Bayesian Variable Selection &
Bayesian Model Averaging

Hoff Chapter 9, Mixtures of g-Priors Liang et al JASA

October 21, 2019

Outline

- Zellner's g-prior in Bayesian Regression
- Model Selection

Conjugate Posterior Distribution

Prior Distribution Normal-Gamma

$$\begin{aligned} \boldsymbol{\beta} \mid \phi &\sim \mathrm{N}(\mathbf{b}_0, (\phi\Phi_0)^{-1}) \\ \phi &\sim \mathbf{G}\left(\frac{\nu_0}{2}, \frac{\nu_0\hat{\sigma}_0^2}{2}\right) \end{aligned}$$

$$\boxed{\begin{aligned} \Phi_n &= \mathbf{X}^T\mathbf{X} + \Phi_0 \\ \mathbf{b}_n &= \Phi_n^{-1}(\mathbf{X}^T\mathbf{X}\hat{\beta} + \Phi_0\mathbf{b}_0) \\ \mathrm{SSE}_n &= \mathrm{SSE} + \mathrm{SS}_0 + \hat{\beta}^T\mathbf{X}^T\mathbf{X}\hat{\beta} + \mathbf{b}_0^T\Phi_0\mathbf{b}_0 - \mathbf{b}_n^T\Phi_n\mathbf{b}_n \\ \nu_n &= n + \nu_0 \\ \hat{\sigma}_n^2 &= \mathrm{SSE}_n/\nu_n \end{aligned}}$$

Posterior Distribution Normal-Gamma

$$\begin{aligned} \boldsymbol{\beta} \mid \phi, \mathbf{Y} &\sim \mathrm{N}(\mathbf{b}_n, (\phi\Phi_n)^{-1}) \\ \phi \mid \mathbf{Y} &\sim \mathbf{G}\left(\frac{n+\nu_0}{2}, \frac{\hat{\sigma}_n^2}{2}\right) \end{aligned}$$

Posterior predictive

Includes limiting cases such as the Independent Jeffrey's prior

Zellner's g-prior II

why center & then add g

Centered model: $\mathbf{Y} = \mathbf{1}_n \alpha + \mathbf{X}_c \boldsymbol{\beta} + \epsilon$

where $\mathbf{X}_c$ is the centered design matrix where all variables have had their mean subtracted

$p(\phi) \propto 1/\phi$ $p(\alpha \mid \phi) \propto 1$ $\boldsymbol{\beta} \mid \alpha, \phi, g \sim \mathrm{N}(0, g\phi^{-1}(\mathbf{X}_c{}'\mathbf{X}_c)^{-1})$

shrink towards zero

$$\alpha \mid \mathbf{Y}, \phi \sim N(\bar{y}, 1/(\phi n))$$

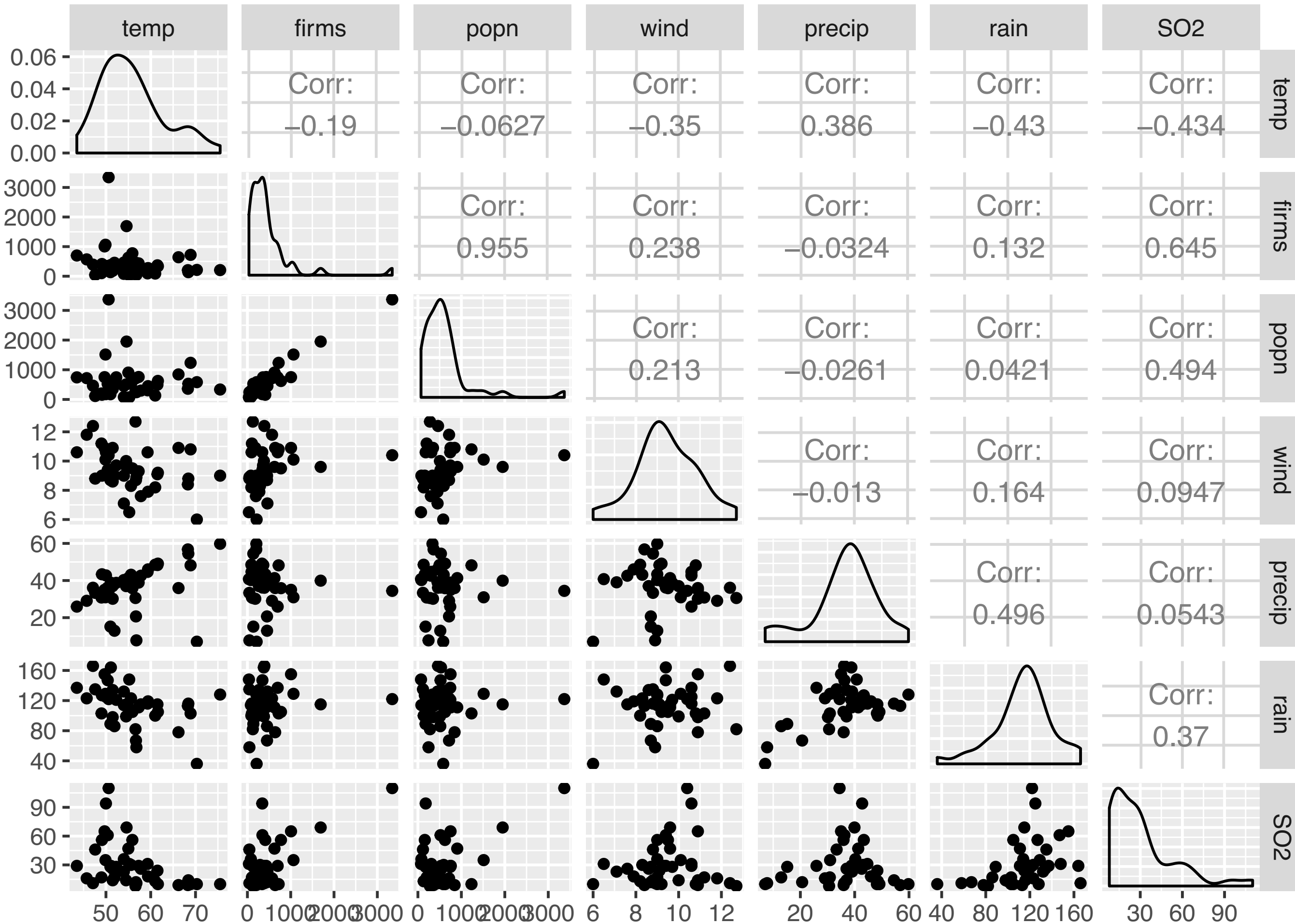
$$\boldsymbol{\beta} \mid \mathbf{Y}, \phi \sim N\left(\frac{g}{1+g}\hat{\boldsymbol{\beta}}, \phi^{-1}\frac{g}{1+g}(\mathbf{X}_c^T\mathbf{X}_c)^{-1}\right)$$

$$\phi \mid \mathbf{Y} \sim \text{Gamma}\left(\frac{n-1}{2}, \frac{\text{SSE} + \frac{1}{1+g}\hat{\boldsymbol{\beta}}^T(\mathbf{X}_c^T\mathbf{X}_c)\hat{\boldsymbol{\beta}}}{2}\right)$$

$$\boldsymbol{\beta} \mid \mathbf{Y} \sim t(n-1, \frac{g}{1+g}\hat{\boldsymbol{\beta}}, \hat{\sigma}_n^2\frac{g}{1+g}(\mathbf{X}_c^T\mathbf{X}_c)^{-1})$$

