# BMA & Distributions

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

October 25, 2021

# USair Data

```r
library(BAS)
data(usair, package="HH")
poll.bma = bas.lm(log(SO2) ~ temp + log(mfgfirms) +
                             log(popn) + wind +
                             precip + raindays,
                  data=usair,
                  prior="g-prior",
                  alpha=nrow(usair), # g = n
                  n.models=2^6,
                  modelprior = uniform(),
                  method="deterministic")
```

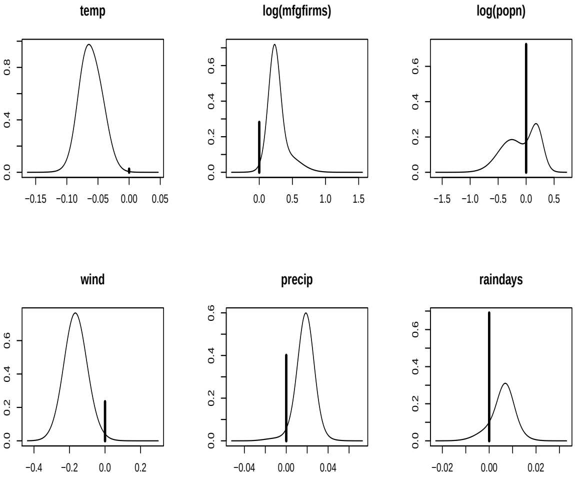
# Summary

```r
summary(poll.bma)

##                  P(B != 0 | Y)  model 1   model 2   model 3    mo
## Intercept            1.0000000 1.000000 1.0000000 1.0000000 1.00
## temp                 0.9755041 1.000000 1.0000000 1.0000000 1.00
## log(mfgfirms)        0.7190313 1.000000 1.0000000 1.0000000 1.00
## log(popn)            0.2756811 0.000000 0.0000000 0.0000000 1.00
## wind                 0.7654485 1.000000 1.0000000 1.0000000 1.00
## precip               0.5993801 1.000000 0.0000000 0.0000000 1.00
## raindays             0.3103574 0.000000 1.0000000 0.0000000 0.00
## BF                          NA 1.000000 0.3022674 0.2349056 0.20
## PostProbs                   NA 0.275800 0.0834000 0.0648000 0.05
## R2                          NA 0.542700 0.5130000 0.4558000 0.55
## dim                         NA 5.000000 5.0000000 4.0000000 6.00
## logmarg                     NA 7.616228 6.4197847 6.1676565 6.05
```

