BMA

Hoff Chapter 9, Liang et al 2007, Hoeting et al (1999), Clyde & George (2004) Statistical Science

October 29, 2015

Prior Distributions

- "Spike and Slab" - Lempers (1971) Mitchell & Beauchamp (1988)

Prior Distributions

- “Spike and Slab” - Lempers (1971) Mitchell & Beauchamp (1988)
- “Spike and Bell” Leamer (1978) in BMA

Prior Distributions

- "Spike and Slab" - Lempers (1971) Mitchell & Beauchamp (1988)
- "Spike and Bell" Leamer (1978) in BMA
- mixture of 2 normals - concentrated and dispersed - SSVS Gibbs Sampler - George & McCulloch (1993)

Prior Distributions

- "Spike and Slab" - Lempers (1971) Mitchell & Beauchamp (1988)
- "Spike and Bell" Leamer (1978) in BMA
- mixture of 2 normals - concentrated and dispersed - SSVS Gibbs Sampler - George & McCulloch (1993)
- Back to "Spike and Bell" Hoeting, Raftery & Madigan MC^3 (1997) and George (1997) collapsed MCMC integrate out β_γ
- Conjugate Normal-Gammas lead to closed form expressions for marginal likelihoods, Zellner's g-prior is the most popular.

Prior Distributions

- "Spike and Slab" - Lempers (1971) Mitchell & Beauchamp (1988)
- "Spike and Bell" Leamer (1978) in BMA
- mixture of 2 normals - concentrated and dispersed - SSVS Gibbs Sampler - George & McCulloch (1993)
- Back to "Spike and Bell" Hoeting, Raftery & Madigan MC^3 (1997) and George (1997) collapsed MCMC integrate out β_γ
- Conjugate Normal-Gammas lead to closed form expressions for marginal likelihoods, Zellner's g-prior is the most popular.

Zellner's g-prior

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

Zellner's g-prior

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

- $p(\alpha) \propto 1$

Zellner's g-prior

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

- $p(\alpha) \propto 1$
- $p(\sigma^2) \propto 1/\sigma^2$

Zellner's g-prior

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

- $p(\alpha) \propto 1$
- $p(\sigma^2) \propto 1/\sigma^2$
- $\boldsymbol{\beta}_\gamma \mid \alpha, \sigma^2, \boldsymbol{\gamma} \sim \mathrm{N}(0, g\sigma^2(\mathbf{X}_\gamma^{c\prime}\mathbf{X}_\gamma^c)^{-1})$

Zellner's g-prior

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

- $p(\alpha) \propto 1$
- $p(\sigma^2) \propto 1/\sigma^2$
- $\boldsymbol{\beta}_\gamma \mid \alpha, \sigma^2, \boldsymbol{\gamma} \sim \mathsf{N}(0, g\sigma^2 ({\mathbf{X}^c_\gamma}' \mathbf{X}^c_\gamma)^{-1})$

which leads to marginal likelihood of $\mathcal{M}_\gamma$ that 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 is the usual R^2 for model $\mathcal{M}_\gamma$.

Zellner's g-prior

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

- $p(\alpha) \propto 1$
- $p(\sigma^2) \propto 1/\sigma^2$
- $\boldsymbol{\beta}_\gamma \mid \alpha, \sigma^2, \boldsymbol{\gamma} \sim \mathrm{N}(0, g\sigma^2(\mathbf{X}^{c\,\prime}_\gamma \mathbf{X}^c_\gamma)^{-1})$

which leads to marginal likelihood of $\mathcal{M}_\gamma$ that 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 is the usual R^2 for model $\mathcal{M}_\gamma$.
Trade-off of model complexity versus goodness of fit

Lastly, assign distribution to space of models

Priors on Model Space

$p(\mathcal{M}_\gamma) \Leftrightarrow p(\boldsymbol{\gamma})$

- $p(\gamma_j = 1) = .5 \Rightarrow P(\mathcal{M}_\gamma) = .5^p$ Uniform on space of models