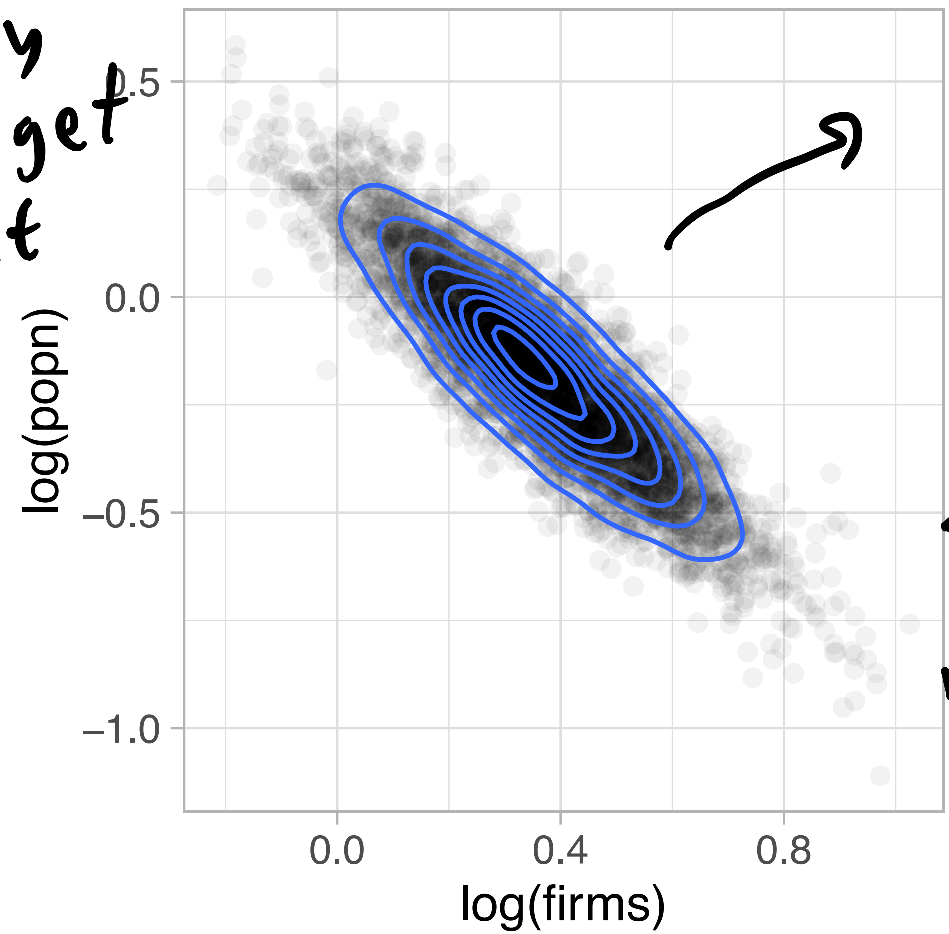
$$\hat{\sigma}_n^2 = \frac{\text{SSE} + \frac{1}{1+g}\hat{\boldsymbol{\beta}}^T(\mathbf{X}_c^T\mathbf{X}_c)\hat{\boldsymbol{\beta}}}{n-1}$$

US Air Example



joint posterior draws of beta's under g-prior

not only do we get $\hat{\beta}$ but we also get $\operatorname{cov}(\hat{\beta})$



β_p, β_f are correlated negatively

Bayesian Variable Selection

$\beta_j \mid Y$ → very correlated $\beta \mid Y$
$\beta_i \mid Y$ i & j are
redundant on

- Avoid the use of redundant variables (problems with interpretations) → need covariance info
- Inclusion of un-necessary terms yields less precise estimates, particularly if explanatory variables are highly correlated with each other
- reduced MSE: reduced variance but possibly higher bias
- it is too "expensive" to use all variables

1) many variables are correlated with noise
2) reduce dim. (cost + variance reduction)
3) redundant variables

Bayesian Model Choice

pull out a subset of p variables

- Models for the variable selection problem are based on a subset of the $\mathbf{X}_1, \ldots \mathbf{X}_p$ variables
- Encode models with a vector $\boldsymbol{\gamma} = (\gamma_1, \ldots \gamma_p)$ where $\gamma_j \in \{0, 1\}$ is an indicator for whether variable $\mathbf{X}_j$ should be included in the model $\mathcal{M}_\gamma$. $\gamma_j = 0 \Leftrightarrow \beta_j = 0$
- Each value of γ represents one of the 2^p models. $\gamma \in \{0,1\}^p$
- Under model $\mathcal{M}_\gamma$:

$$\mathbf{Y} \mid \alpha, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma} \sim \mathrm{N}(\mathbf{1}\alpha + \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma, \sigma^2 \mathbf{I})$$

Where $\mathbf{X}_\gamma$ is design matrix using the columns in $\mathbf{X}$ where $\gamma_j = 1$ and $\boldsymbol{\beta}_\gamma$ is the subset of $\boldsymbol{\beta}$ that are non-zero.

Select non-zero $\gamma_i \to$ non-zero β_i

Posterior Probabilities of Models

marginal likelihood

Bayes rule

- Posterior model probabilities

$$p(\mathcal{M}_j \mid \mathbf{Y}) = \frac{\boxed{p(\mathbf{Y} \mid \mathcal{M}_j)}\, p(\mathcal{M}_j)}{\underbrace{\sum_j p(\mathbf{Y} \mid \mathcal{M}_j) p(\mathcal{M}_j)}_{P(Y)}}$$

total prob.

Marginal likelihod of a model is proportional to

$$p(\mathbf{Y} \mid \mathcal{M}_\gamma) = \underbrace{\iint p(\mathbf{Y} \mid \boldsymbol{\beta}_\gamma, \sigma^2) p(\boldsymbol{\beta}_\gamma \mid \gamma, \sigma^2) p(\sigma^2 \mid \gamma) d\boldsymbol{\beta}\, d\sigma^2}_{\text{integrate out } \beta}$$

- Bayes Factor $BF[i : j]$

$$\frac{P(\mathcal{M}_i \mid \mathbf{Y})}{P(\mathcal{M}_j \mid \mathbf{Y})} = \frac{p(\mathbf{Y} \mid \mathcal{M}_i)}{p(\mathbf{Y} \mid \mathcal{M}_j)} \times \frac{P(\mathcal{M}_i)}{P(\mathcal{M}_j)}$$

no need to compute P(Y)

Posterior Odds = Bayes Factor × Prior odds

- Probability $\beta_j \neq 0$: $\sum_{\mathcal{M}_j : \beta_j \neq 0} p(\mathcal{M}_j \mid \mathbf{Y})$ (marginal posterior inclusion probability) for all models add up

$p(M_j | Y)$ if $\gamma_j \neq 0$

Prior Distributions

- Bayesian Model choice requires proper prior distributions on regression coefficients (exception parameters that are included in all models)
- Vague but proper priors may lead to paradoxes!
- Conjugate Normal-Gammas lead to closed form expressions for marginal likelihoods, Zellner's g-prior is one of the most popular.

