Bayesian Model Averaging and Variable Selection

STA721: Lecture 18

Merlise Clyde
Duke University

https://sta702-F23.github.io/website/

US Air Example

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

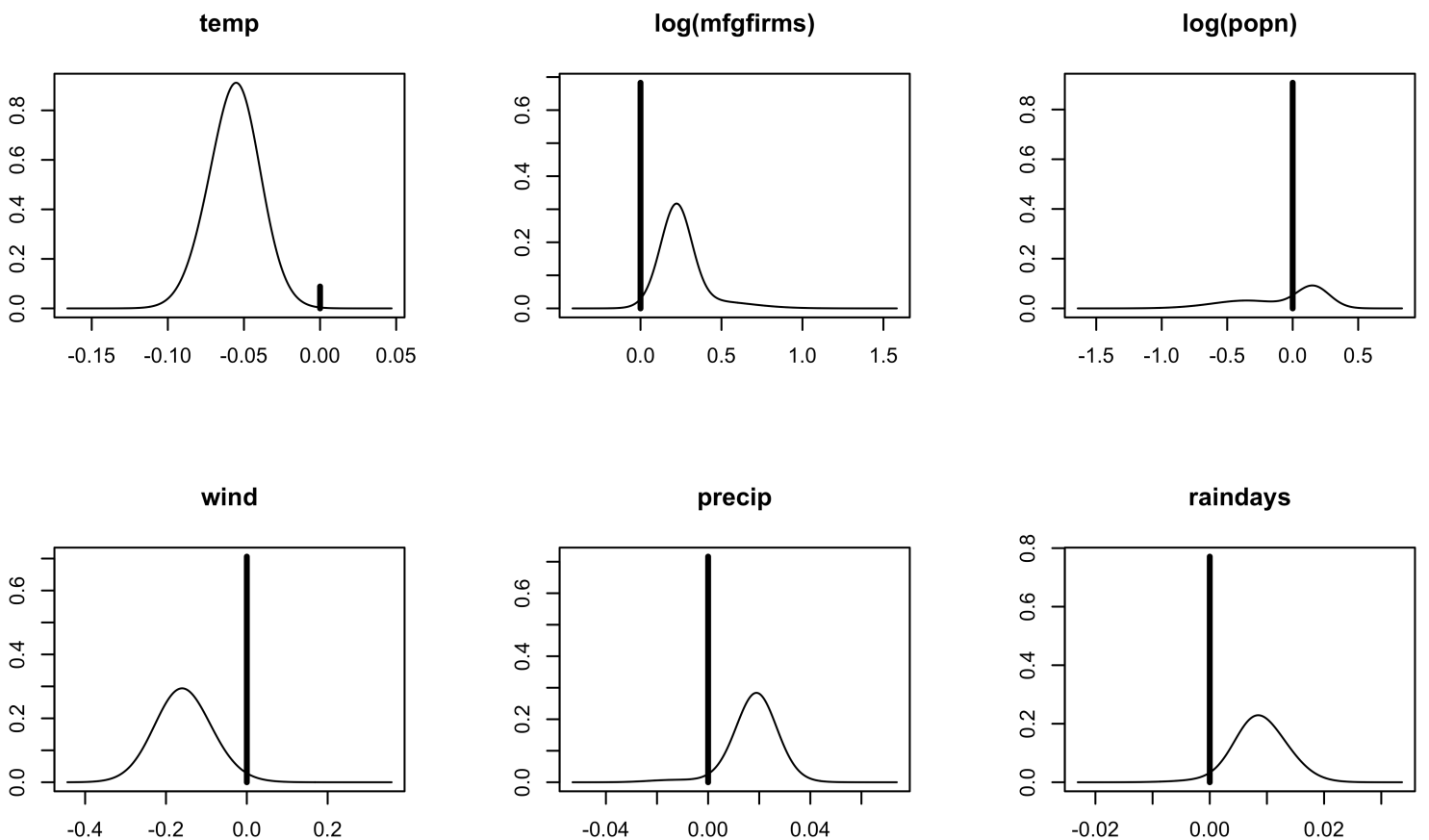

Summary

```
1  summary(poll.bma, n.models=4)
```

	P(B != 0 | Y)	model 1	model 2	model 3	model 4
Intercept	1.00000000	1.00000	1.0000000	1.0000000	1.0000000
temp	0.91158530	1.00000	1.0000000	1.0000000	1.0000000
log(mfgfirms)	0.31718916	0.00000	0.0000000	0.0000000	1.0000000
log(popn)	0.09223957	0.00000	0.0000000	0.0000000	0.0000000
wind	0.29394451	0.00000	0.0000000	0.0000000	1.0000000
precip	0.28384942	0.00000	1.0000000	0.0000000	1.0000000
raindays	0.22903262	0.00000	0.0000000	1.0000000	0.0000000
BF	NA	1.00000	0.3286643	0.2697945	0.2655873
PostProbs	NA	0.29410	0.0967000	0.0794000	0.0781000
R2	NA	0.29860	0.3775000	0.3714000	0.5427000
dim	NA	2.00000	3.0000000	3.0000000	5.0000000
logmarg	NA	3.14406	2.0313422	1.8339656	1.8182487

