(Empirical) Bayes Model Uncertainty Introduction and a New Prior

Víctor Peña, James O. Berger

Department of Statistical Science
Duke University

June 9th, 2016

Outline

1. Why should you care about model uncertainty?
2. Bayesian and Non-Bayesian Approaches
3. Default Priors
4. Our contribution: *a new default prior*

Model Uncertainty (and Selection)

Model Uncertainty

- In applied modeling, we typically report inferences using **a single model** that seems to fit the data well enough.
- However, we tend to ignore that we used some *procedure* (formal or informal) to select it.
- As a result, our inferences are typically **overconfident**, and our confidence and prediction intervals are **too narrow**.
- We typically don't acknowledge that there is **substantial** model uncertainty.
- Excellent introductions (with a Bayesian slant) are Draper (1995), Clyde and George (2004).

How to Deal with Model Uncertainty

- Some couple of ways to deal with model uncertainty:
  - **Nonparametric approaches:** fit a model so big that if the *truth* exists, it must be a particular case of your model.
  - **“Corrected” inferences:** report corrected p-values, intervals, etc. that take model uncertainty into account.
  - **(Discrete) model combination/averaging:** consider a finite (but possibly large) class of models, and combine them.
- Today, we’ll focus on the latter.

Model Averaging

Bayesian Model Averaging

- We have a set of M models $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_M$, with sampling densities $p(y \mid \theta, \mathcal{M}_1)$, $p(y \mid \theta, \mathcal{M}_2)$, ... , $p(y \mid \theta, \mathcal{M}_M)$.
- We're Bayesians now, so we have to specify a full probability model; that is, we need:
  - A pmf on the model space: assign $P(\mathcal{M}_i)$ for $i \in \{1, 2, \ldots, M\}$.
  - Densities for θ given the models: define $p(\theta \mid \mathcal{M}_i)$ for $i \in \{1, 2, \ldots, M\}$.
- If you have prior information (e.g. you might know which variables are likely to be "active"), you can use it; if you don't (or don't want to use it), what can we do? (next section).

Bayesian Model Averaging

- Why is model uncertainty taken care of?
  - The pmf $\{P(\mathcal{M}_1), P(\mathcal{M}_2), \ldots, P(\mathcal{M}_M)\}$ represents our model uncertainty **before** seeing the data.
  - The pmf $\{P(\mathcal{M}_1 \mid y), P(\mathcal{M}_2 \mid y), \ldots, P(\mathcal{M}_M \mid y)\}$ represents our model uncertainty **after** seeing the data.
- For example, if we're doing regression and we have a new individual with covariates X^*, the predictive distribution of her outcome Y^* after seeing the data is the weighted average:

$$p(Y^* \mid X^*, y) = \sum_{i=1}^{M} P(\mathcal{M}_i \mid y)\, p(Y^* \mid \mathcal{M}_i, X^*, y)$$

Non-Bayesian Model Averaging

What if we're not Bayesians?

- We'd like to acknowledge model uncertainty ... *somehow*.
- Back to our regression example, we could see the posterior probabilities as "weights," and evaluate the performance of $p(Y^* \mid X^*, y)$ from a frequentist perspective.

Non-Bayesian Model Averaging

- In general, we can use the Bayesian machinery to come up with ("admissible") frequentist procedures (complete class theorems) that acknowledge model uncertainty.
- Formal criteria for "objective Bayes" model selection (Bayarri et al., 2012) can aid in finding them.
  - These are rules that say: "If a model selection procedure is to be labeled as objective, this should (or shouldn't) happen."
- We can evaluate performance using asymptotics, simulation studies, etc.

Default Priors

Default Bayes Model Averaging (and Selection)

- A very nice introduction to the topic is Berger et al. (2001).
- Recall that we have to specify a pmf on the model space $P(\mathcal{M}_i)$ and densities for the parameters given the models $p(\theta \mid \mathcal{M}_i)$.
- Today, we'll focus on the latter (see Scott et al. (2010) for a discussion on default priors on the model space) in the context of the **normal linear model**.

Marginal Likelihoods

- Posterior probabilities of models depend on the data only through the marginal likelihood of the models $p(y \mid \mathcal{M}_i)$:

$$P(\mathcal{M}_j \mid y) = \frac{P(\mathcal{M}_j)P(y \mid \mathcal{M}_j)}{\sum_{i=1}^{M} P(\mathcal{M}_i)P(y \mid \mathcal{M}_i)}.$$

- The marginal likelihood is found by averaging (or "weighting") the likelihood $p(y \mid \mathcal{M}_j, \theta)$ with respect to $p(\theta \mid \mathcal{M}_j)$:

$$\begin{aligned} p(y \mid \mathcal{M}_j) &= \int_{\Theta} p(y \mid \mathcal{M}_j, \theta)\, p(\theta \mid \mathcal{M}_j)\, \mathrm{d}\theta \\ &= \mathbb{E}_{p(\theta \mid \mathcal{M}_j)}[p(y \mid \mathcal{M}_j, \theta)] \end{aligned}$$