Zellner's g-prior within Models

Centered model:

$$\mathbf{Y} = \mathbf{1}_n\alpha + \mathbf{X}^c_\gamma \boldsymbol{\beta}_\gamma + \epsilon$$

centered

rest. on $\gamma_i = 1$

- Common parameters

$$p(\alpha, \phi) \propto \phi^{-1}$$

Vague precision

- Model Specific parameters

$$\boldsymbol{\beta}_\gamma \mid \alpha, \phi, \boldsymbol{\gamma} \sim \mathsf{N}(0, g\phi^{-1}(\mathbf{X}^{c\,\prime}_\gamma \mathbf{X}^c_\gamma)^{-1})$$

closed form

- Marginal likelihood of $\mathcal{M}_\gamma$ is proportional to

$$p(\mathbf{Y} \mid \mathcal{M}_\gamma) = C(1+g)^{\frac{n-p-1}{2}}(1+g(1-R^2_\gamma))^{-\frac{(n-1)}{2}}$$

where R^2_γ is the usual R^2 for model $\mathcal{M}_\gamma$ and C is a constant that is $p(\mathbf{Y} \mid \mathcal{M}_0)$ (model with intercept alone)

- uniform distribution over space of models $p(\mathcal{M}_\gamma) = 1/(2^p)$

USair Data: Enumeration of All Models

2^p models
so # of comparisons is large

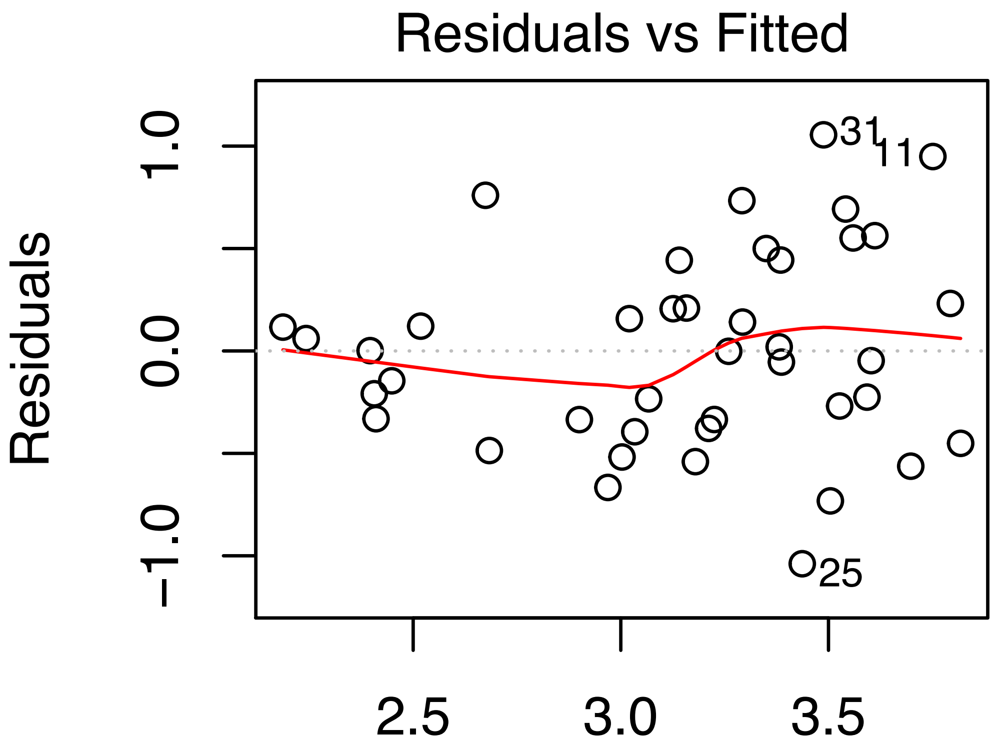
```
library(devtools)

## Loading required package:  usethis

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

## Warning in model.matrix.default(mt, mf,
contrasts):  non-list contrasts argument ignored
```

Cost to compute posterior quantities for each model

$O(p^3 n) \times 2^p$ = 2^{100}

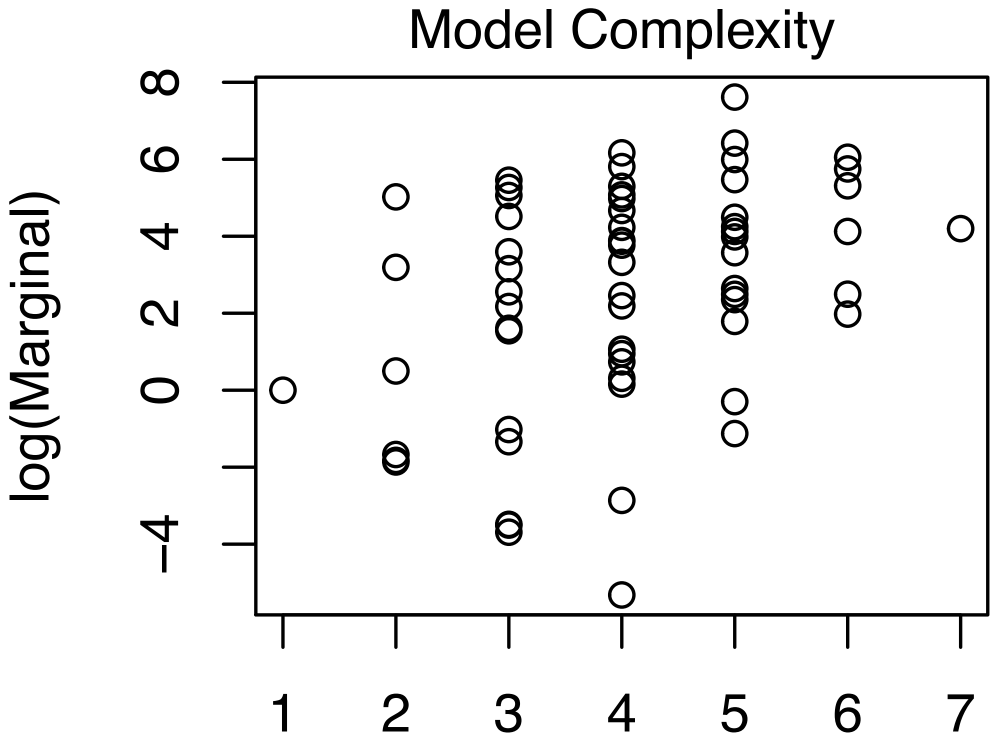
residual plot)

```r
plot(poll.bma, which=1)
```



Model Complexity)

