.999  -0.01715  1.999  1001 1000
delta  0.01607  0.9835     0.03815  -1.885   0.03026  1.876  1001 1000
tau[2] 1.013    1.041      0.03457   0.02281 0.6726   3.838  1001 1000
```

PILOT RUN RESULTS

Model 2: `p[1]<-0.0`

```
node    mean     sd  MC error   2.5%     median   97.5% start sample
mdl     1.0      0.0     3.162E-12 1.0     1.0       1.0   1001 1000
alpha   0.01448  1.005   0.03181  -1.988   0.02122  2.062 1001 1000
beta    0.0384   0.989   0.02649  -1.842   0.0685   2.036 1001 1000
tau[1]  1.04     1.03    0.02945   0.02231 0.696    4.056 1001 1000
-----------------------------------------------------------------------
gamma   7.021E-4 0.04908 0.001408 -0.09903 0.001332 0.098 1001 1000
delta   0.9522   0.0498  0.001652  0.8541  0.9513   1.05  1001 1000
tau[2] 10.41     2.369   0.07938   6.239   10.18   15.39  1001 1000
```


Estimate Pseudopriors:

1. Estimate $N(\mu,\tau)$ pseudopriors by
   (a) μ = posterior mean from pilot run
   (b) τ = (posterior s.d. from pilot run)$^{-2}$
2. Estimate $\Gamma(a,b)$ pseudopriors by
   (a) $E(X)=a/b$, $V(X)=a/b^2 \Rightarrow b=E(X)/V(X)$ and $a=[E(X)]^2/V(X)$
   (b) a = (posterior mean)^2/(posterior s.d.)^2
   (c) b = (posterior mean) /(posterior s.d.)^2

	$m=1$	m=2	
Model 1	(prior)	(pseudoprior)	(Pilot Run)
$\mu_\alpha[m]$	0.0	0.0	0.0008
$\tau_\alpha[m]$	10^{-6}	256	$(0.06047)^{-2}=273.5$
$\mu_\beta[m]$	0.0	1.0	0.9275
$\tau_\beta[m]$	10^{-4}	256	$(0.05988)^{-2}$= 278.9
$r1[m]$	10^{-3}	30	$6.92^2/(1.55)^2=19.9$
$l1[m]$	10^{-3}	4.5	$6.92/(1.55)^2=2.88$
Model 2	(pseudoprior)	(prior)	(Pilot Run)
$\mu_\gamma[m]$	0.0	0.0	0.0007
$\tau_\gamma[m]$	400	10^{-6}	$0.04908^{-2}=415.13$
$\mu_\delta[m]$	0.0	0.0	0.9522
$\tau_\delta[m]$	400	10^{-4}	$0.0498^{-2}=403.22$
$r2[m]$	46	10^{-3}	$(10.41/2.369)^2=19.93$
$l2[m]$	4.5	10^{-3}	$10.41/(2.369)^2=1.85$

	Model 1		Model 2	
Model	$Y_i \sim N(0,\tau_1)$		$Y_i \sim N(0,\tau_2)$	
Structure	$\mu_i=\alpha+\beta x_i$		$\mu_i=\gamma+\delta z_i$	
Prior	$f(\alpha\|m=1)$	$f(\beta\|m=1)$	$f(\gamma\|m=2)$	$f(\delta\|m=2)$
	$N(0,10^{-6})$	$N(0,10^{-4})$	$N(0,10^{-6})$	$N(0,10^{-4})$
	$f(\tau_1\|m=1)=\Gamma(10^{-3},10^{-3})$		$f(\tau_2\|m=2)=\Gamma(10^{-3},10^{-3})$	
Pseudoprior	$f(\alpha\|m=2)$	$f(\beta\|m=2)$	$f(\gamma\|m=1)$	$f(\delta\|m=1)$
	$N(0,256)$	$N(1,256)$	$N(0,400)$	$N(1,400)$
	$f(\tau_1\|m=2)=\Gamma(30,4.5)$		$f(\tau_2\|m=1)=\Gamma(46,4.5)$	


