Lecture 17: Bayesian Variable Selection and Model Averaging

STA702

Merlise Clyde
Duke University

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 ...
 $ age^2  : num  -0.312 -0.867 1.925 2.176 -0.857 ...
```

MCMC with BAS

```r
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 
 19.523   0.951  20.530 
```

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

- increase `MCMC.iterations`?
- check diagnostics

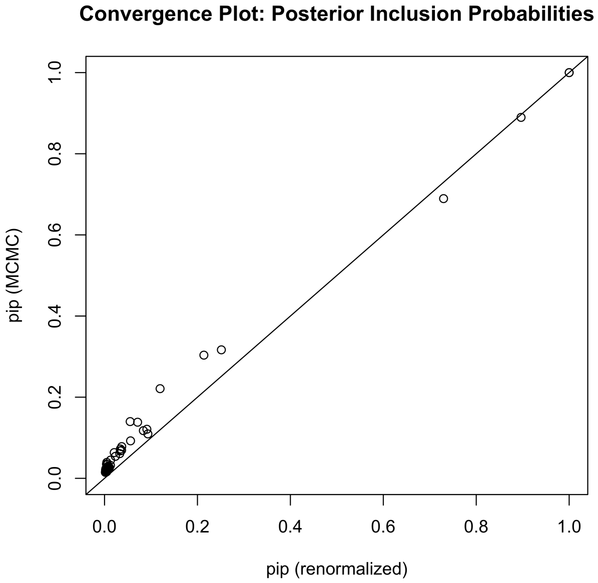
Estimates of Posterior Probabilities

- 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 $S \to \infty$
    - 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

```r
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 $\textsf{E}[\mathbf{Y}^{\text{test}} \mid \mathbf{Y}]$
- Prediction under model averaging

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

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

```
[1] 0.4558026
```

Credible Intervals & Coverage

- posterior predictive distribution

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

- integrate out α and $\boldsymbol{\beta}_{\gamma}$ 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 sammple y

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

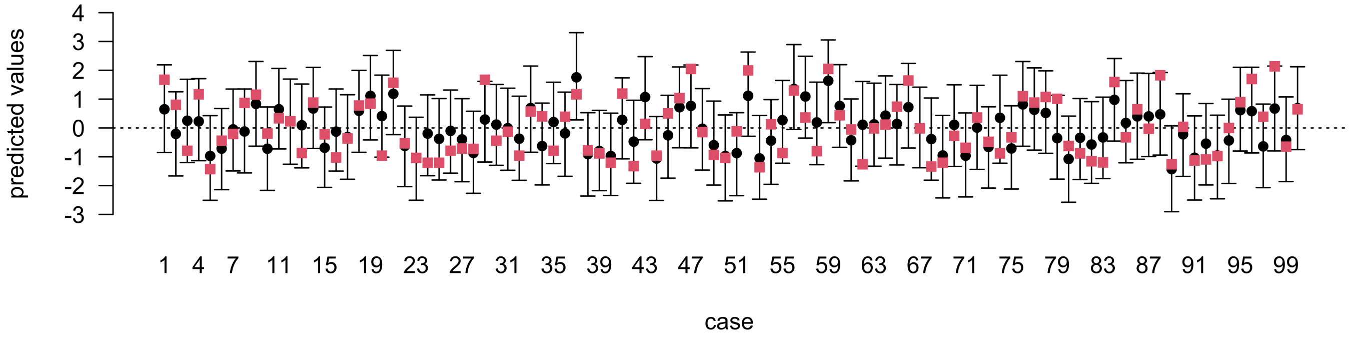
```
[1] 0.99
```

95% Prediction intervals

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

```
NULL
```

```
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 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)

Example

```
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
```

Theory - Consistency of g-priors

- desire that posterior probability of model goes to 1 as $n \to \infty$
  - does not alwyas hold if the null model is true (may be highest posterior probability model)
  - need prior on g to depend on n (rules out EB and fixed g-priors with $g \neq n$)
  - asymptotically BMA collapses to the true model
- other quantities may converge i.e. posterior mean
- what if the true model $\boldsymbol{\gamma}_T$ is not in Γ? What can we say?
  - $\boldsymbol{\mathcal{M}}$-complete; BMA converges to the model that is "closest" to the truth in Kullback-Leibler divergence
  - $\boldsymbol{\mathcal{M}}$-closed; realize that $(p\boldsymbol{\gamma}) = 0 \forall \boldsymbol{\gamma} \in \mathbf{G}$ and is nonsense but know $\boldsymbol{\gamma}_T$, however want to use models in $\mathbf{G}$ only
  - $\boldsymbol{\mathcal{M}}$-open; realize that $(p\boldsymbol{\gamma}) = 0 \forall \boldsymbol{\gamma} \in \mathbf{G}$ and is nonsense but know $\boldsymbol{\gamma}_T$
  - 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 $\boldsymbol{\beta}_\gamma$
  - 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 Y
- functions of X: interactions or nonlinear functions such as splines kernels
- choice of error distribution