Bayesian model averaging

Dr. Jarad Niemi

STAT 544 - Iowa State University

March 9, 2017

Outline

- Bayesian model averaging
- BIC model averaging
- Model search
- Parameter averaging
- Posterior inclusion probability
- Model selection

Bayesian Model Averaging

The posterior predictive distribution

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta$$

assumes there is a true model $p(y|\theta)$ and accounts for the uncertainty in θ.

If you want to account for model uncertainty amongst some set of models $M_1, \ldots, M_h$, you can use the Bayesian model averaged posterior predictive distribution

$$p(\tilde{y}|y) = \sum_{h=1}^{H} p(\tilde{y}|M_h, y)p(M_h|y)$$

where

- $p(M_h|y)$ is the posterior model probability and
- $p(\tilde{y}|M_h, y)$ is the predictive distribution under model M_h.

Normal example

Suppose we have two models:

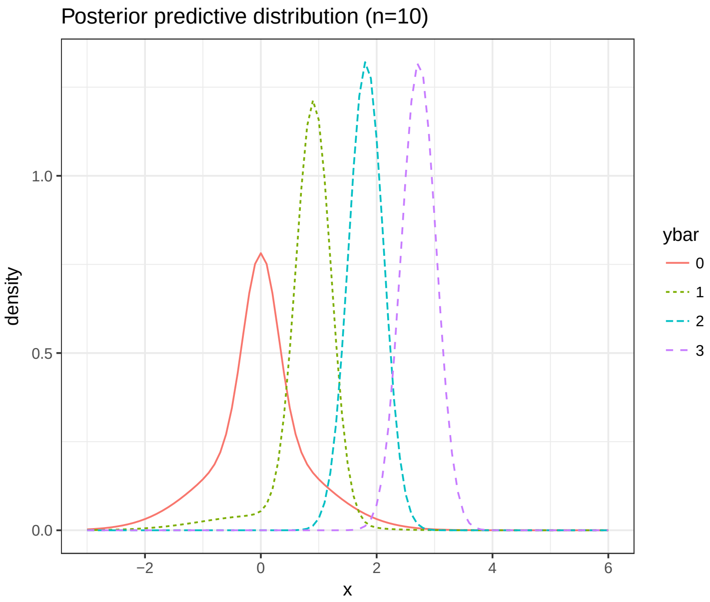
$$
\begin{aligned}
Y_i|M_0 &\stackrel{ind}{\sim} N(0,1) \\
Y_i|M_1,\mu &\stackrel{ind}{\sim} N(\mu,1),\ \mu|M_1 \sim N(0,1)
\end{aligned}
$$

Thus, we have the following posterior predictive distributions

$$
\begin{aligned}
\tilde{y}|y,M_0 &\sim N(0,1) \\
\tilde{y}|y,M_1 &\sim N\left(n\overline{y}[n+1]^{-1}, [n+1]^{-1}+1\right)
\end{aligned}
$$

and the following posterior model probabilities:

$$
\begin{aligned}
p(M_0|y) &\propto N(\overline{y};0,1/n)p(M_0) \\
p(M_1|y) &\propto N(\overline{y};0,2/n)p(M_0)
\end{aligned}
$$


Posterior predictive distribution (n=10)
density
1.0
0.5
0.0
−2
0
2
4
6
x
ybar
0
1
2
3

AIC/BIC model averaging

The generic structure for model averaging is

$$p(\tilde{y}|y) = \sum_{h=1}^{H} p(\tilde{y}|M_h, y) w_h$$

where w_h is the weight for model h.

Here are some possible weights:

- Bayesian model averaging: $w_h = p(M_h|y)$
- AIC model averaging: $w_h = e^{-\Delta_h/2}$ where $\Delta_h = AIC_h - \min AIC$
- AICc model averaging: $w_h = e^{-\Delta_h/2}$ where $\Delta_h = AICc_h - \min AICc$
- BIC model averaging: $w_h = e^{-\Delta_h/2}$ where $\Delta_h = BIC_h - \min BIC$

