

Lecture 17: Bayesian Variable Selection and Model Averaging

STA702

Merlise Clyde
Duke University

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



Diabetes Example

```

1 set.seed(8675309)
2 source("yX.diabetes.train.txt")
3 diabetes.train = as.data.frame(diabetes.train)
4 source("yX.diabetes.test.txt")
5 diabetes.test = as.data.frame(diabetes.test)
6 colnames(diabetes.test)[1] = "y"
7
8 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 ...
 $ age^2  : num  -0.312 -0.867 1.925 2.176 -0.857 ...

```

MCMC with BAS

```
1 library(BAS)
2 diabetes.bas = bas.lm(y ~ ., data=diabetes.train,
3                       prior = "JZS",
4                       method="MCMC",
5                       n.models = 10000,
6                       MCMC.iterations=500000,
7                       thin = 10,
8                       initprobs="eplogp",
9                       force.heredity=FALSE)
```

```
      user  system elapsed
19.523    0.951   20.530
```

```
[1] "number of unique models 5008"
```

- increase `MCMC.iterations`?
- check diagnostics

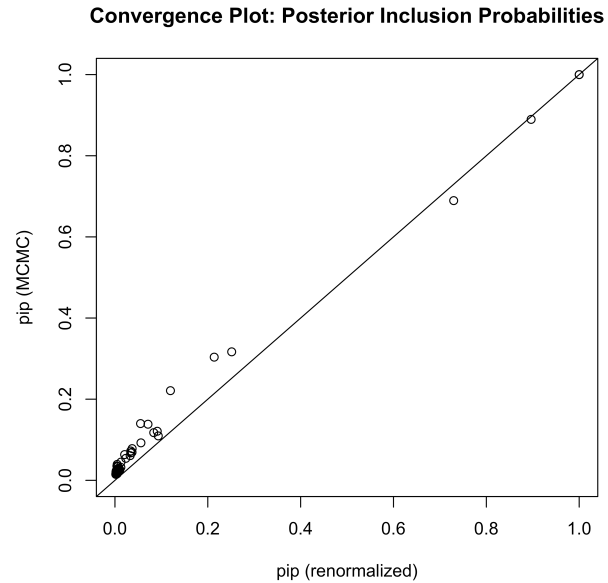
Estimates of Posterior Probabilities

- 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 $S \rightarrow \infty$
 - asymptotically unbiased
- 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 asymptotically unbiased
 - 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}^{test} given \mathbf{Y} under squared error?
- Iterated expectations leads to BMA for $E[\mathbf{Y}^{\text{test}} \mid \mathbf{Y}]$
- Prediction under model averaging

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

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

```
[1] 0.4558026
```

Credible Intervals & Coverage

- posterior predictive distribution

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

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

```
1 ci.bas = confint(pred.bas);
2 coverage = mean(diabetes.test$y > ci.bas[,1] & diabetes.test$y < ci.bas[,2])
3 coverage
```

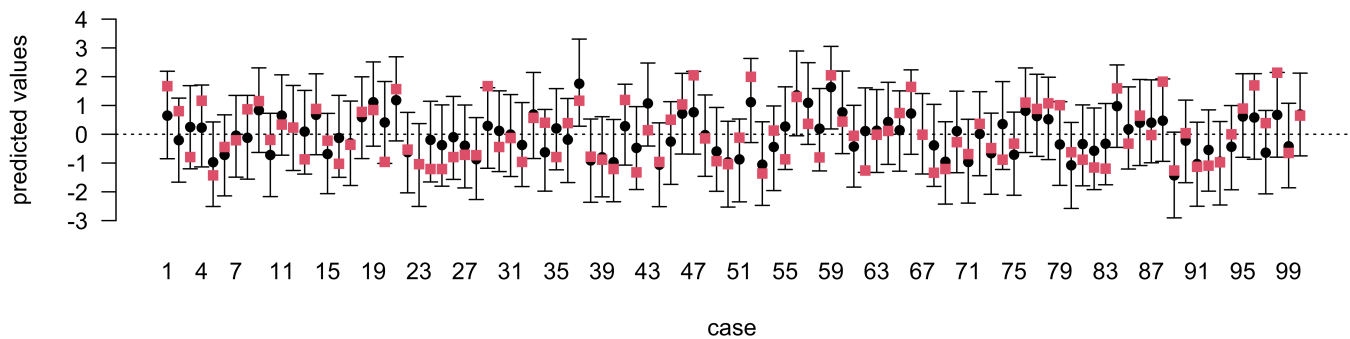
```
[1] 0.99
```

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

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

- 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

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

```
[1] 0.4740667
```

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

```
[1] 0.98
```

Theory

- consistency of mixtures of g-priors
- desire that posterior probability of model goes to 1 as $n \rightarrow \infty$
 - does not always hold if the null model is true (may be highest posterior probability model)
 - need prior on g to depend on n
 - asymptotically BMA collapses to the true model
- other quantities may converge i.e. posterior mean
- what if the true model is not in Γ ? What can we say?
 - \mathcal{M} -closed; BMA converges to the model that is “closest” to the truth in Kullback-Leibler divergence
 - \mathcal{M} -complete
 - \mathcal{M} -open
 - latter is related to “stacking” which is a frequentist method of ensemble learning using cross-validation; see Clyde & Iversen (2013) for the curious

Summary

- Choice of prior on β_γ
 - 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 Y
- functions of X : interactions or nonlinear functions such as splines kernels
- choice of error distribution