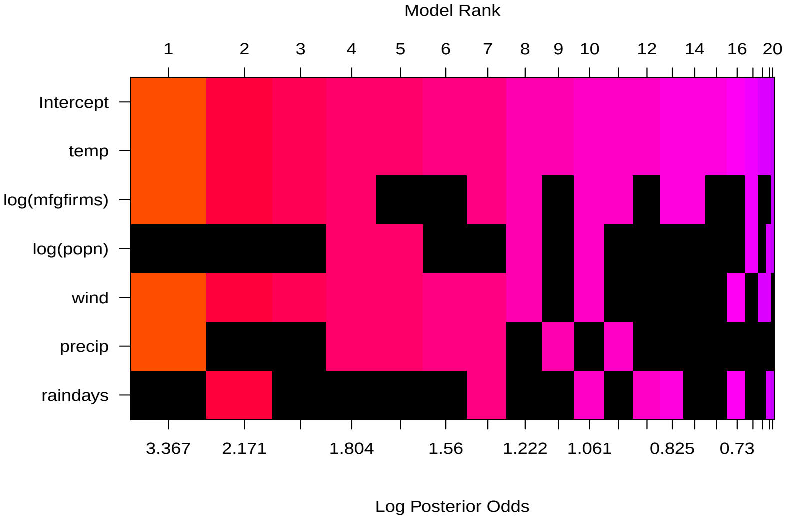
# Plots

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



# Posterior Distribution with Uniform Prior on Model Space

```r
image(poll.bma, rotate=FALSE)
```

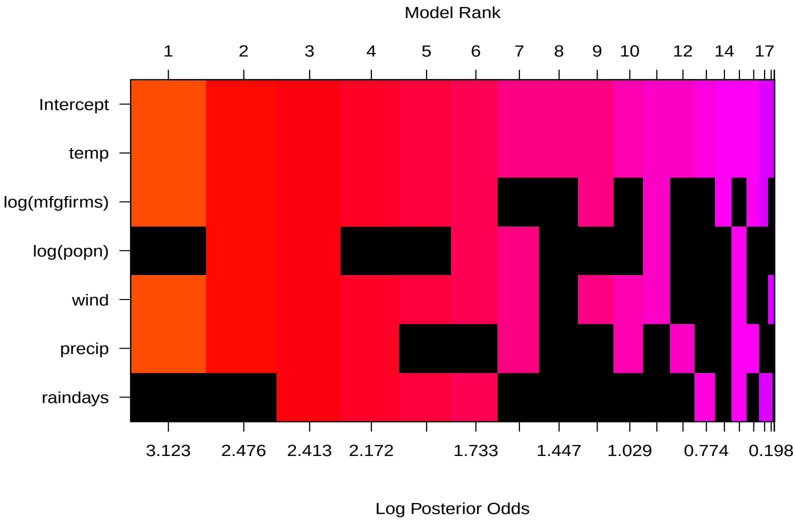


# Posterior Distribution with BB(1,1) Prior on Model Space

```
poll.bb.bma = bas.lm(log(SO2) ~ temp + log(mfgfirms) +
                                log(popn) + wind +
                                precip + raindays,
                     data=usair,
                     prior="g-prior",
                     alpha=nrow(usair),
                     n.models=2^6,  #enumerate
                     modelprior=beta.binomial(1,1))
```

# BB(1,1) Prior on Model Space

```r
image(poll.bb.bma, rotate=FALSE)
```



# Bartlett's Paradox

The Bayes factor for comparing $\gamma$ to the null model:

$$BF(\gamma : \gamma 0) = (1+g)^{(n-1-p_\gamma)/2}(1+g(1-R_\gamma^2))^{-(n-1)/2}$$

# Bartlett's Paradox

The Bayes factor for comparing $\gamma$ to the null model:

$$BF(\gamma : \gamma 0) = (1+g)^{(n-1-p_\gamma)/2}(1+g(1-R^2_\gamma))^{-(n-1)/2}$$

- For fixed sample size $n$ and $R^2_\gamma$, consider taking values of $g$ that go to infinity

# Bartlett's Paradox

The Bayes factor for comparing $\gamma$ to the null model:

$$BF(\gamma : \gamma 0) = (1 + g)^{(n-1-p_\gamma)/2}(1 + g(1 - R^2_\gamma))^{-(n-1)/2}$$

- For fixed sample size $n$ and $R^2_\gamma$, consider taking values of $g$ that go to infinity
- Increasing vagueness in prior

# Bartlett's Paradox

The Bayes factor for comparing $\gamma$ to the null model:

$$BF(\gamma : \gamma 0) = (1 + g)^{(n-1-p_\gamma)/2}(1 + g(1 - R^2_\gamma))^{-(n-1)/2}$$

- For fixed sample size $n$ and $R^2_\gamma$, consider taking values of $g$ that go to infinity
- Increasing vagueness in prior
- What happens to BF as $g \to \infty$?

# Bartlett's Paradox

The Bayes factor for comparing $\gamma$ to the null model:

$$BF(\gamma : \gamma 0) = (1 + g)^{(n-1-p_\gamma)/2}(1 + g(1 - R^2_\gamma))^{-(n-1)/2}$$

- For fixed sample size $n$ and $R^2_\gamma$, consider taking values of $g$ that go to infinity
- Increasing vagueness in prior
- What happens to BF as $g \to \infty$?
- why is this a paradox?