Information criterion

Recall that information criteria have the form:

$$IC = -2\log L(\hat{\theta}) + P$$

where P is a penalty. So if you take

$$w_h = e^{-\Delta_h/2} = e^{-(IC_h - \min IC)/2} \propto e^{-IC_h/2} = L_h(\hat{\theta})e^P.$$

where, if p is the number of parameters, the penalty P is

- AIC: $2p$
- AICc: $2p + 2p(p+1)/(n-p-1)$
- BIC: $p\log(n)$

The BIC is a large sample approximation to the marginal likelihood:

$$-2\log p(y) \approx -2\log p(y|\theta) + p\log(n) + C$$

Regression BMA

A common place to perform Bayesian Model Averaging is in the regression framework:

$$y \sim N(X_\gamma \beta_\gamma, \sigma^2_\gamma \mathrm{I})$$

where γ is a vector indicator of which of the p explanatory variables are included in model γ, e.g.

$$\gamma = (1, 1, 0, \ldots, 0, 1, 0)$$

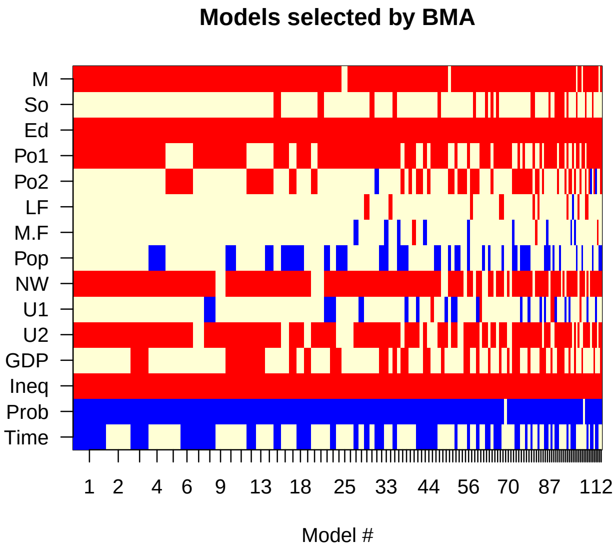
indicates the first, second, ..., and penultimate explanatory variables are included.

BIC model averaging in R

```
library(BMA)
library(MASS)
data(UScrime)
x<- UScrime[,-16]
y<- log(UScrime[,16])
x[,-2]<- log(x[,-2])
lma<- bicreg(x, y, strict = FALSE, OR = 20)
```

```
summary(lma)


##
## Call:
## bicreg(x = x, y = y, strict = FALSE, OR = 20)
##
##
##   115  models were selected
##  Best  5  models (cumulative posterior probability =  0.2039 ):
##
##            p!=0    EV         SD        model 1    model 2    model 3    model 4    model 5
## Intercept  100.0  -23.45301  5.58897  -22.63715  -24.38362  -25.94554  -22.80644  -24.50477
## M           97.3    1.38103  0.53531    1.47803    1.51437    1.60455    1.26830    1.46061
## So          11.7    0.01398  0.05640      .          .          .          .          .
## Ed         100.0    2.12101  0.52527    2.22117    2.38935    1.99973    2.17788    2.39875
## Po1         72.2    0.64849  0.46544    0.85244    0.91047    0.73577    0.98597      .
## Po2         32.0    0.24735  0.43829      .          .          .          .        0.90689
## LF           6.0    0.01834  0.16242      .          .          .          .          .
## M.F          7.0   -0.06285  0.46566      .          .          .          .          .
## Pop         30.1   -0.01862  0.03626      .          .          .       -0.05685      .
## NW          88.0    0.08894  0.05089    0.10888    0.08456    0.11191    0.09745    0.08534
## U1          15.1   -0.03282  0.14586      .          .          .          .          .
## U2          80.7    0.26761  0.19882    0.28874    0.32169    0.27422    0.28054    0.32977
## GDP         31.9    0.18726  0.34986      .          .        0.54105      .          .
## Ineq       100.0    1.38180  0.33460    1.23775    1.23088    1.41942    1.32157    1.29370
## Prob        99.2   -0.24962  0.09999   -0.31040   -0.19062   -0.29989   -0.21636   -0.20614
## Time        43.7   -0.12463  0.17627   -0.28659      .       -0.29682      .          .
##
## nVar                                      8          7          9          8          7
## r2                                      0.842      0.826      0.851      0.838      0.823
## BIC                                   -55.91243  -55.36499  -54.69225  -54.60434  -54.40788
## post prob                               0.062      0.047      0.034      0.032      0.029
```

