Bayesian Model Averaging

Hoff Chapter 9, Hoeting et al 1999, Clyde & George 2004, Liang et al 2008

March 6, 2017

Outline

- Problems with g-priors
- Alternatives: Mixtures of g-priors
- Model Averaging
- Choice of Model

Bayes Factors

- Bayes Factor = ratio of marginal likelihoods

Bayes Factors

- Bayes Factor $=$ ratio of marginal likelihoods
- Posterior odds $=$ Bayes Factor $\times$ Prior odds

Bayes Factors

- Bayes Factor = ratio of marginal likelihoods
- Posterior odds = Bayes Factor × Prior odds
- Posterior Probability

$$P(\mathcal{M}_\gamma \mid \mathbf{Y}) = \frac{BF[\mathcal{M}_\gamma : \mathcal{M}_0] p(\mathcal{M}_\gamma)/p(\mathcal{M}_0)}{\sum_{\mathcal{M}_\gamma \in \Gamma} BF[\mathcal{M}_\gamma : \mathcal{M}_0] p(\mathcal{M}_\gamma)/p(\mathcal{M}_0)}$$

Problem with g-Prior with arbitrary g

The Bayes factor for comparing $\mathcal{M}_\gamma$ to the null model:

Problem with g-Prior with arbitrary g

The Bayes factor for comparing $\mathcal{M}_\gamma$ to the null model:

- Let g be a fixed constant and take n fixed.

Problem with g-Prior with arbitrary g

The Bayes factor for comparing $\mathcal{M}_\gamma$ to the null model:

- Let g be a fixed constant and take n fixed.
- Let $F = \frac{R^2_\gamma / p_\gamma}{(1 - R^2_\gamma)/(n - 1 - p_\gamma)}$

Problem with g-Prior with arbitrary g

The Bayes factor for comparing $\mathcal{M}_\gamma$ to the null model:

- Let g be a fixed constant and take n fixed.
- Let $F = \frac{R^2_\gamma/p_\gamma}{(1-R^2_\gamma)/(n-1-p_\gamma)}$
- As $R^2_\gamma \to 1$, $F \to \infty$ LR test would reject H_0 where F is the usual F statistic for comparing model $\mathcal{M}_\gamma$ to $\mathcal{M}_0$

Problem with g-Prior with arbitrary g

The Bayes factor for comparing $\mathcal{M}_\gamma$ to the null model:

- Let g be a fixed constant and take n fixed.
- Let $F = \frac{R^2_\gamma / p_\gamma}{(1 - R^2_\gamma)/(n - 1 - p_\gamma)}$
- As $R^2_\gamma \to 1$, $F \to \infty$ LR test would reject H_0 where F is the usual F statistic for comparing model $\mathcal{M}_\gamma$ to $\mathcal{M}_0$
- Bayes Factor would go to $(1 + g)^{(n - p_\gamma - 1)/2}$ as $F \to \infty$ (bounded for fixed g, n and p_γ

Problem with g-Prior with arbitrary g

The Bayes factor for comparing $\mathcal{M}_\gamma$ to the null model:

- Let g be a fixed constant and take n fixed.
- Let $F = \frac{R^2_\gamma / p_\gamma}{(1 - R^2_\gamma)/(n - 1 - p_\gamma)}$
- As $R^2_\gamma \to 1$, $F \to \infty$ LR test would reject H_0 where F is the usual F statistic for comparing model $\mathcal{M}_\gamma$ to $\mathcal{M}_0$
- Bayes Factor would go to $(1 + g)^{(n - p_\gamma - 1)/2}$ as $F \to \infty$ (bounded for fixed g, n and p_γ

Bayes and Frequentist would not agree in this limit

Problem with g-Prior with arbitrary g

The Bayes factor for comparing $\mathcal{M}_\gamma$ to the null model:

- Let g be a fixed constant and take n fixed.
- Let $F = \frac{R^2_\gamma / p_\gamma}{(1 - R^2_\gamma)/(n - 1 - p_\gamma)}$
- As $R^2_\gamma \to 1$, $F \to \infty$ LR test would reject H_0 where F is the usual F statistic for comparing model $\mathcal{M}_\gamma$ to $\mathcal{M}_0$
- Bayes Factor would go to $(1 + g)^{(n - p_\gamma - 1)/2}$ as $F \to \infty$ (bounded for fixed g, n and p_γ