```
#       pseudoparameters
        mu.alpha[2]<-0.0;
        mu.beta[2] <-1.0;
        mu.gamma[1]<-0.0;
        mu.delta[1]<-1.0;
        tau.alpha[2]<-256;
        tau.beta[2] <-256;
        tau.gamma[1]<-400;
        tau.delta[1]<-400;
        r1[2]<-30;
        l1[2]<-4.5;
        r2[1]<-46;
        l2[1]<-4.5;
}
```


ESTIMATING BAYES FACTOR

	Pseudopriors	$P(m=1)$	$P(m=1\|y)$	PO	Bayes Factor
1	BUGS	0.5	0.9992	1249	1249
2	PILOT	0.5	0.9996	2499	2499
3	BUGS	0.9995	0.6140	1.591	3180
4	PILOT	0.9995	0.6175	1.614	3227
5	Manual	0.9995	0.6290	1.695	3389
6	CC	0.9995	0.6890	2.215	4420

4.2 Bayesian Variable Selection in BUGS

4.2.1 Illustrative Example: $2 \times 2 \times 2$ Contingency Table

- Data taken from Healy (1988).
- 3-way table
- Factor A =condition of the patient (more or less severe),
- Factor B =if the patient was accepting antitoxin medication
- Factor C (response) = whether the patient survived or not.
- Use a Logistic Regression Model

		Survival(C)	
Condition (A)	Antitoxin (B)	No	Yes
More Severe	Yes	15	6
	No	22	4
Less Severe	Yes	5	15
	No	7	5

Table 3: Example Dataset.

- RESPONSE: Factor C (response) = Survival
- EXPLANATORY TERMS:
  - Factor A =Condition
  - Factor B =Antitoxin
  - Interaction AB = Condition*Antitoxin
- MODELS
  - MODEL 1: AB = A+B+AB
  - MODEL 2: A+B = A+B
  - MODEL 3: A = A
  - MODEL 4: B = B
  - MODEL 5: null= constant

PRIOR DISTRIBUTIONS

- prior variance = 4×2
- prior probability of each model 1/5:
  - $\gamma_{AB} \sim Bernoulli(1/5)$
  - $\gamma_i | \gamma_{AB} \sim Bernoulli(\pi)$, with $\pi = 0.5(1-\gamma_{AB}) + \gamma_{AB}$ for $i \in \{A, B\}$.

DATA IN BUGS

```
r[] n[] x[,1] x[,2] x[,3] x[,4]
 5 12 1 -1   -1  1
 4 26 1  1   -1 -1
15 20 1 -1    1 -1
 6 21 1  1    1  1
```


BUGS CODE FOR SSVS

The Model

```
for (i in 1:N) {
    r[i]~dbin(p[i],n[i]);
    logit(p[i])<-b[1]+ x[,2]* b[2]+ x[,3]*b[3]+  x[,4]*b[4];
                }
```

The Prior on Model Parameters

```
       for (i in 2:N) {
                c[i]<-1000.0
                tau[i]<-pow(c[i],2-2*g[i])/8;
                bpriorm[i]<-0.0;
                b[i]~dnorm(bpriorm[i],tau[i]);
                b[i]~dnorm(bpriorm[i],tau[i]);
        }
```

The Prior on constant Model Parameters (we may use ”non-informative”)

```
tau[1]<-0.1;
bpriorm[1]<-0.0;
b[1]~dnorm(bpriorm[1],tau[1]);
```

The Model Prior (common for all approaches)

```
g[4]~dbern(0.2);
include<-(1-g[4])*0.5+g[4]*1.0
g[2]~dbern(include);
g[3]~dbern(include);
g[1]~dbern(1.0);
```


BUGS CODE FOR KM

The Model