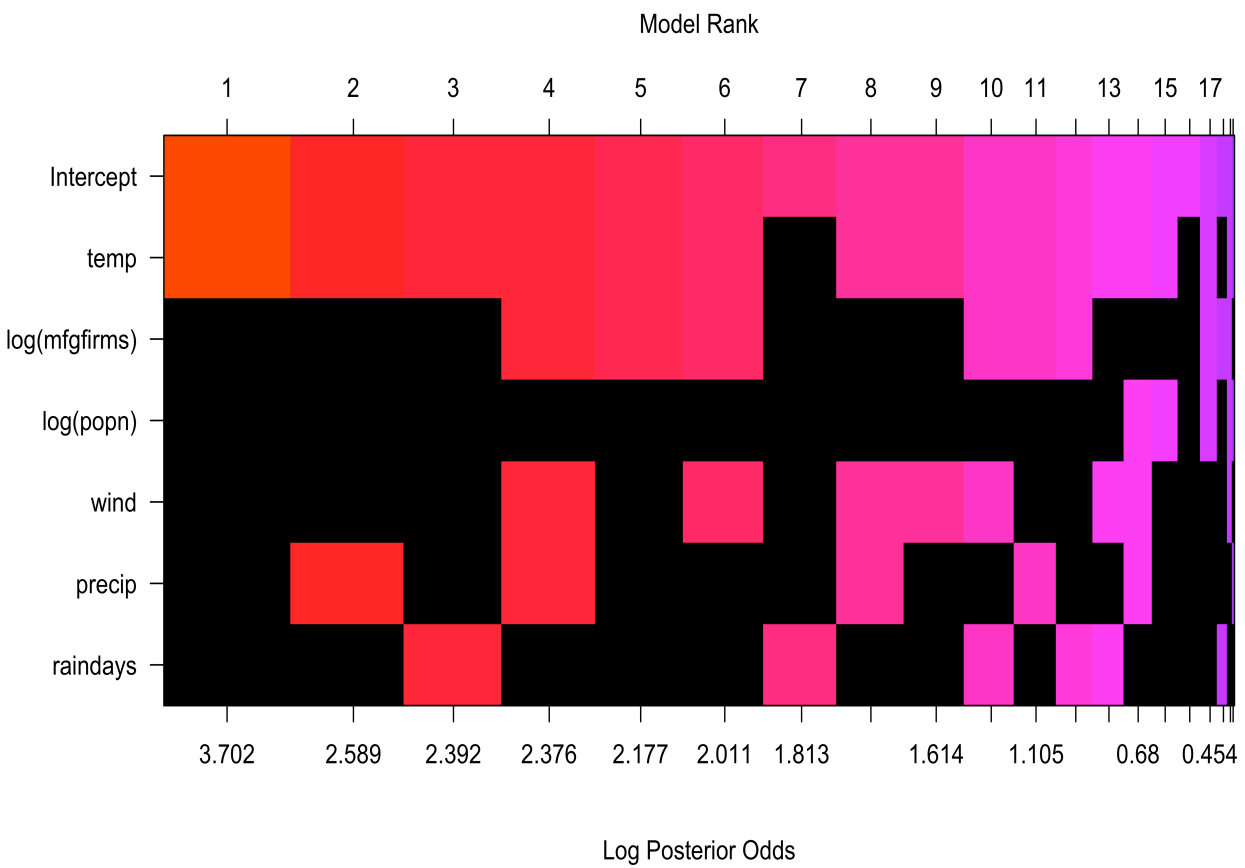
Plots of Coefficients

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



Posterior Distribution with Uniform Prior on Model Space

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

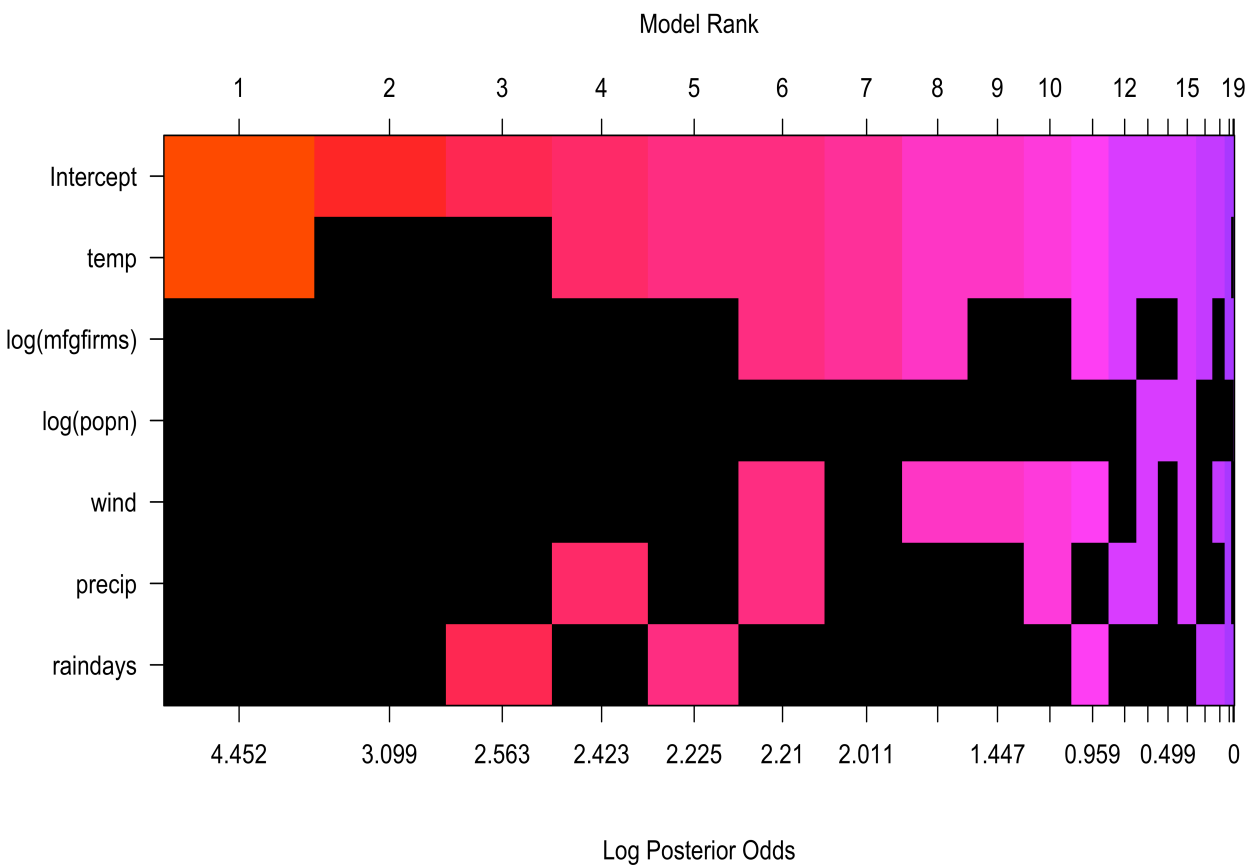


Posterior Distribution with BB(1,1) Prior on Model Space

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


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

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



Diabetes Example

```r
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 ...
 $ hdl     : num  -0.911 1.563 -0.679 -0.757 0.171 ...
 $ tch     : num  -0.0544 -0.8294 -0.0544 0.7205 -0.0544 ...
 $ ltg     : num  0.4181 -1.4349 0.0601 0.4765 -0.6718 ...
 $ glu     : num  -0.371 -1.936 -0.545 -0.197 -0.979 ...
```


MCMC with BAS

```
library(BAS)
diabetes.bas = bas.lm(y ~ ., data=diabetes.train,
                      prior = "JZS",
                      method="MCMC",
                      n.models = 10000,
                      MCMC.iterations=500000,
                      thin = 10,
                      initprobs="eplogp",
                      force.heredity=FALSE)
```

```
   user  system elapsed
 10.538   0.574  11.113
```

```
[1] "number of unique models 5905"
```