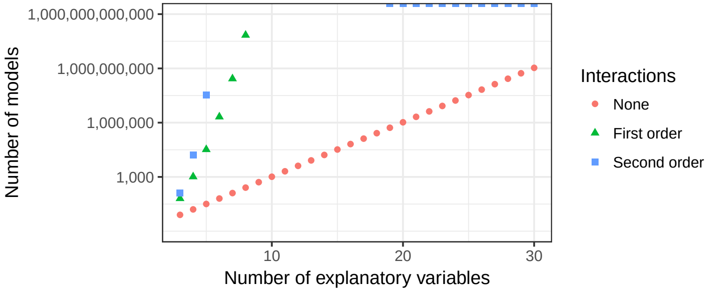
```
imageplot.bma(lma)
```



Model space

For all subsets regression analysis with p (continuous or binary) explanatory variables, we have

- 2^p models with no interactions,
- $2^{\binom{p}{2}}$ times as many models when considering first order interactions,
- $2^{\binom{p}{3}}$ times as many models when considering second order interactions,
- etc.



Model search in R

When model enumeration isn't possible, we resort to model search. There are many ways to search the model space, but one common approach is to use Markov chain Monte Carlo.

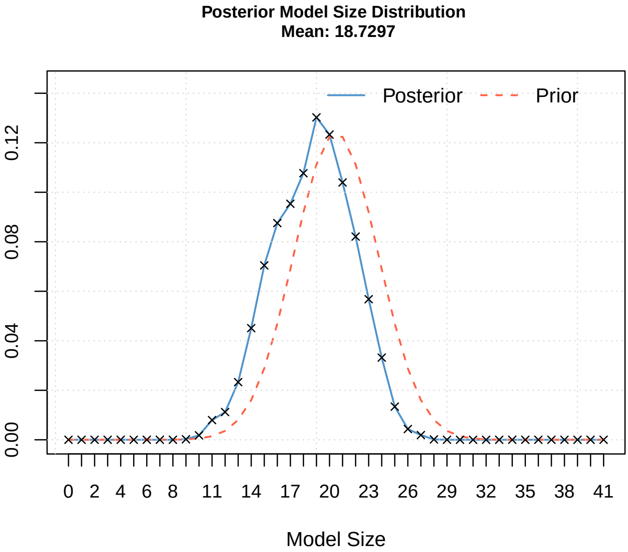
```
library(BMS)
data(datafls)


bma1 = bms(datafls,
           burn = 10000,
           iter = 20000,
           mprior = "uniform", # uniform prior over models
           user.int = FALSE)
```

If there is a uniform prior over models, what is the prior over model size (the number of explanatory variables included)?

```r
summary(bma1)
```

```
## Mean no. regressors               Draws             Burnins                Time  No. models visited
##           "18.7297"             "20000"             "10000"   "3.103621 secs"              "9589"
##      Modelspace 2^K           % visited         % Topmodels            Corr PMP            No. Obs.
##           "2.2e+12"           "4.4e-07"                "14"            "0.2538"                "72"
##         Model Prior             g-Prior     Shrinkage-Stats
##    "uniform / 20.5"               "UIP"         "Av=0.9863"
```