# Information Paradox

The Bayes factor for comparing $\gamma$ to the null model:

$$BF(\gamma : \gamma_0) = (1+g)^{(n-1-p_\gamma)/2}(1+g(1-R^2_\gamma))^{-(n-1)/2}$$

# Information Paradox

The Bayes factor for comparing $\gamma$ to the null model:

$$BF(\gamma : \gamma_0) = (1+g)^{(n-1-p_\gamma)/2}(1+g(1-R_\gamma^2))^{-(n-1)/2}$$

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

# Information Paradox

The Bayes factor for comparing $\gamma$ to the null model:

$$BF(\gamma : \gamma_0) = (1+g)^{(n-1-p_\gamma)/2}(1+g(1-R^2_\gamma))^{-(n-1)/2}$$

- 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 $\gamma$ to the null model:

$$BF(\gamma : \gamma_0) = (1+g)^{(n-1-p_\gamma)/2}(1+g(1-R^2_\gamma))^{-(n-1)/2}$$

- 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 $\gamma_0$ where $F$ is the usual $F$ statistic for comparing model $\gamma$ to $\gamma_0$

# Information Paradox

The Bayes factor for comparing $\gamma$ to the null model:

$$BF(\gamma : \gamma_0) = (1+g)^{(n-1-p_\gamma)/2}(1+g(1-R^2_\gamma))^{-(n-1)/2}$$

- 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 $\gamma_0$ where $F$ is the usual $F$ statistic for comparing model $\gamma$ to $\gamma_0$
- BF converges to a fixed constant $(1+g)^{n-1-p_\gamma/2}$ (does not go to infinity

"Information Inconsistency" see Liang et al JASA 2008

# Mixtures of $g$ priors & Information consistency

- Need $BF \to \infty$ if $\mathrm{R}^2_\gamma \to 1$
- Put a prior on $g$

$$BF(\gamma : \gamma_0) = \frac{C \int (1+g)^{(n-1-p_\gamma)/2}(1+g(1-R^2_\gamma))^{-(n-1)/2}\pi(g)dg}{C}$$

- interchange limit and integration as $R^2 \to 1$ want

$$\mathrm{E}_g[(1+g)^{(n-1-p_\gamma)/2}]$$

  to diverge
- hyper-g prior (Liang et al JASA 2008)

$$p(g) = \frac{a-2}{2}(1+g)^{-a/2}$$

  or $g/(1+g) \sim Beta(1, (a-2)/2)$
- prior expectation converges if $a > n + 1 - p_\gamma$
- Consider minimal model $p_\gamma = 1$ and $n = 3$ (can estimate intercept, one coefficient, and $\sigma^2$, then $a > 3$ integral exists
- For $2 < a \leq 3$ integral diverges and resolves the information paradox!

# Mixtures of $g$ priors & Information consistency

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

## Mixtures of $g$ priors & Information consistency

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

- hyper-g prior (Liang et al JASA 2008)

$$p(g) = \frac{a-2}{2}(1+g)^{-a/2}$$

or $g/(1+g) \sim Beta(1, (a-2)/2)$ need $2 < a \leq 3$

# Mixtures of $g$ priors & Information consistency

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

- hyper-g prior (Liang et al JASA 2008)

$$p(g) = \frac{a-2}{2}(1+g)^{-a/2}$$

  or $g/(1+g) \sim Beta(1, (a-2)/2)$ need $2 < a \leq 3$

- Jeffreys prior on $g$ corresponds to $a = 2$ (improper)

# Mixtures of $g$ priors & Information consistency

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

- hyper-g prior (Liang et al JASA 2008)

$$p(g) = \frac{a-2}{2}(1+g)^{-a/2}$$

  or $g/(1+g) \sim Beta(1, (a-2)/2)$ need $2 < a \leq 3$
- Jeffreys prior on $g$ corresponds to $a = 2$ (improper)
- Hyper-g/n $(g/n)(1+g/n) \sim (Beta(1, (a-2)/2)$

