Mixtures of Prior Distributions

Hoff Chapter 9, Liang et al 2007, Hoeting et al (1999), Clyde & George (2004)

November 3, 2015

Bartlett's Paradox

The Bayes factor for comparing $\mathcal{M}_\gamma$ to the null model:

$$BF(\mathcal{M}_\gamma : \mathcal{M}_0) = (1+g)^{(n-1-p_\gamma)/2}(1+g(1-R^2_\gamma))^{-(n-1)/2}$$

For $g \to \infty$, the $BF \to 0$ for fixed n and R^2_γ

Information Paradox

The Bayes factor for comparing $\mathcal{M}_\gamma$ to the null model:

$$BF(\mathcal{M}_\gamma : \mathbf{\mathcal{M}}_0) = (1+g)^{(n-1-p_\gamma)/2}(1+g(1-R^2))^{-(n-1)/2}$$

- Let g be a fixed constant and take n fixed.
- Let $F = \frac{R^2_\gamma/p_\gamma}{(1-R^2_\gamma)/(n-1-p_\gamma)}$
- As $R^2_\gamma \to 1$, $F \to \infty$ LR test would reject $\mathbf{\mathcal{M}}_0$ where F is the usual F statistic for comparing model $\mathcal{M}_\gamma$ to $\mathbf{\mathcal{M}}_0$
- BF converges to a fixed constant $(1+g)^{-p_\gamma/2}$ (does not go to infinity

"Information Inconsistency" see Liang et al JASA 2008

Mixtures of g priors & Information consistency

Need $BF \to \infty$ if $\mathrm{R}^2 \to 1 \Leftrightarrow \mathrm{E}_g[(1+g)^{-p_\gamma/2}]$ diverges (proof in Liang et al)

- Zellner-Siow Cauchy prior
- hyper-g prior (Liang et al JASA 2008)

$$p(g) = \frac{a-2}{2}(1+g)^{-a/2}$$

  or $g/(1+g) \sim Beta(1,(a-2)/2)$ need $2 < a \leq 3$

- Hyper-g/n $(g/n)(1+g/n) \sim (Beta(1,(a-2)/2)$
- Jeffreys prior on g corresponds to $a = 2$ (improper)
- robust prior (Bayarrri et al Annals of Statistics 2012
- Intrinsic prior (Womack et al JASA 2015)

All have prior tails for β that behave like a Cauchy distribution and (the latter 4) marginal likelihoods that can be computed using special hypergeometric functions (${}_2F_1$, Appell F_1)

Desiderata - Bayarri et al 2012 AoS

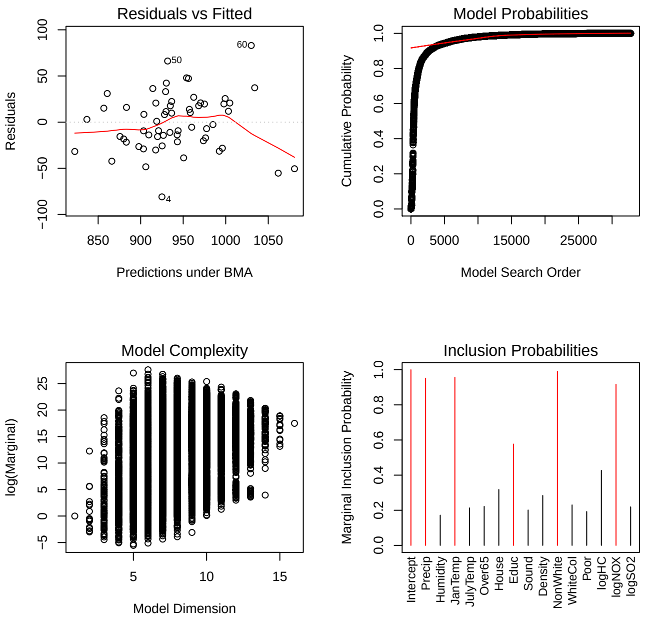
- Proper priors on non-common coefficients
- If LR overwhelmingly rejects a model, Bayesian should also reject
- Selection Consistency: large samples probability of the true model goes to one.
- Intrinsic prior consistency (prior converges to a fixed proper prior as $n \to \infty$
- Invariance (invariance under scale/location changes of data/model leads to $p(\beta_0, \phi) \propto 1/\phi$); other group invariance, rotation invariance.
- predictive matching: predictive distributions match under minimal sample sizes so that $BF = 1$