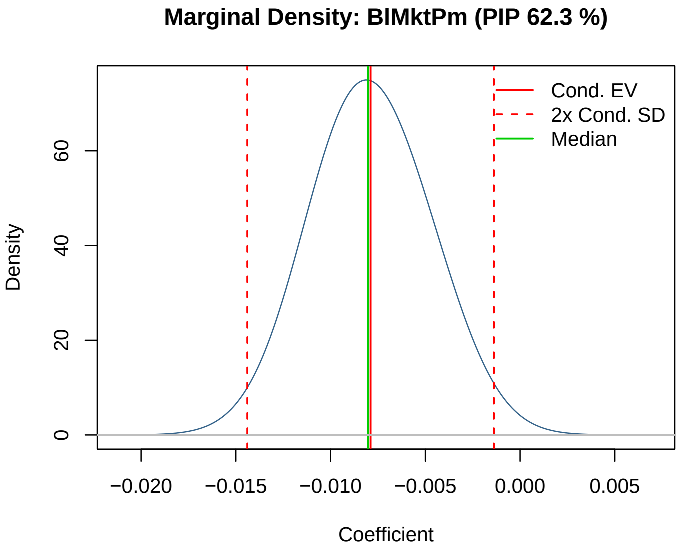
```r
plotModelsize(bma1)
```



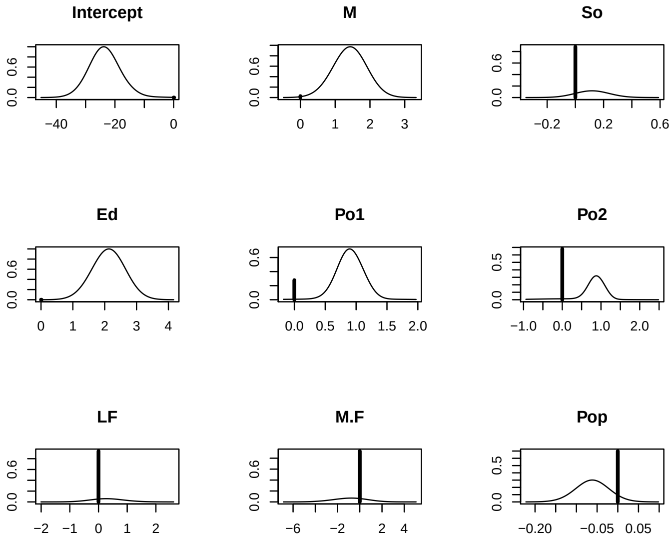
```r
coef(bma1)
```

```
##                 PIP     Post Mean      Post SD Cond.Pos.Sign Idx
## GDP60       1.00000 -1.661913e-02 2.914818e-03    0.00000000  12
## Confucian   1.00000  6.365228e-02 1.366694e-02    1.00000000  19
## LifeExp     0.98220  8.941333e-04 2.659293e-04    1.00000000  11
## EquipInv    0.94660  1.284790e-01 5.319508e-02    1.00000000  38
## Mining      0.88955  3.543130e-02 1.801326e-02    1.00000000  13
## SubSahara   0.85845 -1.562577e-02 8.509442e-03    0.00000000   7
## Hindu       0.78150 -5.361088e-02 3.976809e-02    0.01049264  21
## LabForce    0.71500  1.866232e-07 1.486065e-07    0.98951049  29
## NequipInv   0.68690  3.370527e-02 2.867741e-02    1.00000000  39
## RuleofLaw   0.66110  7.693599e-03 6.805646e-03    1.00000000  26
## BlMktPm     0.64195 -4.878769e-03 4.521999e-03    0.00000000  41
## Muslim      0.63020  8.782964e-03 8.431955e-03    0.99785782  23
## HighEnroll  0.61860 -5.439723e-02 5.316043e-02    0.00913353  30
## EthnoL      0.61495  6.979105e-03 6.874070e-03    0.99747947  20
## EcoOrg      0.55880  1.111902e-03 1.201848e-03    0.99991052  14
## PrScEnroll  0.53155  1.004388e-02 1.160750e-02    0.99115793  10
## Protestants 0.52885 -5.423567e-03 6.280994e-03    0.00000000  25
## LatAmerica  0.52805 -5.364023e-03 6.708976e-03    0.04081053   6
## CivlLib     0.48145 -1.082750e-03 1.496050e-03    0.01588950  34
## Buddha      0.37550  4.035118e-03 6.411328e-03    1.00000000  17
## Spanish     0.36935  3.321626e-03 5.599884e-03    0.97427914   2
## French      0.34830  2.230769e-03 4.135416e-03    0.97674419   3
## YrsOpen     0.34600  3.120581e-03 5.650901e-03    0.95910405  15
## PolRights   0.32550 -3.740090e-04 1.040922e-03    0.14946237  33
## English     0.31735 -2.285464e-03 4.179778e-03    0.00000000  35
## Age         0.27845 -1.186960e-05 2.375877e-05    0.00017957  16
## OutwarOr    0.27630 -8.231808e-04 1.705250e-03    0.01031488   8
## WarDummy    0.27325 -8.195590e-04 1.819551e-03    0.00494053   5
## Brit        0.25710  9.284513e-04 2.820460e-03    0.74640218   4
## PublEdupct  0.24525  4.151736e-02 9.666576e-02    0.94638124  31
## RFEXDist    0.23540 -9.416474e-06 2.274508e-05    0.01253186  37
## Catholic    0.22950 -1.082096e-04 3.497546e-03    0.45337691  18
```

```
density(bma1, reg="BlMktPm")
```



```
plot(lma)
```


Intercept
M
So
Ed
Po1
Po2
LF
M.F
Pop
0.0
0.6
0.5
−40
−20
0
1
2
3
−0.2
0.2
0.6
4
0.0
0.5
1.0
1.5
2.0
−1.0
−2
−1
−6
−0.20
−0.05
0.05

Model averaged parameters

Consider the following set of 4 models with $Y_i \stackrel{ind}{\sim} N(\mu_i, \sigma^2)$ where

$$\begin{array}{lllll}
M_1 : \mu_i & = \beta_0 & & \\
M_2 : \mu_i & = \beta_0 & +\beta_1 X_{i,1} & \\
M_3 : \mu_i & = \beta_0 & & +\beta_2 X_{i,2} \\
M_4 : \mu_i & = \beta_0 & +\beta_1 X_{i,1} & +\beta_2 X_{i,2}
\end{array}$$

It is tempting to want to obtain a model averaged posterior for the coefficients.