# Mixtures of $g$ priors & Information consistency

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

- hyper-g prior (Liang et al JASA 2008)

$$p(g) = \frac{a-2}{2}(1+g)^{-a/2}$$

  or $g/(1+g) \sim Beta(1,(a-2)/2)$ need $2 < a \leq 3$

- Jeffreys prior on $g$ corresponds to $a = 2$ (improper)
- Hyper-g/n $(g/n)(1+g/n) \sim (Beta(1,(a-2)/2)$
- Zellner-Siow Cauchy prior $1/g \sim G(1/2, n/2)$

# Mixtures of $g$ priors & Information consistency

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

- hyper-g prior (Liang et al JASA 2008)

$$p(g) = \frac{a-2}{2}(1+g)^{-a/2}$$

  or $g/(1+g) \sim Beta(1, (a-2)/2)$ need $2 < a \leq 3$
- Jeffreys prior on $g$ corresponds to $a = 2$ (improper)
- Hyper-g/n $(g/n)(1+g/n) \sim (Beta(1, (a-2)/2)$
- Zellner-Siow Cauchy prior $1/g \sim G(1/2, n/2)$
- robust prior (Bayarri et al Annals of Statistics 2012

# Mixtures of $g$ priors & Information consistency

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

- hyper-g prior (Liang et al JASA 2008)

$$p(g) = \frac{a-2}{2}(1+g)^{-a/2}$$

  or $g/(1+g) \sim Beta(1,(a-2)/2)$ need $2 < a \leq 3$
- Jeffreys prior on $g$ corresponds to $a = 2$ (improper)
- Hyper-g/n $(g/n)(1+g/n) \sim (Beta(1,(a-2)/2)$
- Zellner-Siow Cauchy prior $1/g \sim G(1/2, n/2)$
- robust prior (Bayarri et al Annals of Statistics 2012
- Intrinsic prior (Womack et al JASA 2015)

All have prior tails for $\beta$ that behave like a Cauchy distribution and (the latter 4) marginal likelihoods that can be computed using special hypergeometric functions (${}_2F_1$, Appell $F_1$)

# Computation

If $p > 35$ enumeration is difficult

- Gibbs sampler or Random-Walk algorithm on $\gamma$

# Computation

If $p > 35$ enumeration is difficult

- Gibbs sampler or Random-Walk algorithm on $\gamma$
- slow convergence/poor mixing with high correlations

# Computation

If $p > 35$ enumeration is difficult

- Gibbs sampler or Random-Walk algorithm on $\gamma$
- slow convergence/poor mixing with high correlations
- Metropolis Hastings algorithms more flexibility

## Computation

If $p > 35$ enumeration is difficult

- Gibbs sampler or Random-Walk algorithm on $\gamma$
- slow convergence/poor mixing with high correlations
- Metropolis Hastings algorithms more flexibility (swap pairs of variables)

# Diabetes Example from Hoff $p = 64$

```
set.seed(8675309)
source("yX.diabetes.train.txt")
diabetes.train = as.data.frame(diabetes.train)

source("yX.diabetes.test.txt")
diabetes.test = as.data.frame(diabetes.test)
colnames(diabetes.test)[1] = "y"

str(diabetes.train)

## 'data.frame': 342 obs. of  65 variables:
##  $ y      : num  -0.0147 -1.0005 -0.1444 0.6987 -0.2222
##  $ age    : num  0.7996 -0.0395 1.7913 -1.8703 0.113 ...
##  $ sex    : num  1.064 -0.937 1.064 -0.937 -0.937 ...
##  $ bmi    : num  1.296 -1.081 0.933 -0.243 -0.764 ...
##  $ map    : num  0.459 -0.553 -0.119 -0.77 0.459 ...
##  $ tc     : num  -0.9287 -0.1774 -0.9576 0.256 0.0826 ..
##  $ ldl    : num  -0.731 -0.402 -0.718 0.525 0.328 ...
```