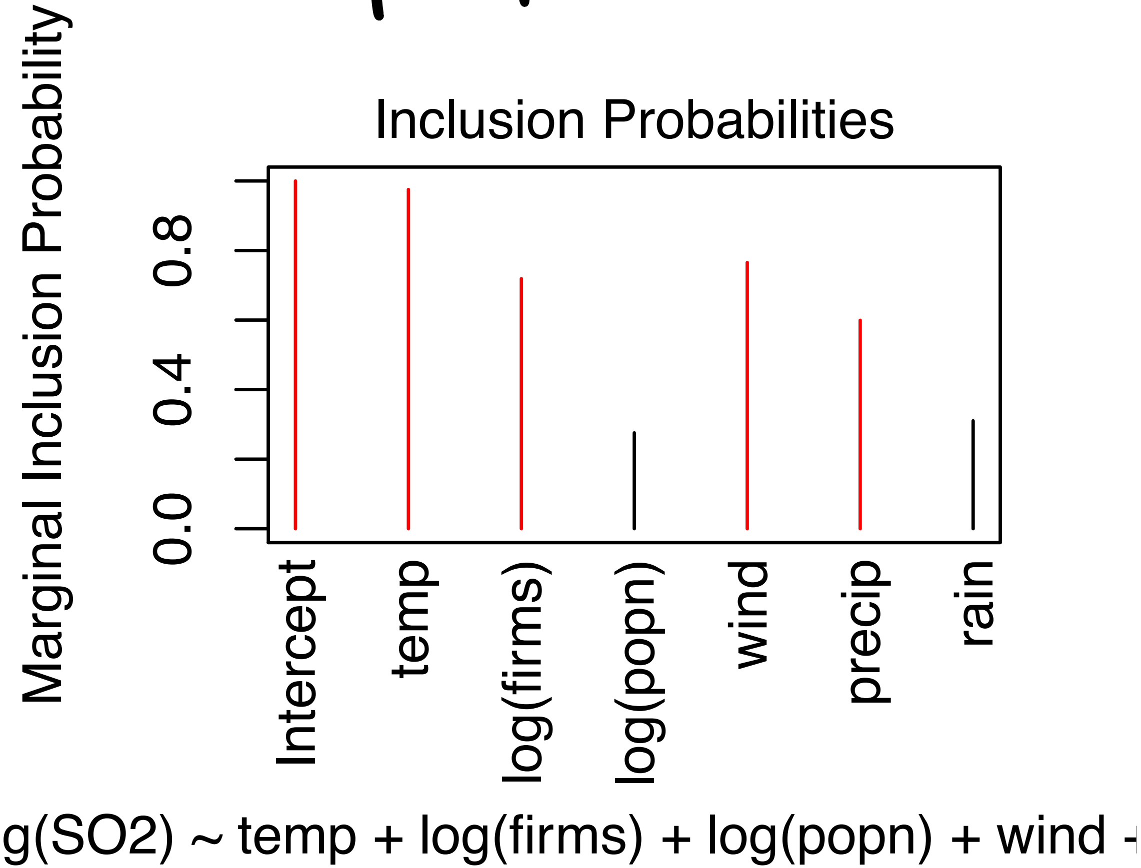
```r
plot(poll.bma, which=3)
```



Inclusion Probabilities)

```
plot(poll.bma, which=4)
```

$p(\gamma_i \neq 0)$



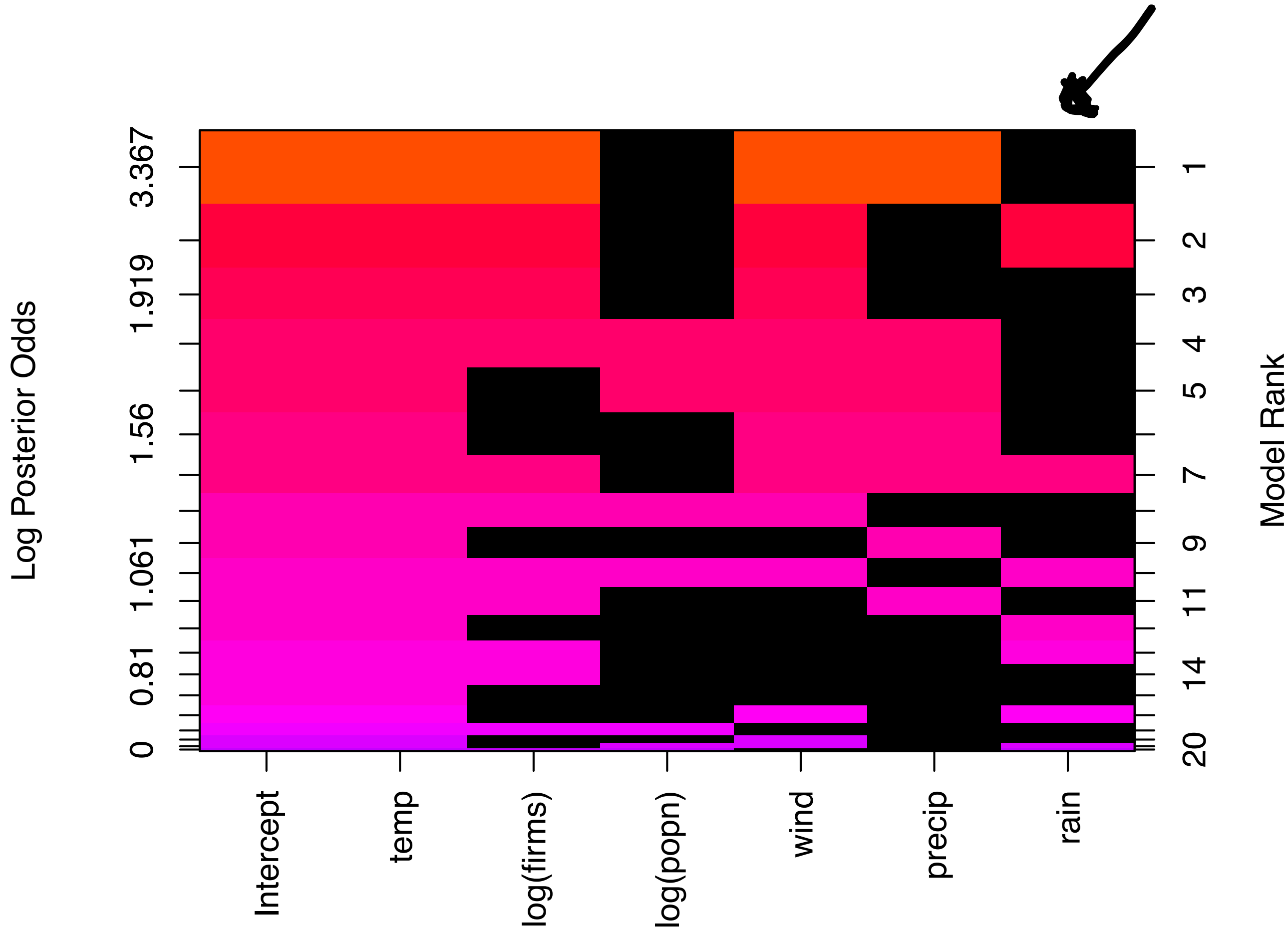
Model Space

```r
summary(poll.bma)
```

```
##                P(B != 0 | Y)  model 1   model 2   model 3
## Intercept          1.0000000 1.000000 1.0000000 1.0000000 1.
## temp               0.9755041 1.000000 1.0000000 1.0000000 1.
## log(firms)         0.7190313 1.000000 1.0000000 1.0000000 1.
## log(popn)          0.2756811 0.000000 0.0000000 0.0000000 1.
## wind               0.7654485 1.000000 1.0000000 1.0000000 1.
## precip             0.5993801 1.000000 0.0000000 0.0000000 1.
## rain               0.3103574 0.000000 1.0000000 0.0000000 0.
## BF                        NA 1.000000 0.3022674 0.2349056 0.
## PostProbs                 NA 0.275800 0.0834000 0.0648000 0.
## R2                        NA 0.542700 0.5130000 0.4558000 0.
## dim                       NA 5.000000 5.0000000 4.0000000 6.
## logmarg                   NA 7.616228 6.4197847 6.1676565 6.
```

Summary

```r
image(poll.bma)
```

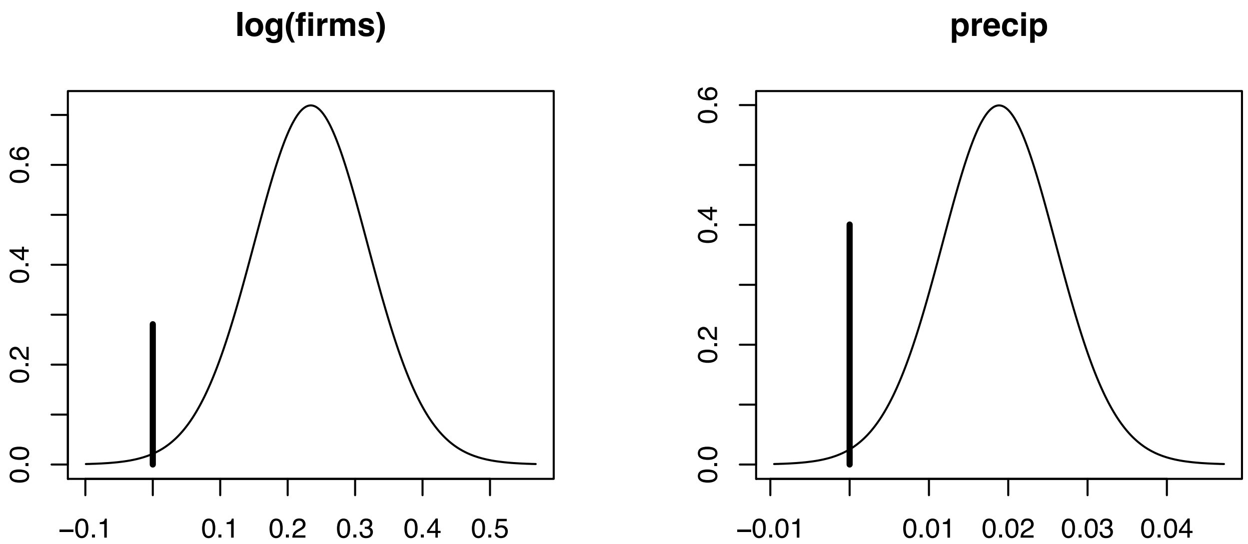


Coefficients

```
beta = coef(poll.bma, n.models=1)
beta

##
##  Marginal Posterior Summaries of Coefficients:
##
##  Using  BMA
##
##  Based on the top  1 models
##              post mean  post SD   post p(B != 0)
## Intercept     3.15300    0.07818   1.00000
## temp         -0.07130    0.01268   0.97550
## log(firms)    0.23428    0.08573   0.71903
## log(popn)     0.00000    0.00000   0.27568
## wind         -0.17998    0.06128   0.76545
## precip        0.01884    0.00729   0.59938
## rain          0.00000    0.00000   0.31036
```