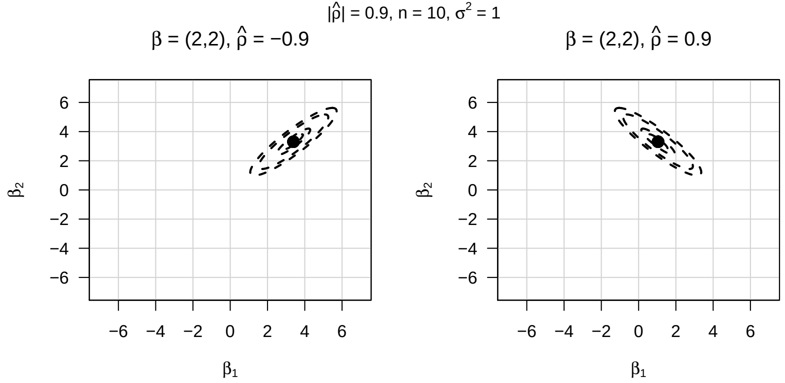
Marginal Likelihoods: Linear Models

- In the context of the **normal linear model** with n observations and p predictors ($n > p$),

$$Y = X\beta + \varepsilon, \qquad \varepsilon \sim \mathrm{N}_n(0_n, \sigma^2 I_n)$$

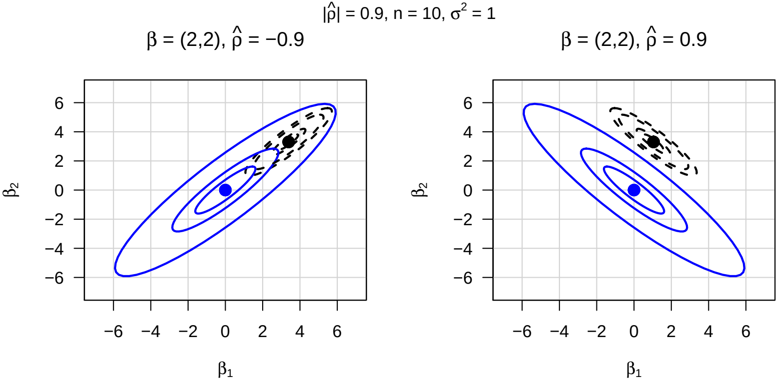
The likelihood of β (given σ^2 and X) is proportional to $\mathrm{N}_p(\widehat{\beta}, \sigma^2(X'X)^{-1})$, where $\widehat{\beta} = (X'X)^{-1}X'Y$.

- Assume $n = 10$, $p = 2$, $\beta = (2, 2)'$, and $\sigma^2 = 1$. Standardize X and consider two cases: sample correlation between predictors $\widehat{\rho} \in \{-0.9, 0.9\}$.
- The contours of the likelihood look like this:



Key Example - Unit Information Prior

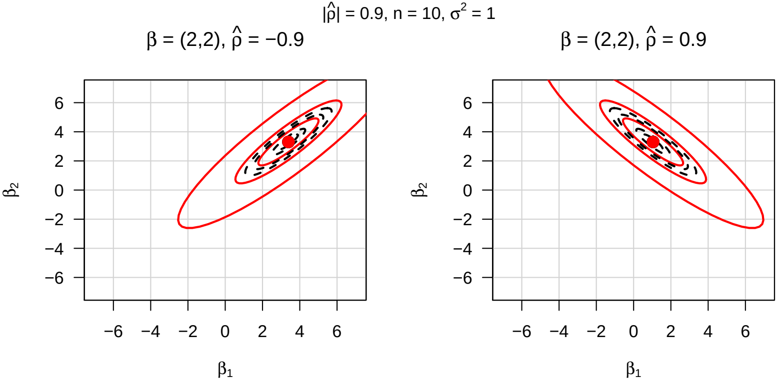
- A commonly used prior is $\beta \mid \sigma^2, X \sim \mathrm{N}_p(0_p, \sigma^2 n(X'X)^{-1})$ (unit information prior).