# MCMC with BAS

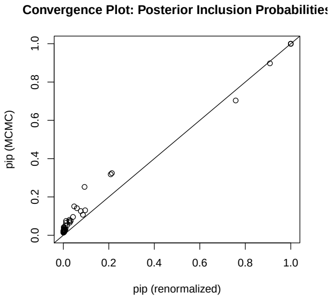
```
diabetes.bas = bas.lm(y ~ ., data=diabetes.train,
                      prior = "JZS",
                      method="MCMC",
                      n.models = 10000,
                      MCMC.iterations=150000,
                      thin = 10,
                      initprobs="eplogp",
                      force.heredity=FALSE)

system.time(bas.lm(y ~ ., data=diabetes.train,
                   prior = "JZS",
                   method="MCMC", n.models = 10000,
                   MCMC.iterations=150000,
                   thin = 10,  initprobs="eplogp",
                   force.heredity=FALSE))

##    user  system elapsed
##   6.881   0.288   7.173
```

# Diagnostics

```
diagnostics(diabetes.bas, type="pip")
```



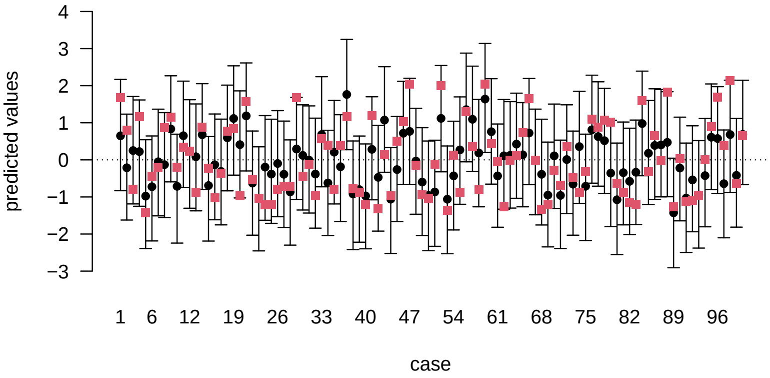
# Prediction

```
pred.bas = predict(diabetes.bas,
                   newdata=diabetes.test,
                   estimator="BMA",
                   se=TRUE)
mean((pred.bas$fit- diabetes.test$y)^2)

## [1] 0.4552798
```

# 95% prediction intervals

```r
ci.bas = confint(pred.bas); plot(ci.bas)
points(diabetes.test$y, col=2, pch=15)
```




coverage is 100

# Selection and Prediction

- BMA - optimal for squared error loss Bayes
- HPM: Highest Posterior Probability model (not optimal for prediction) but for selection
- MPM: Median Probabilty model (select model where PIP ¿ 0.5) (optimal under certain conditions; nested models)
- BPM: Best Probability Model - Model closest to BMA under loss (usually includes more predictors than HPM or MPM)

# Selection

```
pred.bas = predict(diabetes.bas,
                   newdata=diabetes.test,
                   estimator="BPM",
                   se=TRUE)
#MSE
mean((pred.bas$fit- diabetes.test$y)^2)

## [1] 0.4740667

#Coverage
ci.bas = confint(pred.bas)
mean(diabetes.test$y > ci.bas[,1] &
     diabetes.test$y < ci.bas[,2])

## [1] 0.98
```

# Alternatives to MCMC

- "Stochastic Search" (no guarantee samples represent posterior)

# Alternatives to MCMC

- "Stochastic Search" (no guarantee samples represent posterior)
- Variational, EM, etc to find modal model

# Alternatives to MCMC

- "Stochastic Search" (no guarantee samples represent posterior)
- Variational, EM, etc to find modal model
- in BMA all variables are included, but coefficients are shrunk to 0; alternative is to use shrinkage methods without point mass at zero

# Alternatives to MCMC

- "Stochastic Search" (no guarantee samples represent posterior)
- Variational, EM, etc to find modal model
- in BMA all variables are included, but coefficients are shrunk to 0; alternative is to use shrinkage methods without point mass at zero
- 
- If $p > n$, can use a generalized inverse, but requires care for prior on $\gamma$!

Model averaging versus Model Selection – what are objectives?

# 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)
- in BMA all variables are included, but coefficients are shrunk to 0
- Care needed for "causal" questions and confounder adjustment! With confounding, should not use plain BMA. Need to change prior to include potential confounders (advanced topic)