Coefficients

```
par(mfrow=c(2,2));  plot(beta, subset=c(3, 6))
```



Bayesian Confidence Intervals

```
confint(beta)

##                        2.5%       97.5%        beta
## Intercept   2.994993257  3.31101398  3.15300362
## temp       -0.096926645 -0.04567203 -0.07129934
## log(firms)  0.061014518  0.40753936  0.23427694
## log(popn)   0.000000000  0.00000000  0.00000000
## wind       -0.303835463 -0.05612195 -0.17997871
## precip      0.004105874  0.03357242  0.01883915
## rain        0.000000000  0.00000000  0.00000000
## attr(,"Probability")
## [1] 0.95
## attr(,"class")
## [1] "confint.bas"
```

Bayesian Variable Selection and Model Averaging

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

October 23, 2019

Bayesian Model Choice

- Models for the variable selection problem are based on a subset of the $\mathbf{X}_1, \ldots \mathbf{X}_p$ variables

Bayesian Model Choice

- Models for the variable selection problem are based on a subset of the $\mathbf{X}_1, \ldots \mathbf{X}_p$ variables
- Encode models with a vector $\boldsymbol{\gamma} = (\gamma_1, \ldots \gamma_p)$ where $\gamma_j \in \{0, 1\}$ is an indicator for whether variable $\mathbf{X}_j$ should be included in the model $\mathcal{M}_\gamma$. $\gamma_j = 0 \Leftrightarrow \beta_j = 0$

Bayesian Model Choice

- Models for the variable selection problem are based on a subset of the $\mathbf{X}_1, \ldots \mathbf{X}_p$ variables
- Encode models with a vector $\boldsymbol{\gamma} = (\gamma_1, \ldots \gamma_p)$ where $\gamma_j \in \{0, 1\}$ is an indicator for whether variable $\mathbf{X}_j$ should be included in the model $\mathcal{M}_\gamma$. $\gamma_j = 0 \Leftrightarrow \beta_j = 0$
- Each value of $\boldsymbol{\gamma}$ represents one of the 2^p models.

Bayesian Model Choice

- Models for the variable selection problem are based on a subset of the $\mathbf{X}_1, \ldots \mathbf{X}_p$ variables
- Encode models with a vector $\boldsymbol{\gamma} = (\gamma_1, \ldots \gamma_p)$ where $\gamma_j \in \{0, 1\}$ is an indicator for whether variable $\mathbf{X}_j$ should be included in the model $\mathcal{M}_\gamma$. $\gamma_j = 0 \Leftrightarrow \beta_j = 0$
- Each value of $\boldsymbol{\gamma}$ represents one of the 2^p models.
- Under model $\mathcal{M}_\gamma$:

$$\mathbf{Y} \mid \alpha, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma} \sim \mathrm{N}(\mathbf{1}\alpha + \mathbf{X}_{\boldsymbol{\gamma}} \boldsymbol{\beta}_\gamma, \sigma^2 \mathbf{I})$$

  Where $\mathbf{X}_\gamma$ is design matrix using the columns in $\mathbf{X}$ where $\gamma_j = 1$ and $\boldsymbol{\beta}_\gamma$ is the subset of $\boldsymbol{\beta}$ that are non-zero.

Posterior Probabilities of Models

- Posterior model probabilities

$$p(\mathcal{M}_j \mid \mathbf{Y}) = \frac{p(\mathbf{Y} \mid \mathcal{M}_j)p(\mathcal{M}_j)}{\sum_j p(\mathbf{Y} \mid \mathcal{M}_j)p(\mathcal{M}_j)}$$

Posterior Probabilities of Models

- Posterior model probabilities

$$p(\mathcal{M}_j \mid \mathbf{Y}) = \frac{p(\mathbf{Y} \mid \mathcal{M}_j)p(\mathcal{M}_j)}{\sum_j p(\mathbf{Y} \mid \mathcal{M}_j)p(\mathcal{M}_j)}$$

Marginal likelihod of a model is proportional to

$$p(\mathbf{Y} \mid \mathcal{M}_\gamma) = \iiint p(\mathbf{Y} \mid \alpha, \boldsymbol{\beta}_\gamma, \sigma^2)p(\boldsymbol{\beta}_\gamma \mid \boldsymbol{\gamma}, \sigma^2)p(\alpha, \sigma^2 \mid \boldsymbol{\gamma})d\boldsymbol{\beta}\, d\alpha, d$$

Posterior Probabilities of Models

- Posterior model probabilities

$$p(\mathcal{M}_j \mid \mathbf{Y}) = \frac{p(\mathbf{Y} \mid \mathcal{M}_j)p(\mathcal{M}_j)}{\sum_j p(\mathbf{Y} \mid \mathcal{M}_j)p(\mathcal{M}_j)}$$

  Marginal likelihod of a model is proportional to

$$p(\mathbf{Y} \mid \mathcal{M}_\gamma) = \iiint p(\mathbf{Y} \mid \alpha, \boldsymbol{\beta}_\gamma, \sigma^2)p(\boldsymbol{\beta}_\gamma \mid \boldsymbol{\gamma}, \sigma^2)p(\alpha, \sigma^2 \mid \boldsymbol{\gamma})d\boldsymbol{\beta}\, d\alpha, d$$

- Bayes Factor $BF[i:j] = p(\mathbf{Y} \mid \mathcal{M}_i)/p(\mathbf{Y} \mid \mathcal{M}_j)$

$$\frac{P(\mathcal{M}_i \mid \mathbf{Y})}{P(\mathcal{M}_j \mid \mathbf{Y})} = \frac{p(\mathbf{Y} \mid \mathcal{M}_i)}{p(\mathbf{Y} \mid \mathcal{M}_j)} \times \frac{P(\mathcal{M}_i)}{P(\mathcal{M}_j)}$$

  Posterior Odds $=$ Bayes Factor $\times$ Prior odds

Posterior Probabilities of Models

- Posterior model probabilities

$$p(\mathcal{M}_j \mid \mathbf{Y}) = \frac{p(\mathbf{Y} \mid \mathcal{M}_j) p(\mathcal{M}_j)}{\sum_j p(\mathbf{Y} \mid \mathcal{M}_j) p(\mathcal{M}_j)}$$

  Marginal likelihod of a model is proportional to

$$p(\mathbf{Y} \mid \mathcal{M}_\gamma) = \iiint p(\mathbf{Y} \mid \alpha, \boldsymbol{\beta}_\gamma, \sigma^2) p(\boldsymbol{\beta}_\gamma \mid \boldsymbol{\gamma}, \sigma^2) p(\alpha, \sigma^2 \mid \boldsymbol{\gamma}) d\boldsymbol{\beta}\, d\alpha, d$$