- increase `MCMC.iterations`?
- check diagnostics

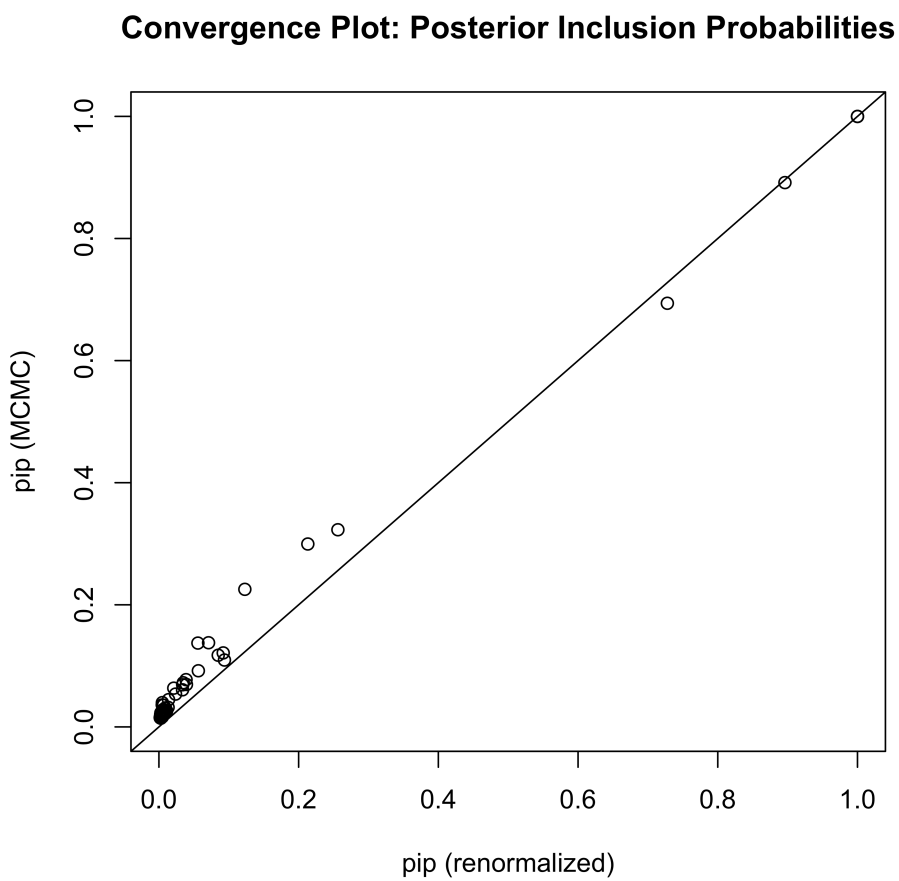
Estimates of Posterior Probabilities

- relative frequencies $\hat{P}_{RF}(\boldsymbol{\gamma} \mid \mathbf{Y}) = \frac{\#\text{ times } \boldsymbol{\gamma} \in S}{S}$
  - ergodic average converges to $p(\boldsymbol{\gamma} \mid \mathbf{Y})$ as $S \to \infty$
  - asymptoptically unbaised
- renormalized posterior probabilities $\hat{P}_{RN}(\boldsymbol{\gamma} \mid \mathbf{Y}) = \frac{p(\mathbf{Y}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma} \in S} p(\mathbf{Y}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})}$
  - also asymptoptically unbaised
  - Fisher consistent (e.g if we happen to enumerate all models in S iterations we recover the truth)
- if we run long enough the two should agree
- also look at other summaries i.e posterior inclusion probabilities

$$\hat{p}(\gamma_j = 1 \mid \mathbf{Y}) = \sum_{S} \gamma_j \hat{P}(\boldsymbol{\gamma} \mid \mathbf{Y})$$

Diagnostic Plot

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



- model probabilities converge much slower!

Out of Sample Prediction

- What is the optimal value to predict $\mathbf{Y}^{\text{test}}$ given $\mathbf{Y}$ under squared error?
- Iterated expectations leads to BMA for $\mathsf{E}[\mathbf{Y}^{\text{test}} \mid \mathbf{Y}]$
- Prediction under model averaging

$$\hat{Y} = \sum_{S} (\hat{\alpha}_{\gamma} + \mathbf{X}_{\gamma}^{\text{test}} \hat{\boldsymbol{\beta}}_{\gamma}) \hat{p}(\boldsymbol{\gamma} \mid \mathbf{Y})$$