Priors on Model Space

$p(\mathcal{M}_\gamma) \Leftrightarrow p(\boldsymbol{\gamma})$

- $p(\gamma_j = 1) = .5 \Rightarrow P(\mathcal{M}_\gamma) = .5^p$ Uniform on space of models
  $p_{\boldsymbol{\gamma}} \sim \text{Bin}(p, .5)$
- $\gamma_j \mid \pi \overset{\text{iid}}{\sim} \text{Ber}(\pi)$ and $\pi \sim \text{Beta}(a, b)$ then $p_{\boldsymbol{\gamma}} \sim \text{BB}(p, a, b)$

$$p(p_{\boldsymbol{\gamma}} \mid p, a, b) = \frac{\Gamma(p+1)\Gamma(p_{\boldsymbol{\gamma}} + a)\Gamma(p - p_{\boldsymbol{\gamma}} + b)\Gamma(a+b)}{\Gamma(p_{\boldsymbol{\gamma}} + 1)\Gamma(p - p_{\boldsymbol{\gamma}} + 1)\Gamma(p + a + b)\Gamma(a)\Gamma(b)}$$

- $p_{\boldsymbol{\gamma}} \sim \text{BB}(p, 1, 1) \sim \text{Unif}(0, p)$
- For sparsity. want b to be a function of p typically.

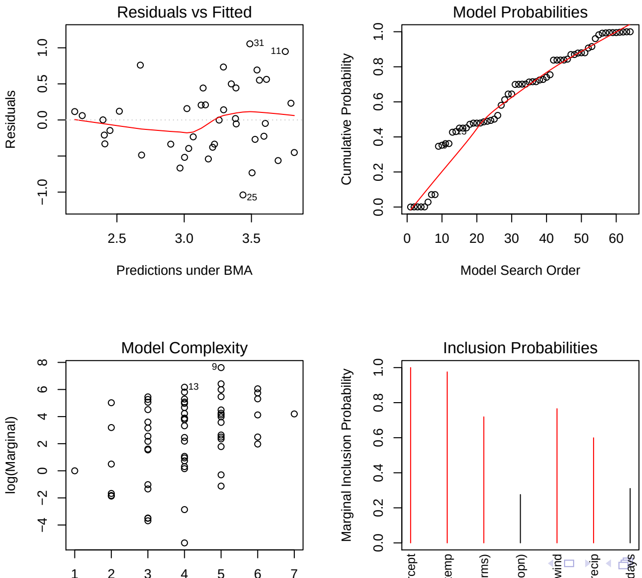
USair Data

```
library(BAS)
poll.bma = bas.lm(log(SO2) ~ temp + log(mgffirms) +
                             log(popn) + wind +
                             precip+ raindays,
                  data=pollution,
                  prior="g-prior",
                  alpha=41, # n
                  n.models=2^6,
                  modelprior = uniform(),
                  initprobs="Uniform")
> poll.bma

 Marginal Posterior Inclusion Probabilities:
 Intercept   temp   log(mfgfirms)  log(popn)   wind  precip
   1.0000   0.9755        0.7190      0.2757  0.7654  0.5994
```

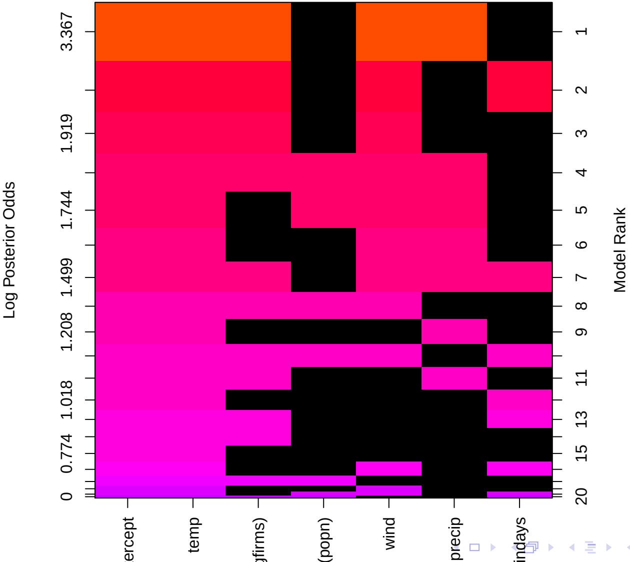
Plots

plot(poll.bma, ask=F)