Bayes and Frequentist would not agree in this limit

"Information paradox"

Resolution of Paradox

Liang et al (2008) show that paradox can be resolved with mixtures of g-priors

$$p(\boldsymbol{\beta}_\gamma \mid \phi) = \int_0^\infty \mathsf{N}(\boldsymbol{\beta}_\gamma; 0, g(\mathbf{X}_\gamma^T \mathbf{X}_\gamma)^{-1}/\phi) p(g) dg$$

Resolution of Paradox

Liang et al (2008) show that paradox can be resolved with mixtures of g-priors

$$p(\boldsymbol{\beta}_\gamma \mid \phi) = \int_0^\infty \mathsf{N}(\boldsymbol{\beta}_\gamma; 0, g(\mathbf{X}_\gamma^T \mathbf{X}_\gamma)^{-1}/\phi) p(g) dg$$

- $BF \to \infty$ if $\mathrm{R}^2 \to 1 \Leftrightarrow \mathrm{E}_g[(1+g)^{-p_\gamma/2}]$ diverges

Resolution of Paradox

Liang et al (2008) show that paradox can be resolved with mixtures of g-priors

$$p(\boldsymbol{\beta}_\gamma \mid \phi) = \int_0^\infty \mathrm{N}(\boldsymbol{\beta}_\gamma; 0, g(\mathbf{X}_\gamma^T \mathbf{X}_\gamma)^{-1}/\phi) p(g) dg$$

- $BF \to \infty$ if $\mathrm{R}^2 \to 1 \Leftrightarrow \mathrm{E}_g[(1+g)^{-p_\gamma/2}]$ diverges
- Zellner-Siow Cauchy prior

$$1/g \sim \mathrm{Gamma}(1/2, n/2)$$

Resolution of Paradox

Liang et al (2008) show that paradox can be resolved with mixtures of g-priors

$$p(\boldsymbol{\beta}_\gamma \mid \phi) = \int_0^\infty \mathrm{N}(\boldsymbol{\beta}_\gamma; 0, g(\mathbf{X}_\gamma^T \mathbf{X}_\gamma)^{-1}/\phi) p(g) dg$$

- $BF \to \infty$ if $\mathrm{R}^2 \to 1 \Leftrightarrow \mathrm{E}_g[(1+g)^{-p_\gamma/2}]$ diverges
- Zellner-Siow Cauchy prior

$$1/g \sim \mathrm{Gamma}(1/2, n/2)$$

- Hyper-g $p(g) \propto (1+g)^{a/2-1}$ if $2 < a \leq 3$

$$\frac{g}{1+g} \sim \mathrm{Beta}(1, \frac{a}{2} - 1)$$

Resolution of Paradox

Liang et al (2008) show that paradox can be resolved with mixtures of g-priors

$$p(\boldsymbol{\beta}_\gamma \mid \phi) = \int_0^\infty \mathsf{N}(\boldsymbol{\beta}_\gamma; 0, g(\mathbf{X}_\gamma^T \mathbf{X}_\gamma)^{-1}/\phi) p(g) dg$$

- $BF \to \infty$ if $\mathrm{R}^2 \to 1 \Leftrightarrow \mathsf{E}_g[(1+g)^{-p_\gamma/2}]$ diverges
- Zellner-Siow Cauchy prior

$$1/g \sim \mathsf{Gamma}(1/2, n/2)$$

- Hyper-g $p(g) \propto (1+g)^{a/2-1}$ if $2 < a \leq 3$

$$\frac{g}{1+g} \sim \mathsf{Beta}(1, \frac{a}{2} - 1)$$

- "hyper-g/n"

Resolution of Paradox

Liang et al (2008) show that paradox can be resolved with mixtures of g-priors

$$p(\boldsymbol{\beta}_\gamma \mid \phi) = \int_0^\infty \mathsf{N}(\boldsymbol{\beta}_\gamma; 0, g(\mathbf{X}_\gamma^T \mathbf{X}_\gamma)^{-1}/\phi) p(g) dg$$