Model averaged parameters (cont.)

Perhaps we can write a model averaged posterior for a parameter as

$$p(\beta_1|y) = \sum_{h=1}^{H} p(\beta_1|y, M_h)p(M_h|y)$$

But β_1 means something entirely different in these models:

- In model M_2, β_1 is the effect of a one unit increase in $X_{i,1}$ on the expected response.
- In model M_4, β_1 is the effect of a one unit increase in $X_{i,1}$ on the expected response after adjusting for $X_{i,2}$.

More accurate model

Consider the following set of 4 models with $Y_i \stackrel{ind}{\sim} N(\mu_i, \sigma^2)$ where

$$
\begin{array}{llll}
M_1 : \mu_i & = \alpha_0 & & \\
M_2 : \mu_i & = \beta_0 & +\beta_1 X_{i,1} & \\
M_3 : \mu_i & = \gamma_0 & & +\gamma_2 X_{i,2} \\
M_4 : \mu_i & = \delta_0 & +\delta_1 X_{i,1} & +\delta_2 X_{i,2}
\end{array}
$$

Now it seems clear that we cannot average these parameters.

Assessing explanatory variable importance

To obtain some measure of how important a particular explanatory variable is we can find its posterior inclusion probability, i.e. the probability it is non-zero:

$$p(\beta_j \neq 0) = \sum_{h:\beta_j \neq 0} p(M_h|y)$$

which is just the sum of the model probabilities for the models where β_j is not zero.

```r
summary(lma)
```

```
##
## Call:
## bicreg(x = x, y = y, strict = FALSE, OR = 20)
##
##
##   115  models were selected
##  Best  5  models (cumulative posterior probability =  0.2039 ):
##
##            p!=0    EV         SD        model 1    model 2    model 3    model 4    model 5
## Intercept  100.0  -23.45301  5.58897  -22.63715  -24.38362  -25.94554  -22.80644  -24.50477
## M           97.3    1.38103  0.53531    1.47803    1.51437    1.60455    1.26830    1.46061
## So          11.7    0.01398  0.05640       .          .          .          .          .
## Ed         100.0    2.12101  0.52527    2.22117    2.38935    1.99973    2.17788    2.39875
## Po1         72.2    0.64849  0.46544    0.85244    0.91047    0.73577    0.98597       .
## Po2         32.0    0.24735  0.43829       .          .          .          .        0.90689
## LF           6.0    0.01834  0.16242       .          .          .          .          .
## M.F          7.0   -0.06285  0.46566       .          .          .          .          .
## Pop         30.1   -0.01862  0.03626       .          .          .       -0.05685       .
## NW          88.0    0.08894  0.05089    0.10888    0.08456    0.11191    0.09745    0.08534
## U1          15.1   -0.03282  0.14586       .          .          .          .          .
## U2          80.7    0.26761  0.19882    0.28874    0.32169    0.27422    0.28054    0.32977
## GDP         31.9    0.18726  0.34986       .          .        0.54105       .          .
## Ineq       100.0    1.38180  0.33460    1.23775    1.23088    1.41942    1.32157    1.29370
## Prob        99.2   -0.24962  0.09999   -0.31040   -0.19062   -0.29989   -0.21636   -0.20614
## Time        43.7   -0.12463  0.17627   -0.28659       .       -0.29682       .          .
##
## nVar                                      8          7          9          8          7
## r2                                      0.842      0.826      0.851      0.838      0.823
## BIC                                  -55.91243  -55.36499  -54.69225  -54.60434  -54.40788
## post prob                               0.062      0.047      0.034      0.032      0.029
```

```r
coef(bma1)
```