Marginal likelihoods depend on the sign of $\widehat{\rho}$

Key Example - BIC

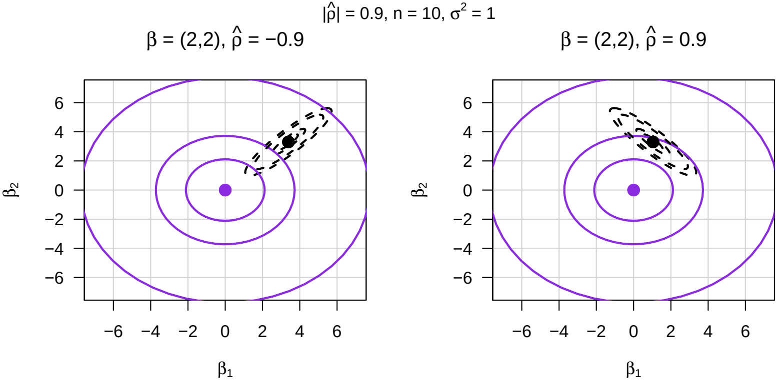
- Another prior is $\beta \mid \sigma^2, X \sim \mathrm{N}_p(\widehat{\beta}, \sigma^2 n(X'X)^{-1})$ ("BIC" prior).



Seems too aggressive

Key Example - Independence

- Yet another prior is $\beta \mid \sigma^2, X \sim \mathrm{N}_p(0_p, \sigma^2 n\, I_p)$ ("independent" prior).



Posterior prob. depends on units, ignores "shape" of likelihood

Motivation of our work

- Can we find a compromise between $\mathrm{N}_p(0_p, \sigma^2 n(X'X)^{-1})$ (UIP) and $\mathrm{N}_p(\widehat{\beta}, \sigma^2 n(X'X)^{-1})$ (BIC)?
- Idea: Can we define a prior that is centered at 0_p, tries to "catch" the likelihood, and is at least as disperse (in some sense) as $\sigma^2 n(X'X)^{-1}$?

Our Work

Our Formal Setup

- Variable selection in linear models of the form:

$$
\begin{aligned}
Y \mid X_0, X_i, \beta_0, \beta, \sigma^2 &\sim N_n(X_0\beta_0 + X_i\beta_i, \sigma^2 I_n), \\
\beta_i \mid \sigma^2, W_i &\sim N_p(0_p, \sigma^2 W_i) \\
\pi(\beta_0, \sigma^2) &\propto 1/\sigma^2,\ \ X_0' X = 0_{p_0 \times p}.
\end{aligned}
$$

- Y is an $n \times 1$ vector, X_0 is an $n \times p_0$ matrix with "common predictors", X_i is an $n \times p_i$ design matrix with model-specific predictors, and the predictors in X_0 and X_i are orthogonal. Throughout, assume $n > p_0 + p$.
- The use of $\pi(\beta_0, \sigma^2) \propto 1/\sigma^2$ can be justified by invariance arguments given in Bayarri et al. (2012).

What is W_i?

- We want a prior that tries to "catch" the likelihood, but it is at most as informative as the UIP.
- We set the matrix W_i to

$$\widehat{W_i} = \arg\max_{W_i \succeq n(X'X)^{-1}} m_i(Y \mid W_i),$$

where

$$m_i(Y \mid W_i) = \int f(Y \mid X_0, X_i, \beta_0, \beta_i, \sigma^2)\pi(\beta_0, \beta_i, \sigma^2)\, \mathrm{d}\beta_0\, \mathrm{d}\beta_i\, \mathrm{d}\sigma^2$$

and $A \succeq B$ if $A - B$ is positive semidefinite (Loewner ordering).

- Quite surprisingly, $\widehat{W_i}$ has a closed-form expression!

Why is $\widehat{W}$ sensible?

Why is $n(X'X)^{-1}$ a reasonable lower bound?

- Expected information of β is $(X'X)/\sigma^2$, so $(X'X)/(n\sigma^2)$ contains (roughly) the same information as a “typical” observation in the sample (Hoff, 2009).
- Reasonable default choice given predictive matching results in Bayarri et al. (2012).

Why is $\widehat{W}$ sensible?

What does $W \succeq n(X'X)^{-1}$ mean?

- It implies $\mathrm{tr}(W) \geq \mathrm{tr}(n(X'X)^{-1})$ and $\det(W) \geq \det(n(X'X)^{-1})$. Traces and determinants are sometimes used for measuring "total variability" and/or "size" of matrices.
- If π_1 is the UIP and π_2 is the W-prior, $E_{\pi_1} f(\beta) \leq E_{\pi_2} f(\beta)$ for convex f (Müller, 2001). For example, this is true for volume of HPD sets and L^p norms.
- If σ^2 is known, $W \succeq n(X'X)^{-1}$ implies that W leads to inferences for β that are, in some sense, at least as good as those with $n(X'X)^{-1}$ (Hansen and Torgersen, 1974; Goel and Ginebra, 2003)

What is $\widehat{W}$?