- Bayes Factor $BF[i:j] = p(\mathbf{Y} \mid \mathcal{M}_i)/p(\mathbf{Y} \mid \mathcal{M}_j)$

$$\frac{P(\mathcal{M}_i \mid \mathbf{Y})}{P(\mathcal{M}_j \mid \mathbf{Y})} = \frac{p(\mathbf{Y} \mid \mathcal{M}_i)}{p(\mathbf{Y} \mid \mathcal{M}_j)} \times \frac{P(\mathcal{M}_i)}{P(\mathcal{M}_j)}$$

  Posterior Odds = Bayes Factor $\times$ Prior odds

- Probability $\beta_j \neq 0$: $\sum_{\mathcal{M}_j : \beta_j \neq 0} p(\mathcal{M}_j \mid \mathbf{Y})$ (marginal posterior inclusion probability)

Posterior Probabilities of Models

- Posterior model probabilities

$$p(\mathcal{M}_j \mid \mathbf{Y}) = \frac{p(\mathbf{Y} \mid \mathcal{M}_j) p(\mathcal{M}_j)}{\sum_j p(\mathbf{Y} \mid \mathcal{M}_j) p(\mathcal{M}_j)}$$

  Marginal likelihod of a model is proportional to

$$p(\mathbf{Y} \mid \mathcal{M}_\gamma) = \iiint p(\mathbf{Y} \mid \alpha, \boldsymbol{\beta}_\gamma, \sigma^2) p(\boldsymbol{\beta}_\gamma \mid \boldsymbol{\gamma}, \sigma^2) p(\alpha, \sigma^2 \mid \boldsymbol{\gamma}) d\boldsymbol{\beta}\, d\alpha, d$$

- Bayes Factor $BF[i : j] = p(\mathbf{Y} \mid \mathcal{M}_i)/p(\mathbf{Y} \mid \mathcal{M}_j)$

$$\boxed{\frac{P(\mathcal{M}_i \mid \mathbf{Y})}{P(\mathcal{M}_j \mid \mathbf{Y})}} = \frac{p(\mathbf{Y} \mid \mathcal{M}_i)}{p(\mathbf{Y} \mid \mathcal{M}_j)} \times \frac{P(\mathcal{M}_i)}{P(\mathcal{M}_j)}$$

  Posterior Odds $=$ Bayes Factor $\times$ Prior odds

- Probability $\beta_j \neq 0$: $\sum_{\mathcal{M}_j : \beta_j \neq 0} p(\mathcal{M}_j \mid \mathbf{Y})$ (marginal posterior inclusion probability)

Zellner's g-prior within Models

Centered model:

$$\mathbf{Y} = \mathbf{1}_n \alpha + \mathbf{X}^c_{\boldsymbol{\gamma}} \boldsymbol{\beta}_\gamma + \epsilon$$

- Common parameters

$$p(\alpha, \phi) \propto \phi^{-1}$$

Zellner's g-prior within Models

Centered model:

$$\mathbf{Y} = \mathbf{1}_n \alpha + \mathbf{X}^c_{\boldsymbol{\gamma}} \boldsymbol{\beta}_\gamma + \epsilon$$

- Common parameters

$$p(\alpha, \phi) \propto \phi^{-1}$$

- Model Specific parameters

$$\boldsymbol{\beta}_\gamma \mid \alpha, \phi, \boldsymbol{\gamma} \sim \mathsf{N}(0, g\phi^{-1}({\mathbf{X}^c_{\boldsymbol{\gamma}}}' \mathbf{X}^c_{\boldsymbol{\gamma}})^{-1})$$

Zellner's g-prior within Models

Centered model:

$$\mathbf{Y} = \mathbf{1}_n\alpha + \mathbf{X}^c_{\boldsymbol{\gamma}}\boldsymbol{\beta}_\gamma + \epsilon$$

- Common parameters

$$p(\alpha, \phi) \propto \phi^{-1}$$

- Model Specific parameters

$$\boldsymbol{\beta}_\gamma \mid \alpha, \phi, \boldsymbol{\gamma} \sim \mathsf{N}(0, g\phi^{-1}({\mathbf{X}^c_{\boldsymbol{\gamma}}}'\mathbf{X}^c_{\boldsymbol{\gamma}})^{-1})$$

- Marginal likelihood of $\mathcal{M}_\gamma$ is proportional to

$$p(\mathbf{Y} \mid \mathcal{M}_\gamma) = C(1+g)^{\frac{n-p-1}{2}}(1 + g(1 - R^2_\gamma))^{-\frac{(n-1)}{2}}$$

where R^2_γ is the usual R^2 for model $\mathcal{M}_\gamma$ and C is a constant that is $p(\mathbf{Y} \mid \mathcal{M}_0)$ (model with intercept alone)

Zellner's g-prior within Models

Centered model:

$$\mathbf{Y} = \mathbf{1}_n \alpha + \mathbf{X}^c_{\boldsymbol{\gamma}} \boldsymbol{\beta}_\gamma + \epsilon$$

- Common parameters

$$p(\alpha, \phi) \propto \phi^{-1}$$

- Model Specific parameters

$$\boldsymbol{\beta}_\gamma \mid \alpha, \phi, \boldsymbol{\gamma} \sim \mathsf{N}(0, g\phi^{-1}(\mathbf{X}^{c\,\prime}_{\boldsymbol{\gamma}} \mathbf{X}^c_{\boldsymbol{\gamma}})^{-1})$$

BF

- Marginal likelihood of $\mathcal{M}_\gamma$ is proportional to

$$p(\mathbf{Y} \mid \mathcal{M}_\gamma) = C \boxed{(1+g)^{\frac{n-p-1}{2}}} (1 + g(1 - R^2_\gamma))^{-\frac{(n-1)}{2}}$$

  where R^2_γ is the usual R^2 for model $\mathcal{M}_\gamma$ and C is a constant that is $p(\mathbf{Y} \mid \mathcal{M}_0)$ (model with intercept alone)
- uniform distribution over space of models $p(\mathcal{M}_\gamma) = 1/(2^p)$

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 $\mathcal{M}_0$

LR test: $\ell = \log \frac{\ell(M_1)}{\ell(M_2)}$

frequentist

$\ell = \log \frac{p(M_\gamma \mid Y)}{p(M_0 \mid Y)} \to F$

idea if you have a good model selection crit. $\ell = \log \frac{p(me)}{\ldots}$ me = better than

λ log $\frac{\quad}{p(M_0)}$ as $n \to \infty$ null

Problem with g-Prior with arbitrary g

$$\ell = \log \frac{p(M_t)}{p(M_{tt})} \to \infty \quad \to 0$$

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"

what is g?

