Bayesian Model Uncertainty

STA721: Lecture 19

Merlise Clyde
Duke University

Recap Diabetes Data

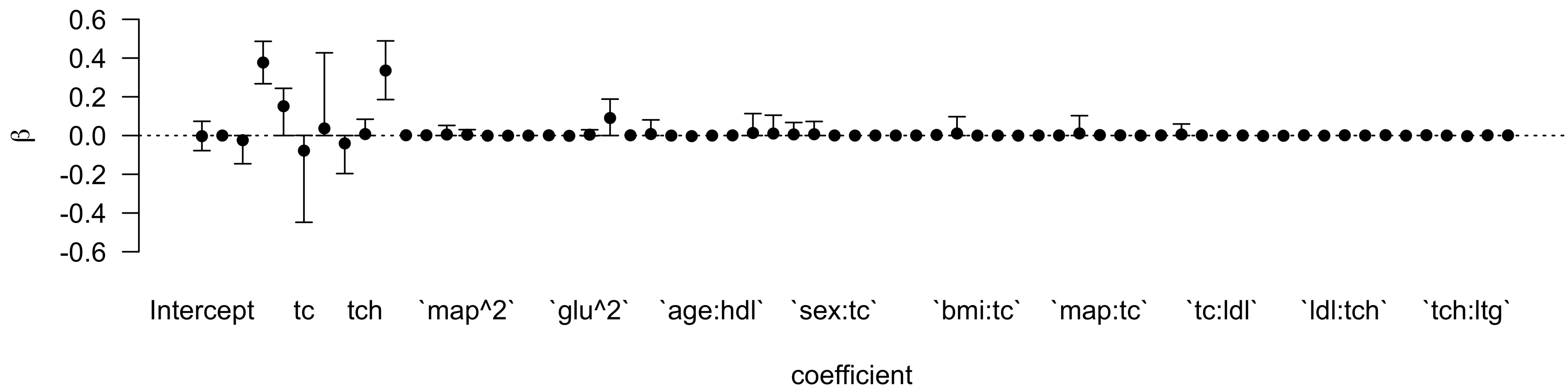
```r
set.seed(8675309)
source("yX.diabetes.train.txt")
diabetes.train = as.data.frame(diabetes.train)
source("yX.diabetes.test.txt")
diabetes.test = as.data.frame(diabetes.test)
colnames(diabetes.test)[1] = "y"

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 ...
```

Credible Intervals under BMA

```r
coef.diabetes = coefficients(diabetes.bas)
ci.coef.bas = confint(coef.diabetes, level=0.95)
plot(ci.coef.bas)
```

```
NULL
```



- uses Monte Carlo simulations from the posteriors of the coefficients
- uses HPD intervals from the CODA package to compute intervals

Out of Sample Prediction

- What is the optimal value to predict $\mathbf{Y}^{\text{test}}$ given $\mathbf{Y}$ under squared error?
- BMA is optimal prediction for squared error loss with Bayes

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

- Iterated expectations leads to BMA for $\mathsf{E}[\mathbf{Y}^{\text{test}} \mid \mathbf{Y}]$
- Prediction under model averaging

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

```
[1] 0.4556414
```

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 HPD or quantiles from samples of y