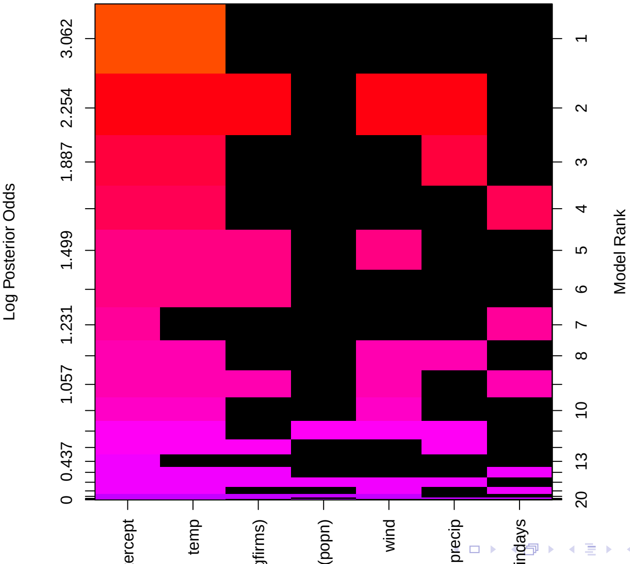
Posterior Distribution with Uniform Prior on Model Space

image(poll.bma)



Posterior Distribution with BB(1,p) Prior on Model Space

image(poll-bb.bma)



Jeffreys Scale of Evidence

- Bayes Factor = ratio of marginal likelihoods

Jeffreys Scale of Evidence

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

Jeffreys Scale of Evidence

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

$B = BF[\mathcal{M}_0 : \mathcal{M}_\gamma]$ and $1/B = BF[\mathcal{M}_\gamma : \mathcal{M}_0]$

Jeffreys Scale of Evidence

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

$B = BF[\mathcal{M}_0 : \mathcal{M}_\gamma]$ and $1/B = BF[\mathcal{M}_\gamma : \mathcal{M}_0]$

Bayes Factor	Interpretation
$B \geq 1$	H_0 supported
$1 > B \geq 10^{-\frac{1}{2}}$	minimal evidence against H_0
$10^{-\frac{1}{2}} > B \geq 10^{-1}$	substantial evidence against H_0
$10^{-1} > B \geq 10^{-2}$	strong evidence against H_0
$10^{-2} > B$	decisive evidence against H_0

Jeffreys Scale of Evidence

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

$B = BF[\mathcal{M}_0 : \mathcal{M}_\gamma]$ and $1/B = BF[\mathcal{M}_\gamma : \mathcal{M}_0]$

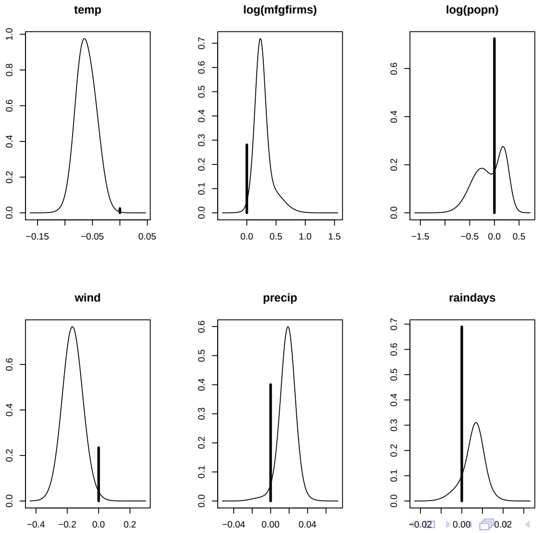
Bayes Factor	Interpretation
$B \geq 1$	H_0 supported
$1 > B \geq 10^{-\frac{1}{2}}$	minimal evidence against H_0
$10^{-\frac{1}{2}} > B \geq 10^{-1}$	substantial evidence against H_0
$10^{-1} > B \geq 10^{-2}$	strong evidence against H_0
$10^{-2} > B$	decisive evidence against H_0

in context

of testing one hypothesis with equal prior odds

Coefficients

```
beta = coef(poll.bma)
par(mfrow=c(2,3));  plot(beta, subset=2:7,ask=F)
```