$\frac{g}{1+g} \hat{\beta}$ is post. mean

Resolution of Paradox

prior over g

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

similar to ϕ

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

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

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

- $BF \to \infty$ 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 $2 < a \leq 3$

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

Information paradox is about robust priors. putting a dist

Resolution of Paradox

over g is a good idea

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

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

$$\frac{g}{1+g} \approx 1$$

$$=0$$

- $BF \to \infty$ 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 $2 < a \leq 3$

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

- "hyper-g/n"

Resolution of Paradox

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

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

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

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

g depends on n

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

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

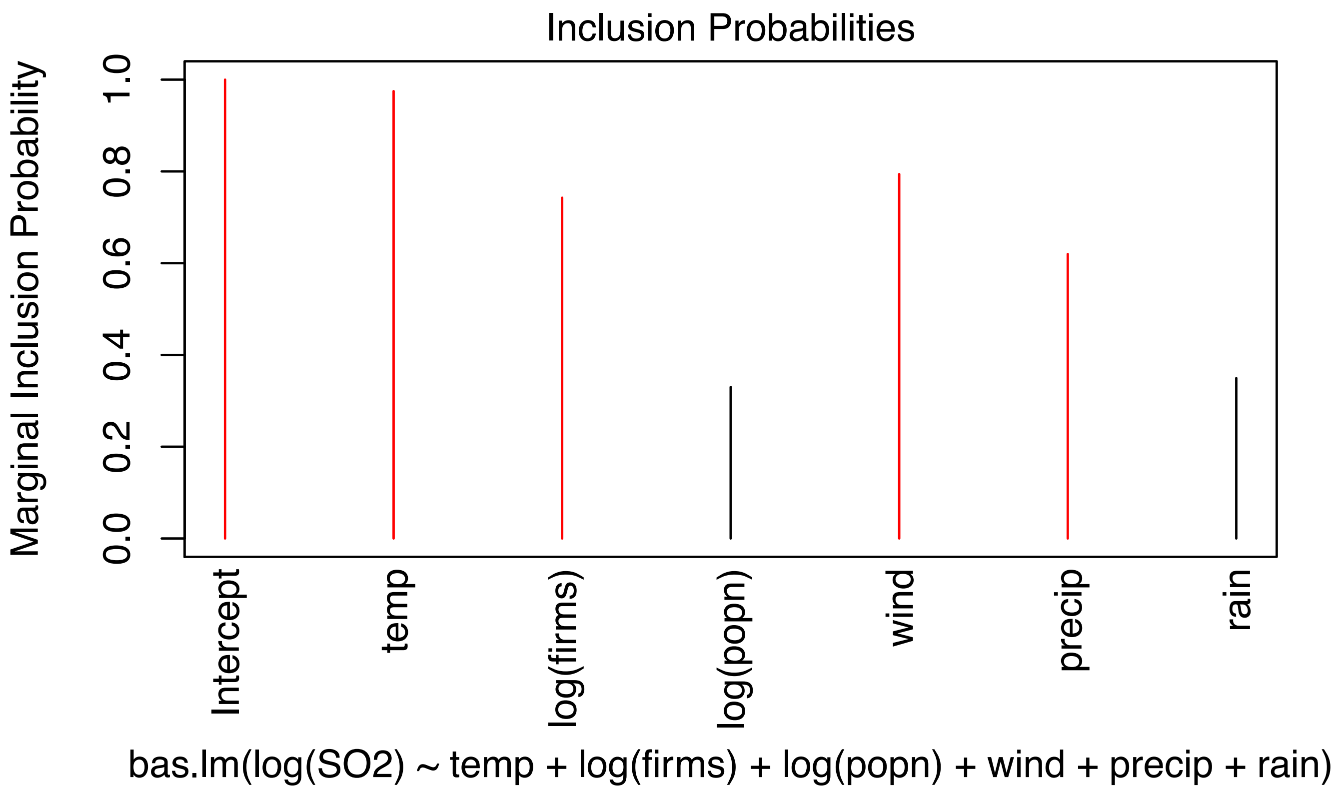
- "hyper-g/n"
- robust prior (Bayarri et al Annals of Statistics 2012)

Example

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

## Warning in model.matrix.default(mt, mf,
contrasts): non-list contrasts argument ignored
```

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

post μ | data

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

with expectation expressed as a weighted average

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

contrast with model selection

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 $\mathbf{Y}^*$

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

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

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

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

if your models are concentrated around $\mathcal{M}_t$

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

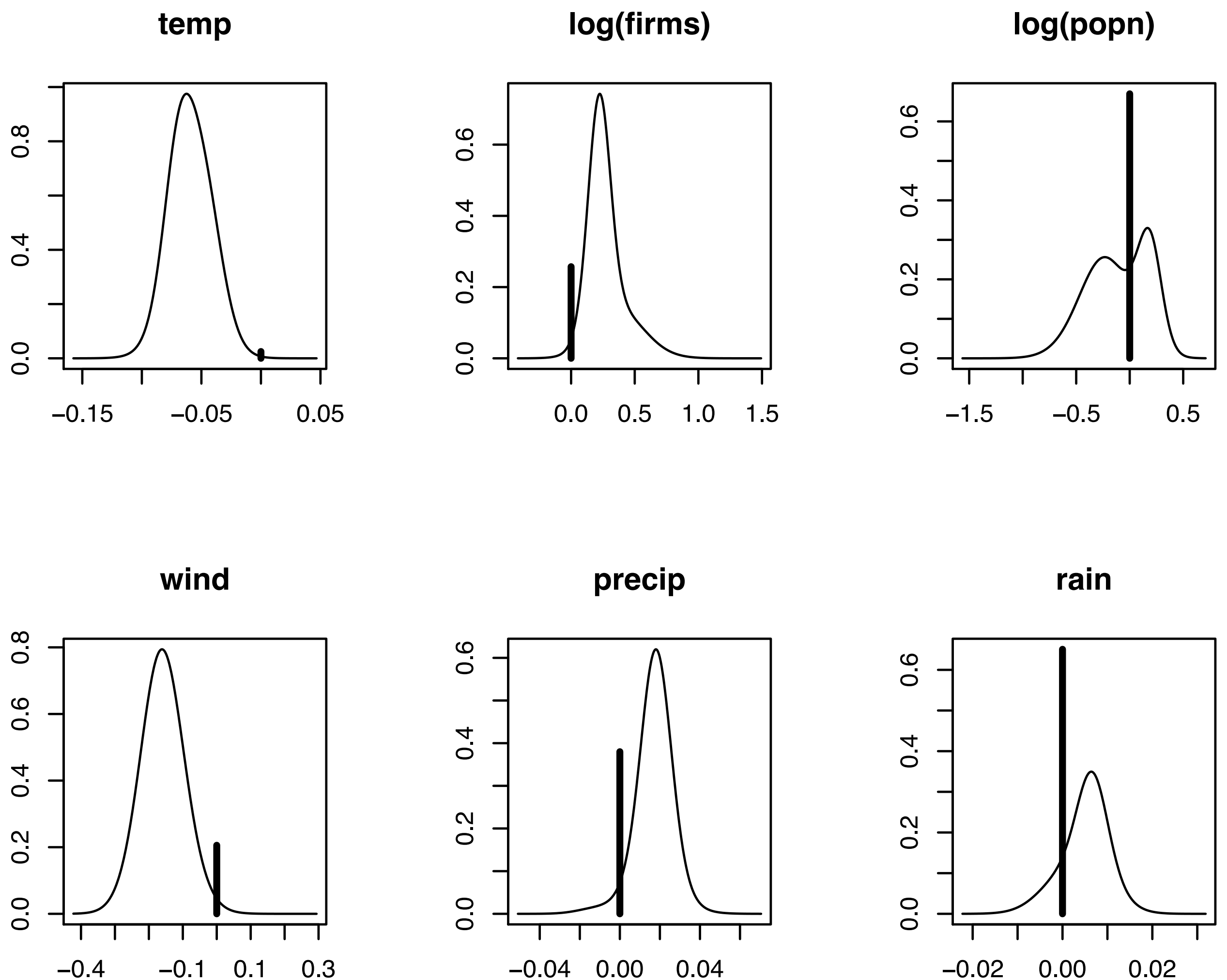
$$\mathsf{E}[\mu \mid \mathbf{Y}] = \mathbf{1}\hat{\alpha} + \mathbf{X} \sum \mathsf{E}[\beta \mid \mathbf{Y}, \mathcal{M}_\gamma] p(\mathcal{M}_\gamma \mid \mathbf{Y})$$

- 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 - when there is model uncertainty

Coefficients under BMA

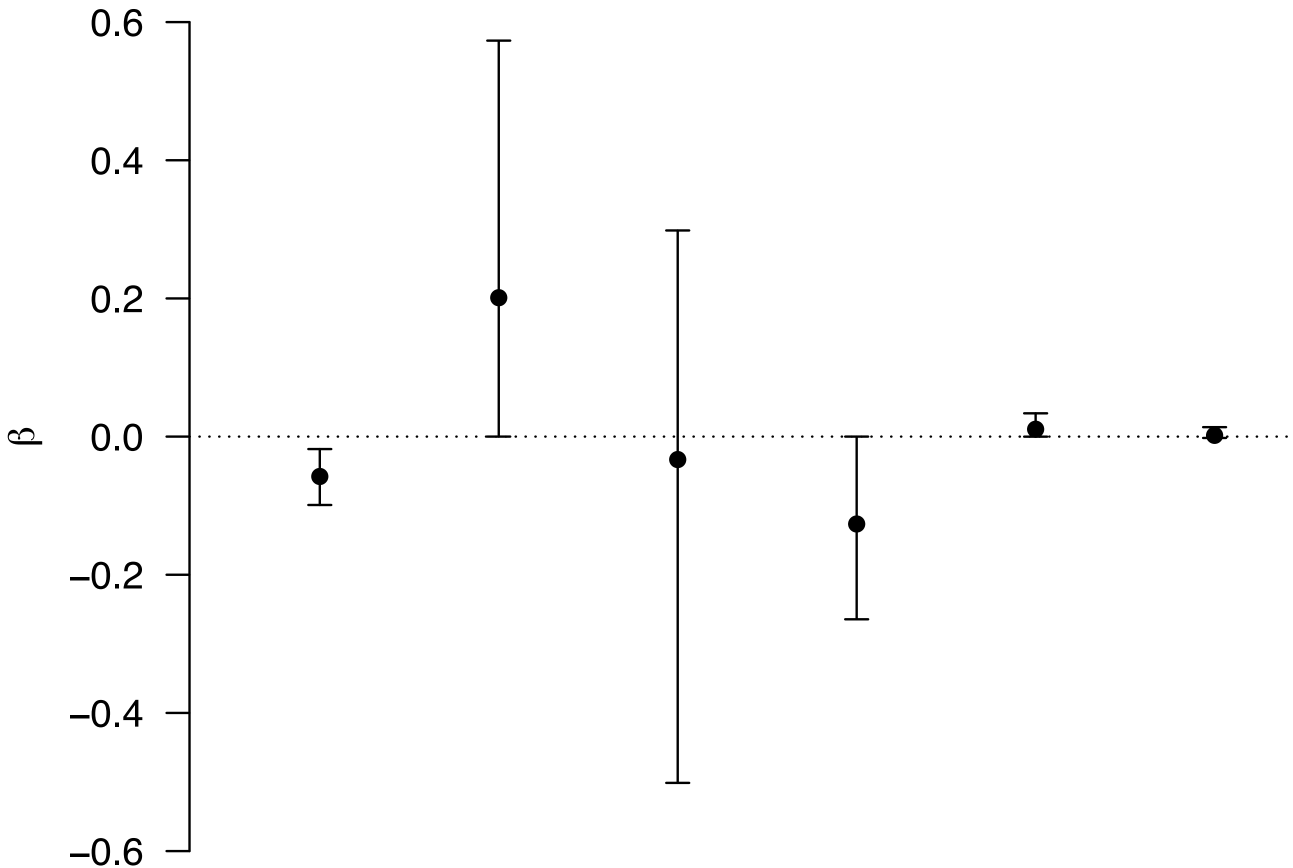
```
beta.ZS = coef(poll.ZS)
beta.ZS