- $BF \to \infty$ if $\mathsf{R}^2 \to 1 \Leftrightarrow \mathsf{E}_g[(1+g)^{-p_\gamma/2}]$ diverges
- Zellner-Siow Cauchy prior

$$1/g \sim \text{Gamma}(1/2, n/2)$$

- Hyper-g $p(g) \propto (1+g)^{a/2-1}$ if $2 < a \leq 3$

$$\frac{g}{1+g} \sim \text{Beta}(1, \frac{a}{2} - 1)$$

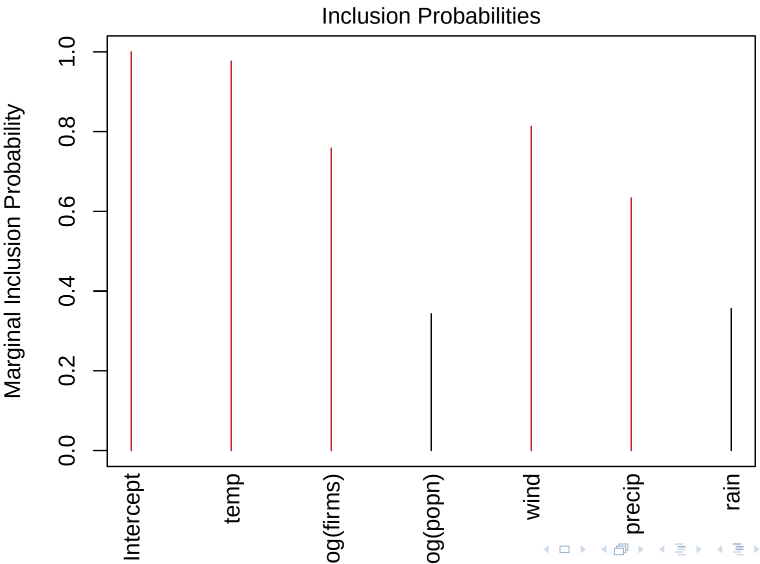
- "hyper-g/n"
- robust prior (Bayarri et al Annals of Statistics 2012)

Example

```r
library(BAS)
poll.ZS = bas.lm(log(SO2) ~ temp + log(firms) +
                             log(popn) + wind +
                             precip+ rain,
                   data=usair,
                   prior="ZS-null",
                   alpha=41,     # g = n
                   n.models=2^7,# enumerate (can omit)
                   modelprior=uniform(),
                   method="deterministic")    # fast enumera
```

use ‘prior = ”hyper-g”‘ and ‘a = 3‘ for hyper-g or ‘prior = ”hyper-g/n”‘ and ‘a=3‘ for hyper-g/n

```r
plot(poll.ZS, which=4)
```



Bayesian Model Averaging

- Posterior for $\mu = \mathbf{1}\alpha + \mathbf{X}\boldsymbol{\beta}$ is a mixture distribution

$$p(\mu \mid \mathbf{Y}) = \sum p(\mu \mid \mathbf{Y}, \mathcal{M}_\gamma) p(\mathcal{M}_\gamma \mid \mathbf{Y})$$

Bayesian Model Averaging

- Posterior for $\mu = \mathbf{1}\alpha + \mathbf{X}\boldsymbol{\beta}$ is a mixture distribution

$$p(\mu \mid \mathbf{Y}) = \sum p(\mu \mid \mathbf{Y}, \mathcal{M}_\gamma) p(\mathcal{M}_\gamma \mid \mathbf{Y})$$

  with expectation expressed as a weighted average

$$\mathsf{E}[\mu \mid \mathbf{Y}] = \mathbf{1}\hat{\alpha} + \mathbf{X} \sum \mathsf{E}[\boldsymbol{\beta} \mid \mathbf{Y}, \mathcal{M}_\gamma] p(\mathcal{M}_\gamma \mid \mathbf{Y})$$

Bayesian Model Averaging

- Posterior for $\mu = \mathbf{1}\alpha + \mathbf{X}\beta$ is a mixture distribution

$$p(\mu \mid \mathbf{Y}) = \sum p(\mu \mid \mathbf{Y}, \mathcal{M}_\gamma) p(\mathcal{M}_\gamma \mid \mathbf{Y})$$

  with expectation expressed as a weighted average

$$\mathsf{E}[\mu \mid \mathbf{Y}] = \mathbf{1}\hat{\alpha} + \mathbf{X} \sum \mathsf{E}[\beta \mid \mathbf{Y}, \mathcal{M}_\gamma] p(\mathcal{M}_\gamma \mid \mathbf{Y})$$