```
for (i in 1:N) {
    r[i]~dbin(p[i],n[i]);
    logit(p[i])<-b[1] +  x[,2]* g[2]* b[2]
                      +  x[,3]* g[3]* b[3]
                      +  x[,4]* g[4]* b[4];
                }
```

The Prior on Model Parameters

```
for (i in 2:N) {
         tau[i]<-1/8;
         bpriorm[i]<-0.0;
         b[i]~dnorm(bpriorm[i],tau[i]); }
```


BUGS CODE FOR GVS

The Model

```
for (i in 1:N) {
    r[i]~dbin(p[i],n[i]);
    logit(p[i])<-b[1] +  x[,2]* g[2]* b[2]
                      +  x[,3]* g[3]* b[3]
                      +  x[,4]* g[4]* b[4];  }
```

The Prior on Model Parameters

```
       for (i in 2:N) {
#                tau[i]<-pow(100,1-g[i])/8;
#                bpriorm[i]<-0.0;
                tau[i]<-g[i]/8+(1-g[i])/(se[i]*se[i]);
                bpriorm[i]<-mean[i]*(1-g[i]);
                b[i]~dnorm(bpriorm[i],tau[i]); }
```


In model specification we may use

```
for (i in 1:N) {for (j in 1:N) {
            z[i,j]<-x[i,j]*b[j]*g[j]
    }}
for (i in 1:N) {
    r[i]~dbin(p[i],n[i]);
    logit(p[i])<-sum(z[i,]);
                }
```


ESTIMATING POSTERIOR PROBABILITIES IN BUGS

```
#       defining model code
#       0 for  constant, 1 for [A], 2 for [B], 3 for [A][B],
#       6 for [AB]
#
        mdl<-g[2]+2*g[3]+3*g[4];
        pmdl[1]<-equals(mdl,0)
        pmdl[2]<-equals(mdl,1)
        pmdl[3]<-equals(mdl,2)
        pmdl[4]<-equals(mdl,3)
        pmdl[5]<-equals(mdl,6)
```



```
burn-in period:  10,000 iterations.
SSVS -> 500,000 iterations
Kuo and Mallick's method ->500,000 iterations
GVS -> and 100,000 iterations
```

Models	SSVS	KM	GVS
1	0.2	0.5	0.5
A	48.0	49.2	49.3
B	1.0	1.2	1.2
$A+B$	45.3	44.0	43.9
AB	5.5	5.2	5.1

Table 4: Posterior model probabilities (%) for logistic regression. SSVS: Stochastic Search Variable Selection; KM: Kuo and Mallick's Unconditional Priors approach; GVS: Gibbs Variable Selection.

5 Bayesian Model Diagnostics in BUGS

[BUGS manual: page 40]

1. Residuals
2. Model Comparison
3. Goodness of fit

EXAMPLE: line.bug

```
model
    {
        for( i in 1 : N ) {
            Y[i] ~ dnorm(mu[i],tau)
            mu[i] <- alpha + beta * (x[i] - xbar)
        }
        tau ~ dgamma(0.001,0.001) sigma <- 1 / sqrt(tau)
        alpha ~ dnorm(0.0,1.0E-6)
        beta ~ dnorm(0.0,1.0E-6)
    }

Data:  list(x = c(1, 2, 3, 4, 5),
        Y= c(1, 3, 3, 3, 5), xbar = 3, N = 5)