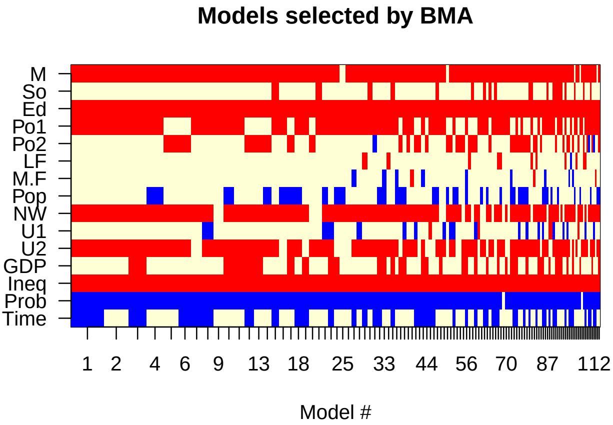
```
##                PIP     Post Mean      Post SD Cond.Pos.Sign Idx
## SubSahara   1.0000 -1.735413e-02 6.694849e-03    0.00000000   7
## LifeExp     1.0000  8.840297e-04 2.134909e-04    1.00000000  11
## GDP60       1.0000 -1.647188e-02 2.850791e-03    0.00000000  12
## Confucian   1.0000  6.491797e-02 1.314022e-02    1.00000000  19
## Mining      0.9960  4.080184e-02 1.327238e-02    1.00000000  13
## EquipInv    0.9720  1.288812e-01 4.827352e-02    1.00000000  38
## BlMktPm     0.8000 -6.541257e-03 4.316219e-03    0.00000000  41
## Hindu       0.7960 -5.686685e-02 3.658916e-02    0.00000000  21
## EthnoL      0.7855  8.987239e-03 6.517067e-03    1.00000000  20
## LabForce    0.7495  1.960980e-07 1.392868e-07    0.99933289  29
## RuleofLaw   0.7175  8.095049e-03 6.415198e-03    1.00000000  26
## PrScEnroll  0.6575  1.184038e-02 1.141578e-02    1.00000000  10
## Muslim      0.6460  8.747158e-03 7.568580e-03    1.00000000  23
## Protestants 0.6285 -6.043979e-03 5.977904e-03    0.00000000  25
## HighEnroll  0.6285 -6.042433e-02 5.489875e-02    0.00000000  30
## NequipInv   0.6185  2.773995e-02 2.725924e-02    1.00000000  39
## CivlLib     0.5710 -1.338009e-03 1.497246e-03    0.00000000  34
## EcoOrg      0.5515  1.044343e-03 1.134098e-03    1.00000000  14
## LatAmerica  0.4935 -4.694489e-03 6.014098e-03    0.00000000   6
## English     0.3415 -2.217346e-03 3.920341e-03    0.00146413  35
## Buddha      0.3285  3.166856e-03 5.550291e-03    0.99238965  17
## Age         0.3150 -1.410528e-05 2.526417e-05    0.00000000  16
## PublEdupct  0.2970  4.956740e-02 1.061428e-01    0.89898990  31
## RFEXDist    0.2935 -1.068002e-05 2.205369e-05    0.00000000  37
## Spanish     0.2835  1.851451e-03 3.913293e-03    0.98236332   2
## PolRights   0.2300 -1.446898e-04 8.571074e-04    0.31739130  33
## Brit        0.2230  2.282650e-04 1.893096e-03    0.59417040   4
## Catholic    0.2060 -4.743627e-04 2.931547e-03    0.37135922  18
## French      0.1820  8.885071e-04 2.509963e-03    1.00000000   3
## YrsOpen     0.1800  1.694299e-03 4.319902e-03    0.95833333  15
## Jewish      0.1750  2.562079e-04 4.089714e-03    0.64571429  22
## OutwarOr    0.1600 -5.137417e-04 1.391118e-03    0.00000000   8
```

Multiple posterior inclusion probability

If explanatory variables are correlated, then it is possible to have low posterior incluion probability for the correlated explanatory variable, but the probability of at least one of the explanatory variables being included is high.
For example,

$$P(\beta_i \neq 0 \text{ or } \beta_j \neq 0|y) = \sum_{h:\beta_i \neq 0 \text{ or } \beta_j \neq 0} p(M_h|y)$$

```
imageplot.bma(lma)
```




```
cor(UScrime$Po1, UScrime$Po2)

## [1] 0.9935865
```

Model selection

Sometimes, we will want to select a model. Selecting model M_h is clearly justified if $p(M_h|y) \approx 1$.

If forced to choose a model, it might seem that choosing the model with the highest $p(M_h|y)$ would be the way to go, but Barbieri and Berger (2004) show that if prediction is the goal, then the median probability model is better. The median probability model is the model that includes all explanatory variables whose posterior inclusion probability is greater than $1/2$.