- Predictive Distribution for $\mathbf{Y}^*$

$$p(\mathbf{Y}^* \mid \mathbf{Y}) = \sum p(\mathbf{Y}^* \mid \mathbf{Y}, \mathcal{M}_\gamma) p(\mathcal{M}_\gamma \mid \mathbf{Y})$$

Bayesian Model Averaging

- Posterior for $\mu = \mathbf{1}\alpha + \mathbf{X}\boldsymbol{\beta}$ is a mixture distribution

$$p(\mu \mid \mathbf{Y}) = \sum p(\mu \mid \mathbf{Y}, \mathcal{M}_\gamma) p(\mathcal{M}_\gamma \mid \mathbf{Y})$$

with expectation expressed as a weighted average

$$\mathsf{E}[\mu \mid \mathbf{Y}] = \mathbf{1}\hat{\alpha} + \mathbf{X} \sum \mathsf{E}[\boldsymbol{\beta} \mid \mathbf{Y}, \mathcal{M}_\gamma] p(\mathcal{M}_\gamma \mid \mathbf{Y})$$

- Predictive Distribution for $\mathbf{Y}^*$

$$p(\mathbf{Y}^* \mid \mathbf{Y}) = \sum p(\mathbf{Y}^* \mid \mathbf{Y}, \mathcal{M}_\gamma) p(\mathcal{M}_\gamma \mid \mathbf{Y})$$

- Posterior Distribution of β_j

$$p(\beta_j \mid \mathbf{Y}) = p(\gamma_j = 0 \mid \mathbf{Y}) \delta_0(\beta) + \sum p(\beta_j \mid \mathbf{Y}, \mathcal{M}_\gamma) \gamma_j p(\mathcal{M}_\gamma \mid \mathbf{Y})$$

Estimator

- Find $\hat{\mu}$ that minimizes posterior expected loss

$$\mathsf{E}[(\mu - \hat{\mu})^T(\mu - \hat{\mu}) \mid \mathbf{Y}]$$

Estimator

- Find $\hat{\mu}$ that minimizes posterior expected loss

$$\mathsf{E}[(\mu - \hat{\mu})^T(\mu - \hat{\mu}) \mid \mathbf{Y}]$$

- Solution is posterior mean under BMA

$$\mathsf{E}[\mu \mid \mathbf{Y}] = \mathbf{1}\hat{\alpha} + \mathbf{X}\sum \mathsf{E}[\boldsymbol{\beta} \mid \mathbf{Y}, \mathcal{M}_\gamma] p(\mathcal{M}_\gamma \mid \mathbf{Y})$$

Estimator

- Find $\hat{\mu}$ that minimizes posterior expected loss

$$\mathsf{E}[(\mu - \hat{\mu})^T(\mu - \hat{\mu}) \mid \mathbf{Y}]$$

- Solution is posterior mean under BMA

$$\mathsf{E}[\mu \mid \mathbf{Y}] = \mathbf{1}\hat{\alpha} + \mathbf{X} \sum \mathsf{E}[\boldsymbol{\beta} \mid \mathbf{Y}, \mathcal{M}_\gamma] p(\mathcal{M}_\gamma \mid \mathbf{Y})$$

- If one model has probability 1, then BMA is equivalent to using the highest posterior probability model