Information Paradox

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

$$BF(\mathcal{M}_\gamma : \boldsymbol{\mathcal{M}}_0) = (1+g)^{(n-1-p_\gamma)/2}(1+g(1-R^2))^{-(n-1)/2}$$

Information Paradox

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

$$BF(\mathcal{M}_\gamma : \boldsymbol{\mathcal{M}}_0) = (1+g)^{(n-1-p_\gamma)/2}(1+g(1-R^2))^{-(n-1)/2}$$

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

Information Paradox

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

$$BF(\mathcal{M}_\gamma : \boldsymbol{\mathcal{M}}_0) = (1+g)^{(n-1-p_\gamma)/2}(1+g(1-R^2))^{-(n-1)/2}$$

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

Information Paradox

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

$$BF(\mathcal{M}_\gamma : \mathfrak{M}_0) = (1+g)^{(n-1-p_\gamma)/2}(1+g(1-R^2))^{-(n-1)/2}$$

- 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 $\mathfrak{M}_0$ where F is the usual F statistic for comparing model $\mathcal{M}_\gamma$ to $\mathfrak{M}_0$

Information Paradox

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

$$BF(\mathcal{M}_\gamma : \mathbf{\mathcal{M}}_0) = (1+g)^{(n-1-p_\gamma)/2}(1+g(1-R^2))^{-(n-1)/2}$$

- 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 $\mathbf{\mathcal{M}}_0$ where F is the usual F statistic for comparing model $\mathcal{M}_\gamma$ to $\mathbf{\mathcal{M}}_0$
- BF converges to a fixed constant $(1+g)^{-p_\gamma/2}$ (does not go to infinity

"Information Inconsistency" see Liang et al JASA 2008

Mixtures of g priors & Information consistency

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

Mixtures of g priors & Information consistency

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

- Zellner-Siow Cauchy prior

Mixtures of g priors & Information consistency

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

- Zellner-Siow Cauchy prior
- hyper-g prior or hyper-g/n (Liang et al JASA 2008)

Mixtures of g priors & Information consistency

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

- Zellner-Siow Cauchy prior
- hyper-g prior or hyper-g/n (Liang et al JASA 2008)
- robust prior (Bayarrri et al Annals of Statistics 2012

Mixtures of g priors & Information consistency

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

- Zellner-Siow Cauchy prior
- hyper-g prior or hyper-g/n (Liang et al JASA 2008)
- robust prior (Bayarrri et al Annals of Statistics 2012

All have tails that behave like a Cauchy distribution

Mortality & Pollution

- Data from Statistical Sleuth 12.17

Mortality & Pollution

- Data from Statistical Sleuth 12.17
- 60 cities

Mortality & Pollution

- Data from Statistical Sleuth 12.17
- 60 cities
- response Mortality

Mortality & Pollution

- Data from Statistical Sleuth 12.17
- 60 cities
- response Mortality
- measures of HC, NOX, SO2

Mortality & Pollution

- Data from Statistical Sleuth 12.17
- 60 cities
- response Mortality
- measures of HC, NOX, SO2
- Is pollution associated with mortality after adjusting for other socio-economic and meteorological factors?

Mortality & Pollution

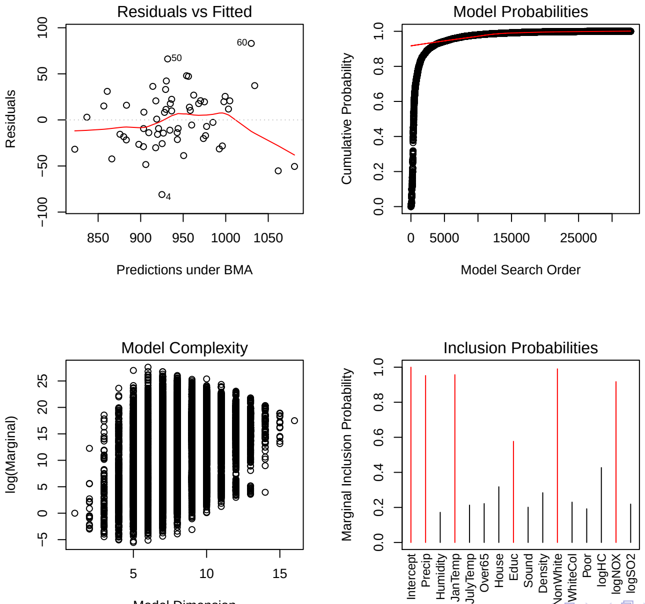
- Data from Statistical Sleuth 12.17
- 60 cities
- response Mortality
- measures of HC, NOX, SO2
- Is pollution associated with mortality after adjusting for other socio-economic and meteorological factors?
- 15 predictor variables implies $2^{15} = 32,768$ possible models