Inits: list(alpha = 0, beta = 0, tau = 1)
```


5.1 Checking Residuals

Assess the predictive distribution of residual functions:

1. Residual: $r_i = y_i - E(y_i)$
   `resid[i]<-y[i]-mu[i];`
2. Standardized Residual: $sr_i = r_i/\sqrt{V(y_i)} = (y_i - E(y_i))/\sqrt{V(y_i)}$
   `sresid[i]<-r[i]*sqrt(tau);`
3. Chance of more extreme observation: $min(P(Y_i < y_i), P(Y_i > y_i))$
   `Y.rep[i]<-dnorm(mu[i],tau);`
   `p.smaller[i]<-step(y[i],Y.rep[i]);`
4. Chance of more surprising observation: $P(Y_i : P(Y_i) < P(y_i))$
5. Predictive ordinate of y_i: $P(y_i)$
   `like[i]<-sqrt(tau/(2*PI))*exp(-0.5*pow(sresid[i],2));`
   `p.inv[i]<-1/like[i];`

5.2 Model Comparison

1. Bayes Factor: we have already done some examples
2. Cross-Validatory Measures:
   - sum of squares of residuals
   - Negative cross-validatory log-likelihood:
     $$NCV = -\sum_{i=1}^{n} \log f(y_i|\underline{y}_{\setminus i}) = -\sum_{i=1}^{n} \log(1/mean(p.inv[i]))$$
     [pseudo-Bayes factor].
3. Deviance: Calculate $D = -2\log f(\underline{y}|\underline{\theta})$ and consider the minimum (non-hierarchical) and/or mean (hierarchical).
   `like[i]<-sqrt(tau/(2*PI))*exp(-0.5*pow(sresid[i],2));`
   `log.like[i]<-log(like[i]);`
   `deviance<- -2*sum(log.like[])`

5.3 Goodness of Fit

BAYESIAN p-values

1. Consider a statistic $d(\underline{y})$
2. Find the predictive distribution of $d(\underline{y})$
3. Estimate the probability $P(d(Y) < d(\underline{y}))$.

IN BUGS

1. Generate predictive sample $\underline{y}^{rep}$
2. Calculate $d^{rep} = d(\underline{y}^{rep})$
3. Set p.d=1 if $d^{rep} > d$
4. Estimate p-value by posterior mean of p.d

5.4 Goodness of Fit

BUGS CODE: P-value for Skewness

```
for (i in 1:N) {
    Y.rep[i]<-dnorm(mu[i],tau);
    m3[i]<-power(sresid[i],3);
    m3.rep[i]<-power( (Y.rep[i]-mu[i])*sqrt(tau),3); }
    skew.obs<-sum(m3[])/N ;
    skew.rep<-sum(m3.rep[])/N ;
    p.skew<-step(skew.rep-skew.obs);
```

BUGS CODE: P-value for Kurtosis

```
for (i in 1:N) {
    Y.rep[i]<-dnorm(mu[i],tau);
    m4[i]<-power(sresid[i],4);
    m4.rep[i]<-power( (Y.rep[i]-mu[i])*sqrt(tau),4); }
    kur.obs<-sum(m4[])/N ;
    ku.rep<-sum(m4.rep[])/N ;
    p.kur<-step(kur.rep-kur.obs);
```


RESULTS (1000 burnin, 10000 iterations)

```
            Original   With Outlier
deviance    12.92         25.34
p.skew       0.498         0.43
p.kur        0.733         0.70
                                   1/mean(p.in[i])
p.inv[1]     5.32         25.83     0.188   0.039
p.inv[2]     6.82        136.20     0.147   0.007
p.inv[3]     2.85         10.35     0.351   0.097
p.inv[4]     6.89         11.85     0.145   0.084
p.inv[5]     5.12         14.44     0.195   0.069
--------------------------------------------------
NCV                                 8.20     15.69
                                  min(p.smaller, 1-p.smaller)
p.smaller[1]  0.356        0.300    0.356        0.300
p.smaller[2]  0.799        0.853    0.201        0.147
p.smaller[3]  0.502        0.400    0.488        0.400
p.smaller[4]  0.202        0.352    0.202        0.352
p.smaller[5]  0.659        0.552    0.341        0.448
```



```
resid[1]     -0.40        -2.00
resid[2]      0.80         3.59
resid[3]     -0.00        -0.81
resid[4]     -0.80        -1.22
resid[5]      0.39         0.37

sresid[1]    -0.51        -0.73
sresid[2]     1.00         1.30
sresid[3]    -0.00        -0.29
sresid[4]    -1.01        -0.44
sresid[5]     0.50         0.14
```


END OF TUTORIAL