Estimator

- Find $\hat{\mu}$ that minimizes posterior expected loss

$$\mathsf{E}[(\mu - \hat{\mu})^T(\mu - \hat{\mu}) \mid \mathbf{Y}]$$

- Solution is posterior mean under BMA

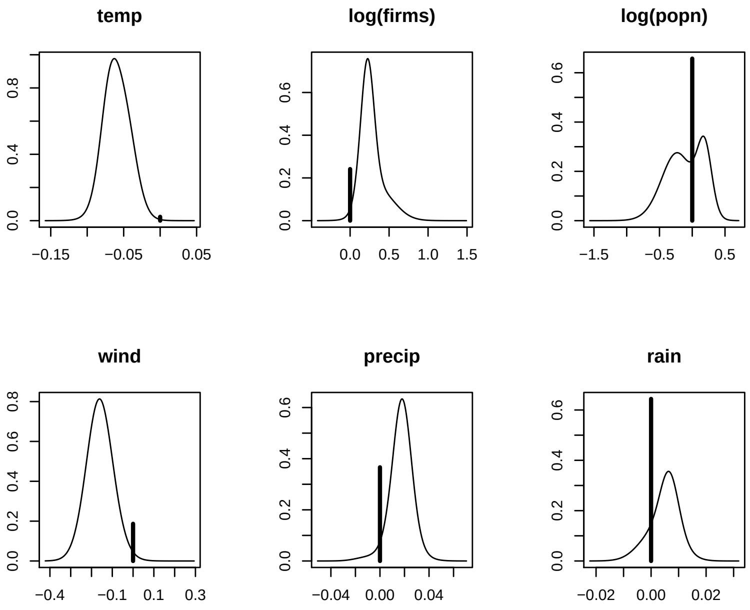
$$\mathsf{E}[\mu \mid \mathbf{Y}] = \mathbf{1}\hat{\alpha} + \mathbf{X} \sum \mathsf{E}[\boldsymbol{\beta} \mid \mathbf{Y}, \mathcal{M}_\gamma] p(\mathcal{M}_\gamma \mid \mathbf{Y})$$

- If one model has probability 1, then BMA is equivalent to using the highest posterior probability model
- incorporates estimates from other models when there is substantial uncertainty

Coefficients under BMA

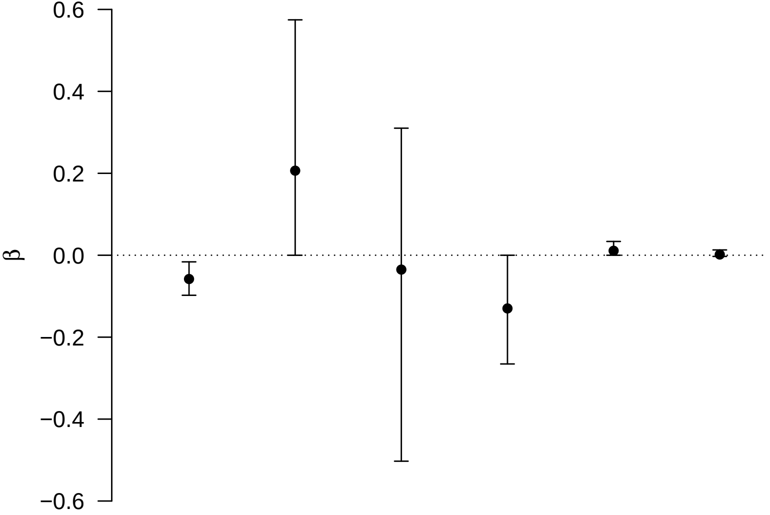
```
beta.ZS = coef(poll.ZS)
beta.ZS

##
##  Marginal Posterior Summaries of Coefficients:
##
##  Based on the top  64 models
##              post mean  post SD    post p(B != 0)
## Intercept     3.153004   0.082226   1.000000
## temp         -0.058053   0.020325   0.976833
## log(firms)    0.206384   0.177253   0.758554
## log(popn)    -0.035074   0.174760   0.342677
## wind         -0.129875   0.085195   0.813330
## precip        0.010898   0.011327   0.633639
## rain          0.001759   0.004034   0.356085
```

Posterior of Coefficients under BMA



Credible Intervals for Coefficients under BMA

```
plot(confint(beta.ZS, parm=2:7))
```



Selection and Model Uncertainty

- Select a model and $\hat{\mu}$ that minimizes posterior expected loss

$$\mathsf{E}[(\mu - \hat{\mu})^T(\mu - \hat{\mu}) \mid \mathbf{Y}]$$

Selection and Model Uncertainty

- Select a model and $\hat{\mu}$ that minimizes posterior expected loss

$$\mathsf{E}[(\mu - \hat{\mu})^T(\mu - \hat{\mu}) \mid \mathbf{Y}]$$

- BMA is "best" estimator without selection

Selection and Model Uncertainty

- Select a model and $\hat{\mu}$ that minimizes posterior expected loss

$$\mathsf{E}[(\mu - \hat{\mu})^T(\mu - \hat{\mu}) \mid \mathbf{Y}]$$

- BMA is "best" estimator without selection
- Best model and estimator is the posterior mean under the model that is closest to BMA under squared error loss

$$(\hat{\mu}_{BMA} - \hat{\mu}_{\mathcal{M}_\gamma})^T(\hat{\mu}_{BMA} - \hat{\mu}_{\mathcal{M}_\gamma})$$