Mortality & Pollution

- Data from Statistical Sleuth 12.17
- 60 cities
- response Mortality
- measures of HC, NOX, SO2
- Is pollution associated with mortality after adjusting for other socio-economic and meteorological factors?
- 15 predictor variables implies $2^{15} = 32,768$ possible models
- Use Zellner-Siow Cauchy prior $1/g \sim G(1/2, n/2)$

```
mort.bma = bas.lm(MORTALITY ~ ., data=mortality,
                  prior="ZS-null",
                  alpha=60,  n.models=2^15,
                  update=100, initprobs="eplogp")
```

Posterior Distributions



Posterior Probabilities

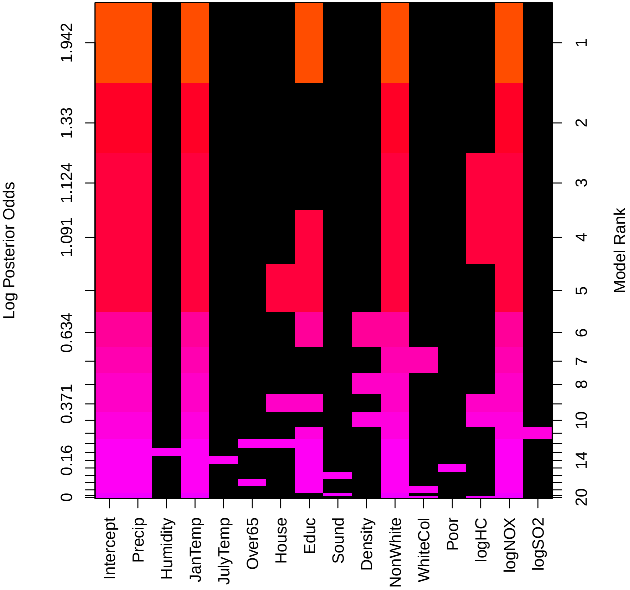
- What is the probability that there is no pollution effect?
- Sum posterior model probabilities over all models that include no pollution variables

  ```
  > which.mat = list2matrix.which(mort.bma,1:(2^15))
  > poll.in = (which.mat[, 14:16] %*% rep(1, 3)) > 0
  > sum(poll.in * mort.bma$postprob)
  [1] 0.9889641
  ```

- Posterior probability no effect is 0.011
- Posterior Odds that there is an effect $(1 - .011)/(.011) = 89.$
- Prior Odds $7 = (1 - .5^3)/.5^3$
- Bayes Factor for a pollution effect $89.9/7 = 12.8$
- Bayes Factor for NOX based on marginal inclusion probability $0.917/(1 - 0.917) = 11.0$
- Marginal inclusion probability for logHC $= 0.427144$ $(BF = .745)$
- Marginal inclusion probability for logSO2 $= 0.218978$ $(BF = .280)$