Mortality & Pollution

- Data from Statistical Sleuth 12.17
- 60 cities
- response Mortality
- measures of HC, NOX, SO2
- Is pollution associated with mortality after adjusting for other socio-economic and meteorological factors?
- 15 predictor variables implies $2^{15} = 32,768$ possible models
- Use Zellner-Siow Cauchy prior $1/g \sim G(1/2, n/2)$

```
mort.bma = bas.lm(MORTALITY ~ ., data=mortality,
                  prior="ZS-null",
                  alpha=60,  n.models=2^15,
                  update=100, initprobs="eplogp")
```

Posterior Distributions



Posterior Probabilities

- What is the probability that there is no pollution effect?

Posterior Probabilities

- What is the probability that there is no pollution effect?
- Sum posterior model probabilities over all models that include no pollution variables

Posterior Probabilities

- What is the probability that there is no pollution effect?
- Sum posterior model probabilities over all models that include no pollution variables

```
> which.mat = list2matrix.which(mort.bma,1:(2^15))
> poll.in = (which.mat[, 14:16] %*% rep(1, 3)) > 0
> sum(poll.in * mort.bma$postprob)
[1] 0.9889641
```

Posterior Probabilities

- What is the probability that there is no pollution effect?
- Sum posterior model probabilities over all models that include no pollution variables

```
> which.mat = list2matrix.which(mort.bma,1:(2^15))
> poll.in = (which.mat[, 14:16] %*% rep(1, 3)) > 0
> sum(poll.in * mort.bma$postprob)
[1] 0.9889641
```

- Posterior probability is 0.011

Posterior Probabilities

- What is the probability that there is no pollution effect?
- Sum posterior model probabilities over all models that include no pollution variables

```
> which.mat = list2matrix.which(mort.bma,1:(2^15))
> poll.in = (which.mat[, 14:16] %*% rep(1, 3)) > 0
> sum(poll.in * mort.bma$postprob)
[1] 0.9889641
```

- Posterior probability is 0.011
- Odds that there is an effect $(1 - .011)/(.011) = 89$.

Posterior Probabilities

- What is the probability that there is no pollution effect?
- Sum posterior model probabilities over all models that include no pollution variables

```
> which.mat = list2matrix.which(mort.bma,1:(2^15))
> poll.in = (which.mat[, 14:16] %*% rep(1, 3)) > 0
> sum(poll.in * mort.bma$postprob)
[1] 0.9889641
```

- Posterior probability is 0.011
- Odds that there is an effect $(1 - .011)/(.011) = 89$.
- Prior Odds $7 = (1 - .5^3)/.5^3$

Posterior Probabilities

- What is the probability that there is no pollution effect?
- Sum posterior model probabilities over all models that include no pollution variables

```
> which.mat = list2matrix.which(mort.bma,1:(2^15))
> poll.in = (which.mat[, 14:16] %*% rep(1, 3)) > 0
> sum(poll.in * mort.bma$postprob)
[1] 0.9889641
```

- Posterior probability is 0.011
- Odds that there is an effect $(1 - .011)/(.011) = 89.$
- Prior Odds $7 = (1 - .5^3)/.5^3$
- Bayes Factor for a pollution effect $89.9/7 = 12.8$

Posterior Probabilities

- What is the probability that there is no pollution effect?
- Sum posterior model probabilities over all models that include no pollution variables

```
> which.mat = list2matrix.which(mort.bma,1:(2^15))
> poll.in = (which.mat[, 14:16] %*% rep(1, 3)) > 0
> sum(poll.in * mort.bma$postprob)
[1] 0.9889641
```