- $\widehat{W}$ can be written as

$$\boxed{\begin{aligned} \widehat{W} &= a\,\widehat{\beta}\widehat{\beta}' + n(X'X)^{-1}, \\ a &= \max\left(0, (n - p_0 - 1)/\mathrm{SSE} - (n+1)/\mathrm{SSR}\right) \end{aligned}}$$

where $\mathrm{SSR} = \widehat{\beta}'(X'X)\widehat{\beta}$, $\mathrm{SSE} = Y'(I_n - P_{X_0} - P_X)Y$ (P_{X_0}, and P_X are perpendicular projection operators with onto the column spaces of X_0 and X, respectively)

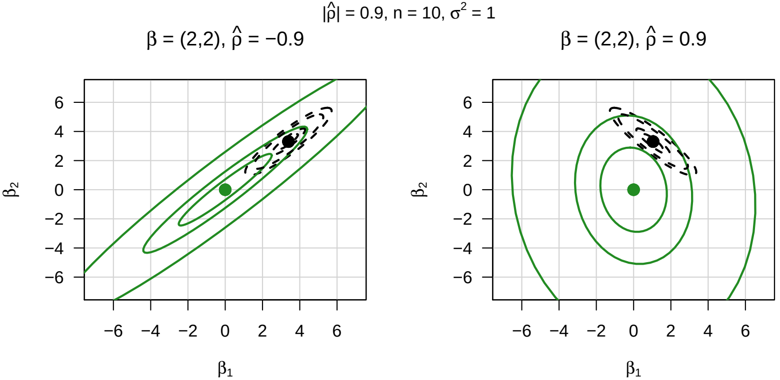
Interpreting $\widehat{W}$

$$\boxed{\begin{aligned} \widehat{W} &= a\widehat{\beta}\widehat{\beta}' + n(X'X)^{-1}, \\ a &= \max\left(0, (n - p_0 - 1)/\mathrm{SSE} - (n+1)/\mathrm{SSR}\right) \end{aligned}}$$

- The global maximum over W is proportional to the rank 1 matrix $\widehat{\beta}\widehat{\beta}'$. Therefore, $\widehat{W}$ is a linear combination of the global maximum and the lower bound $n(X'X)^{-1}$.
- $\widehat{W}$ is equal to $n(X'X)^{-1}$ when SSE (residual sum of squares) is big relative to SSR (explained/“regression” sum of squares).

Key Example

- Our prior is $\beta \sim \mathrm{N}_p(0_p, \sigma^2 \widehat{W})$.



Example: Two Predictors

- Assume $p = 2$, $\beta = (2, 2)'$, and $\sigma^2 = 1$. Standardize X and consider two cases: sample correlation between predictors $\widehat{\rho} \in \{-0.9, 0.9\}$.
- Average posterior probability of the true model after $B = 10^4$ simulations:

	$\widehat{\rho} = 0.9$				$\widehat{\rho} = -0.9$			
n	BIC	$\widehat{W}$	UIP	Ind.	BIC	$\widehat{W}$	UIP	Ind.
20	0.953	0.917	0.546	0.761	0.905	0.818	0.811	0.704
25	0.988	0.980	0.778	0.930	0.973	0.949	0.949	0.984
30	0.996	0.994	0.925	0.983	0.990	0.984	0.984	0.972

Very similar to BIC

Properties

- In general, very close to BIC.
- If the *truth* is contained on our list of models, its posterior probability converges to 1.
- Posterior probabilities are invariant with respect to measurement units
- In the context of estimation, the resulting posterior mean is minimax with respect to scaled squared loss.
- It has other properties not discussed here (e.g. information consistency)

Conclusions

- Model uncertainty is important, but often ignored.
- There are approaches at the interface of Bayesian and non-Bayesian statistics with good properties.
- I presented an intuitively appealing approach, which behaves very similarly to BIC.
- BIC is perceived to be very aggressive, but it might not be.

References I

Bayarri, M., Berger, J., Forte, A., García-Donato, G., et al. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics*, 40(3):1550–1577.

Berger, J. O., Pericchi, L. R., Ghosh, J., Samanta, T., De Santis, F., Berger, J., and Pericchi, L. (2001). Objective bayesian methods for model selection: introduction and comparison. *Lecture Notes-Monograph Series*, pages 135–207.

Clyde, M. and George, E. I. (2004). Model uncertainty. *Statistical science*, pages 81–94.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 45–97.

References II

Goel, P. K. and Ginebra, J. (2003). When is one experiment 'always better than'another? *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(4):515–537.

Hansen, O. H. and Torgersen, E. N. (1974). Comparison of linear normal experiments. *The Annals of Statistics*, pages 367–373.

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*. Springer Science & Business Media.

Müller, A. (2001). Stochastic ordering of multivariate normal distributions. *Annals of the Institute of Statistical Mathematics*, 53(3):567–575.

Scott, J. G., Berger, J. O., et al. (2010). Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38(5):2587–2619.