95% Prediction intervals

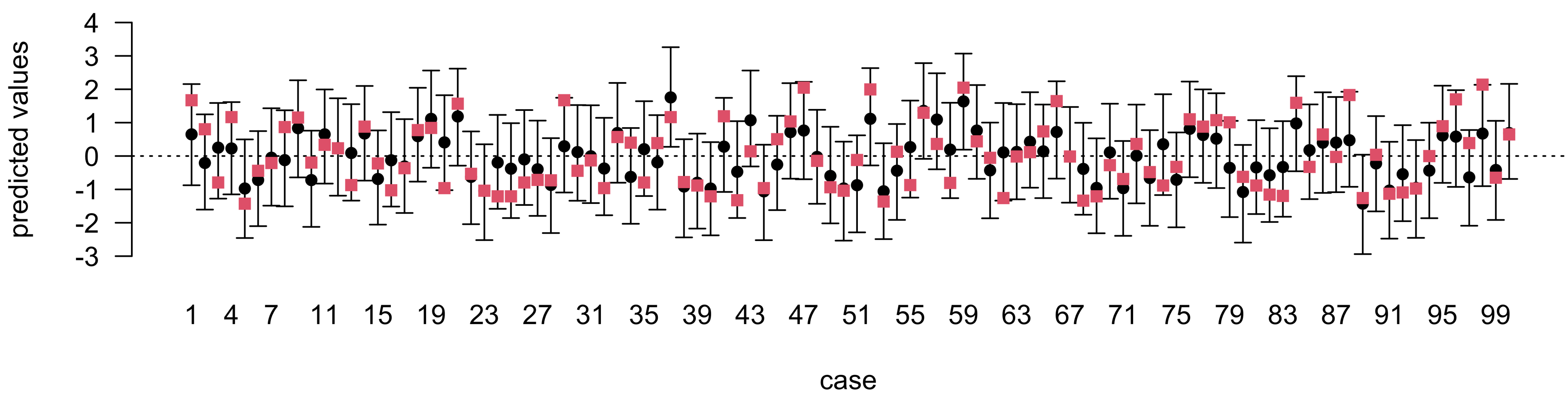
```
ci.bas = confint(pred.bas);
coverage = mean(diabetes.test$y > ci.bas[,1] & diabetes.test$y <
coverage
```

```
[1] 0.99
```

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

```
NULL
```

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



Selection and Prediction

- BMA is optimal for squared error loss Bayes
- 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)
- costs of using variables in prediction?

Example

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

```
[1] 0.4556414
```

```
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 always 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

Model Paradigms

- 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;
  - know $\boldsymbol{\gamma}_T \notin \mathbf{G}$ so that $(p_{\boldsymbol{\gamma}}) = 0 \; \forall \boldsymbol{\gamma} \in \mathbf{G}$ but want to use models in $\mathbf{G}$
  - Predictive distribution $p(\mathbf{Y}^* \mid \mathbf{Y}, \boldsymbol{\gamma}_T)$ is available
- $\boldsymbol{\mathcal{M}}$-open;
  - know $\boldsymbol{\gamma}_T \notin \mathbf{G}$ so that $(p_{\boldsymbol{\gamma}}) = 0 \; \forall \boldsymbol{\gamma} \in \mathbf{G}$ but want to use models in $\mathbf{G}$
  - Predictive distribution $p(\mathbf{Y}^* \mid \mathbf{Y}, \boldsymbol{\gamma}_T)$ is not available. (too complicated to use, etc)

$\mathcal{M}$-Open and M-Complete Prediction

Clyde & Iversen (2013) pdf motivate a solution via decision theory

- Use the models in $\mathbf{G}$ to predict $\mathbf{Y}^*$ given $\mathbf{Y}$ under squared error loss

$$E[\mathbf{Y}^*, \sum_{\gamma \in \mathbf{G}} \omega_\gamma \hat{\mathbf{Y}}^*_\gamma \mid \mathbf{Y}] = \int (\mathbf{Y}^* - \sum_{\gamma \in \mathbf{G}} \omega_\gamma \hat{\mathbf{Y}}^*_\gamma)^2 p(\mathbf{Y}^* \mid \mathbf{Y})$$

- Still use a weighted sum of predictions or densities from models in $\mathbf{G}$ but now the weights are not probabilities but are chosen to minimize the loss function
  - uses additional constraints of penalties on the weights as part of the loss function
  - need to approximate the predictive distribution for $\mathbf{Y}^* \mid \mathbf{Y}$ (via an approximate Dirichlet Process Model)
  - latter is related to “stacking” (Wolpert 1972) which is a frequentist method of ensemble learning using cross-validation;

Summary

- Choice of prior on $\boldsymbol{\beta}_\gamma$
  - multivariate Spike & Slab
  - products of independent Spike & Slab priors
  - intermediates block g-priors
  - non-semi-conjugate
  - non-local priors
  - shrinkage priors without point-masses
- priors on the models (sensitivity)
- computation (MCMC, “stochastic search”, adaptive MH, variational, orthogonal data augmentation, reversible jump-MCMC)
- decision theory - select a model or “average” over all models
- asymptotic properties - large n and large $p > n$

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
- outliers