- Posterior probability is 0.011
- Odds that there is an effect $(1 - .011)/(.011) = 89.$
- Prior Odds $7 = (1 - .5^3)/.5^3$
- Bayes Factor for a pollution effect $89.9/7 = 12.8$
- Bayes Factor for NOX based on marginal inclusion probability $0.917/(1 - 0.917) = 11.0$

Posterior Probabilities

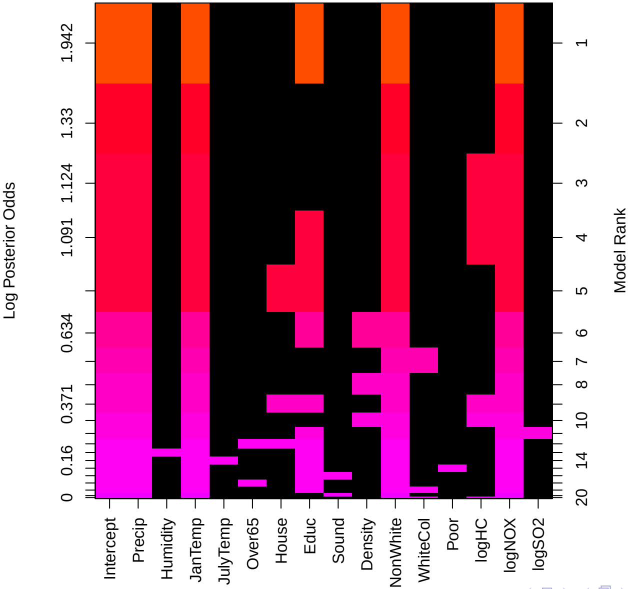
- What is the probability that there is no pollution effect?
- Sum posterior model probabilities over all models that include no pollution variables

```
> which.mat = list2matrix.which(mort.bma,1:(2^15))
> poll.in = (which.mat[, 14:16] %*% rep(1, 3)) > 0
> sum(poll.in * mort.bma$postprob)
[1] 0.9889641
```

- Posterior probability is 0.011
- Odds that there is an effect $(1 - .011)/(.011) = 89.$
- Prior Odds $7 = (1 - .5^3)/.5^3$
- Bayes Factor for a pollution effect $89.9/7 = 12.8$
- Bayes Factor for NOX based on marginal inclusion probability $0.917/(1 - 0.917) = 11.0$
- Marginal inclusion probability for logHC $= 0.427144$
- Marginal inclusion probability for logSO2 $= 0.218978$

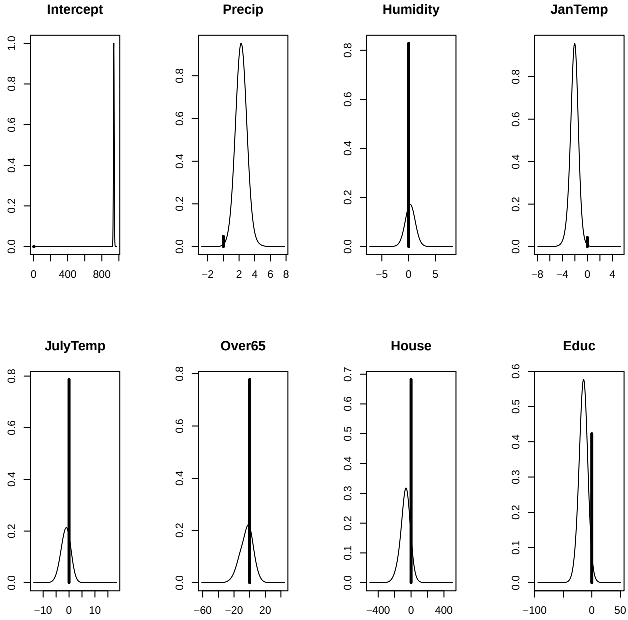
Bayes Factors are not additive! Better to work with probabilities.