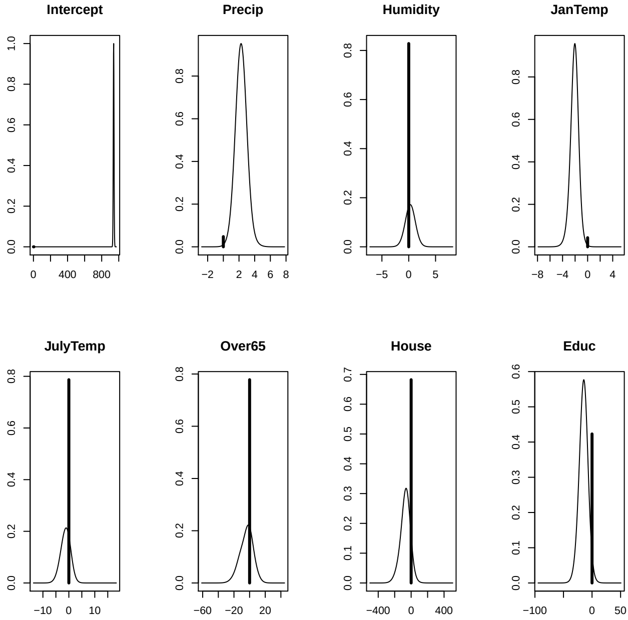
Bayes Factors are not additive!

Model Space

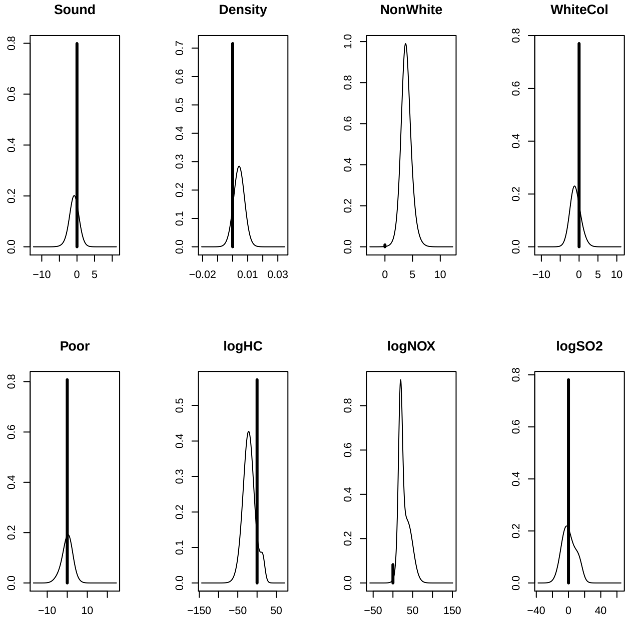


Log Posterior Odds
1.942
1.33
1.124
1.091
0.634
0.371
0.16
0
Model Rank
1
2
3
4
5
6
7
8
10
14
20
Intercept
Precip
Humidity
JanTemp
JulyTemp
Over65
House
Educ
Sound
Density
NonWhite
WhiteCol
Poor
logHC
logNOX
logSO2

Coefficients



Coefficients



Effect Estimation

- Coefficients in each model are adjusted for other variables in the model
- OLS: leave out a predictor with a non-zero coefficient then estimates are biased!
- Model Selection in the presence of high correlation, may leave out "redundant" variables;
- improved MSE for prediction (Bias-variance tradeoff)
- Bayes is biased anyway so should we care?
- What is meaning of $\sum_{\boldsymbol{\gamma}} \beta_{j\boldsymbol{\gamma}} \gamma_j P(\mathcal{M}_\gamma \mid \mathbf{Y})$

With confounding, should not use plain BMA. Need to change prior?

Other Problems

- Computational if $p > 35$ enumeration is difficult
  - Gibbs sampler or Random-Walk algorithm on γ
  - poor convergence/mixing with high correlations
  - Metropolis Hastings algorithms more flexibility (method="MCMC")
  - "Stochastic Search" (no guarantee samples represent posterior)
  - Variational, EM, etc to find modal model
  - in BMA all variables are included, but coefficients are shrunk to 0; alternative is to use Shrinkage methods
  - Models with Non-estimable parameters? (use generalized inverse)
- Prior Choice: Choice of prior distributions on $\boldsymbol{\beta}$ and on γ

Model averaging versus Model Selection – what are objectives?

BAS Algorithm - Clyde, Ghosh, Littman - JCGS

Sampling w/out Replacing