##
##  Marginal Posterior Summaries of Coefficients:
##
##  Using  BMA
##
##  Based on the top  64 models
##               post mean  post SD    post p(B != 0)
## Intercept      3.153004   0.082496   1.000000
## temp          -0.057725   0.020401   0.974978
## log(firms)     0.201049   0.177190   0.742681
## log(popn)     -0.033245   0.172185   0.330113
## wind          -0.126515   0.086429   0.794158
## precip         0.010662   0.011308   0.620004
## rain           0.001780   0.004037   0.349521
```

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

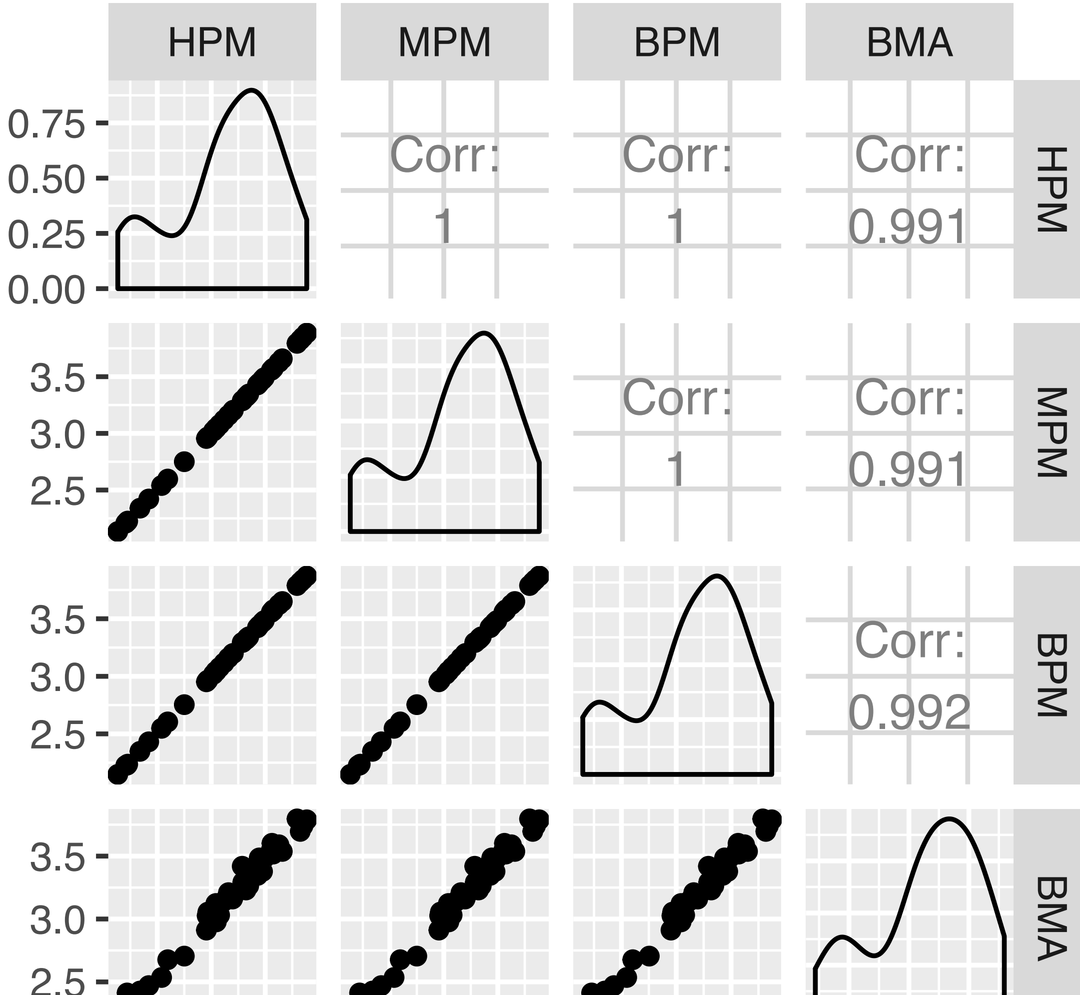
$$(\hat{\mu}_{BMA} - \hat{\mu}_{\mathcal{M}_\gamma})^T(\hat{\mu}_{BMA} - \hat{\mu}_{\mathcal{M}_\gamma})$$

- 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
```

```
## Warning in model.matrix.default(mt, mf,
contrasts): non-list contrasts argument ignored
```



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 but can use $g = n$