Model Space

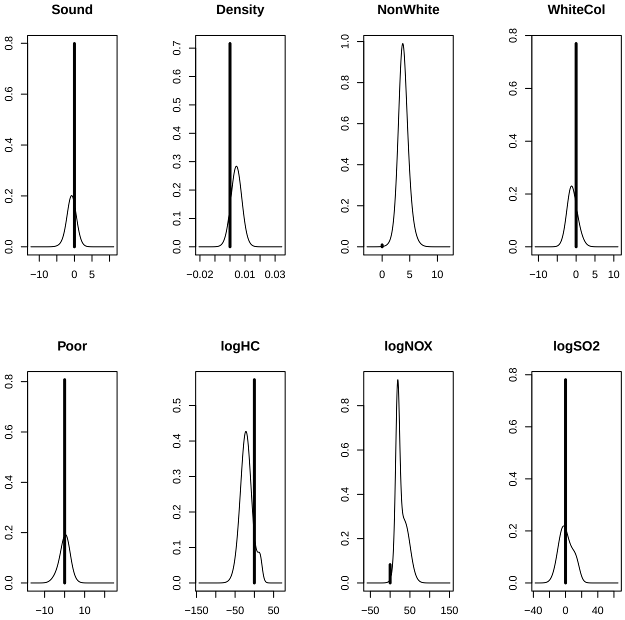


Log Posterior Odds
1.942
1.33
1.124
1.091
0.634
0.371
0.16
0
Model Rank
1
2
3
4
5
6
7
8
10
14
20
Intercept
Precip
Humidity
JanTemp
JulyTemp
Over65
House
Educ
Sound
Density
NonWhite
WhiteCol
Poor
logHC
logNOX
logSO2

Coefficients



Coefficients



Effect Estimation

- Coefficients in each model are adjusted for other variables in the model
- OLS: leave out a predictor with a non-zero coefficient then estimates are biased!
- Model Selection in the presence of high correlation, may leave out "redundant" variables;
- improved MSE for prediction (Bias-variance tradeoff)
- Bayes is biased anyway so should we care?
- What is meaning of $\sum_{\gamma} \beta_{j\gamma} \gamma_j P(\mathcal{M}_\gamma \mid \mathbf{Y})$

Problem with confounding! Need to change prior?

Other Problems

- Computational

Other Problems

- Computational if $p > 35$ enumeration is difficult

Other Problems

- Computational if $p > 35$ enumeration is difficult
  - Gibbs sampler or Random-Walk algorithm on γ

Other Problems

- Computational if $p > 35$ enumeration is difficult
  - Gibbs sampler or Random-Walk algorithm on γ
  - poor convergence/mixing with high correlations

Other Problems

- Computational if $p > 35$ enumeration is difficult
  - Gibbs sampler or Random-Walk algorithm on γ
  - poor convergence/mixing with high correlations
  - Metropolis Hastings algorithms more flexibility

Other Problems

- Computational if $p > 35$ enumeration is difficult
  - Gibbs sampler or Random-Walk algorithm on γ
  - poor convergence/mixing with high correlations
  - Metropolis Hastings algorithms more flexibility
  - "Stochastic Search" (no guarantee samples represent posterior)

Other Problems

- Computational if $p > 35$ enumeration is difficult
  - Gibbs sampler or Random-Walk algorithm on γ
  - poor convergence/mixing with high correlations
  - Metropolis Hastings algorithms more flexibility
  - "Stochastic Search" (no guarantee samples represent posterior)
  - in BMA all variables are included, but coefficients are shrunk to 0; alternative is to use Shrinkage methods

Other Problems

- Computational if $p > 35$ enumeration is difficult
  - Gibbs sampler or Random-Walk algorithm on γ
  - poor convergence/mixing with high correlations
  - Metropolis Hastings algorithms more flexibility
  - "Stochastic Search" (no guarantee samples represent posterior)
  - in BMA all variables are included, but coefficients are shrunk to 0; alternative is to use Shrinkage methods
- Prior Choice: Choice of prior distributions on β and on γ

Other Problems

- Computational if $p > 35$ enumeration is difficult
  - Gibbs sampler or Random-Walk algorithm on γ
  - poor convergence/mixing with high correlations
  - Metropolis Hastings algorithms more flexibility
  - "Stochastic Search" (no guarantee samples represent posterior)
  - in BMA all variables are included, but coefficients are shrunk to 0; alternative is to use Shrinkage methods
- Prior Choice: Choice of prior distributions on β and on γ

Model averaging versus Model Selection – what are objectives?