Selection and Model Uncertainty

- Select a model and $\hat{\mu}$ that minimizes posterior expected loss

$$\mathsf{E}[(\mu - \hat{\mu})^T(\mu - \hat{\mu}) \mid \mathbf{Y}]$$

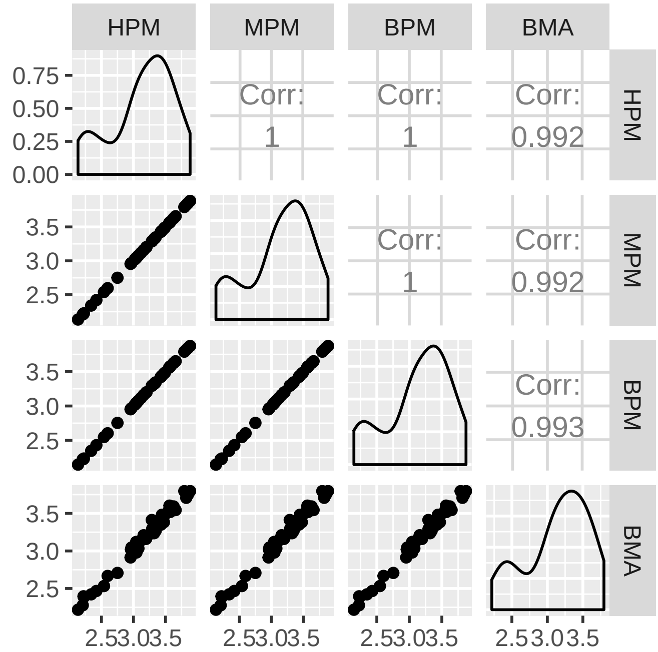
- BMA is "best" estimator without selection
- Best model and estimator is the posterior mean under the model that is closest to BMA under squared error loss

$$(\hat{\mu}_{BMA} - \hat{\mu}_{\mathcal{M}_\gamma})^T(\hat{\mu}_{BMA} - \hat{\mu}_{\mathcal{M}_\gamma})$$

- Often contains more predictors than the HPM or Median Probability Model

Best Predictive Model

```
#BPM
BPM = predict(poll.ZS, estimator = "BPM")
BPM$bestmodel

## [1] 0 1 2 4 5 6

(poll.ZS$namesx[attr(BPM$fit, 'model') +1])[-1]

## [1] "temp"        "log(firms)" "wind"        "precip"

#HPM
HPM = predict(poll.ZS, estimator = "HPM")
HPM$bestmodel

## [1] 0 1 2 4 5
```



HPM
MPM
BPM
BMA
Corr:
1
Corr:
1
Corr:
0.992
Corr:
1
Corr:
0.992
Corr:
0.993
0.75
0.50
0.25
0.00
3.5
3.0
2.5
2.5
3.0
3.5

Summary

- BMA shown in practice to have better out of sample predictions than selection (in many cases)

Summary

- BMA shown in practice to have better out of sample predictions than selection (in many cases)
- avoids selecting a single model and accounts for out uncertainty

Summary

- BMA shown in practice to have better out of sample predictions than selection (in many cases)
- avoids selecting a single model and accounts for out uncertainty
- if one model dominates BMA is very close to selection (asymptotically will put probability one on model that is "closest" to the true model)

Summary

- BMA shown in practice to have better out of sample predictions than selection (in many cases)
- avoids selecting a single model and accounts for out uncertainty
- if one model dominates BMA is very close to selection (asymptotically will put probability one on model that is "closest" to the true model)
- MCMC allows one to implement without enumerating all models

Summary

- BMA shown in practice to have better out of sample predictions than selection (in many cases)
- avoids selecting a single model and accounts for out uncertainty
- if one model dominates BMA is very close to selection (asymptotically will put probability one on model that is "closest" to the true model)
- MCMC allows one to implement without enumerating all models
- BMA depends on prior on coefficients, variance and models (sensitivity to choice?)

Summary

- BMA shown in practice to have better out of sample predictions than selection (in many cases)
- avoids selecting a single model and accounts for out uncertainty
- if one model dominates BMA is very close to selection (asymptotically will put probability one on model that is "closest" to the true model)
- MCMC allows one to implement without enumerating all models
- BMA depends on prior on coefficients, variance and models (sensitivity to choice?)
- Mixtures of g priors preferred to usual g prior