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

```
[1] 0.4556414
```


Credible Intervals & Coverage

- posterior predictive distribution

$$p(\mathbf{y}^{\text{test}} \mid \mathbf{y}) = \sum_{\boldsymbol{\gamma}} p(\mathbf{y}^{\text{test}} \mid \mathbf{y}, \boldsymbol{\gamma}) p(\boldsymbol{\gamma} \mid \mathbf{y})$$

- integrate out α and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ to get a normal predictive given ϕ and $\boldsymbol{\gamma}$ (and $\mathbf{y}$)
- integrate out ϕ to get a t distribution given $\boldsymbol{\gamma}$ and $\mathbf{y}$
- credible intervals via sampling
  - sample a model from $p(\boldsymbol{\gamma} \mid \mathbf{y})$
  - conditional on a model sample $y \sim p(\mathbf{y}^{\text{test}} \mid \mathbf{y}, \boldsymbol{\gamma})$
  - compute quantiles from sample y

```
ci.bas = confint(pred.bas);
coverage = mean(diabetes.test$y > ci.bas[,1] & diabetes.test$y <
coverage
```

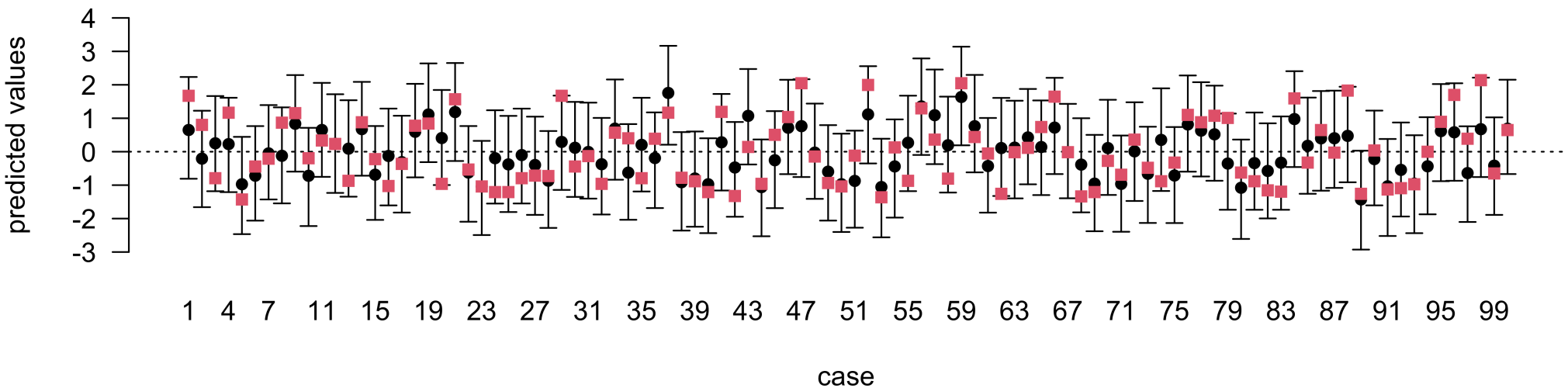
```
[1] 1
```


95% Prediction intervals

```
1 plot(ci.bas)
```

```
NULL
```

```
1 points(diabetes.test$y, col=2, pch=15)
```



Selection and Prediction

- BMA - optimal for squared error loss Bayes

$$\mathsf{E}[\|\mathbf{Y}^{\text{test}} - a\|^2 \mid \mathbf{y}] = \mathsf{E}[\|\mathbf{Y}^{\text{test}} - \mathsf{E}[\mathbf{Y}^{\text{test}} \mid \mathbf{y}]\|^2 \mid \mathbf{y}] + \|\mathsf{E}[\mathbf{Y}^{\text{test}} \mid \mathbf{y}] - a\|^2$$

- What if we want to use only a single model for prediction under squared error loss?
- HPM: Highest Posterior Probability model is optimal for selection, but not prediction
- MPM: Median Probability 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)

Example

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

```
[1] 0.4740667
```

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

```
[1] 0.98
```


Theory - Consistency of g-priors

Summary

- Choice of prior on $\boldsymbol{\beta}_\gamma$
  - orthogonally invariant priors - multivariate Spike & Slab
  - products of independent Spike & Slab priors
  - non-semi-conjugate
- priors on the models (sensitivity)
- computation (MCMC, "stochastic search", variational, orthogonal data augmentation, reversible jump-MCMC)
- posterior summaries - select a model or "average" over all models

Other aspects of model selection?

- transformations of $\mathbf{Y}$
- functions of $\mathbf{X}$: interactions or nonlinear functions such as splines kernels
- choice of error distribution