Lecture 3: Essentials of Bayesian Model Uncertainty

Jim Berger

Duke University

CBMS Conference on Model Uncertainty and Multiplicity
July 23-28, 2012

Outline

- Possible goals for Bayesian model uncertainty analysis
- Formulation and motation for Bayesian model uncertainty
- Model averaging
- Attractions of the Bayesian approach
- Challenges with the Bayesian approach
- Approaches to prior choice for model uncertainty
- Selecting a single model for prediction

I. Possible Goals for Bayesian Model Uncertainty Analysis

Goal 1. Selection of the true model from a collection of models that contains the truth.

- decision-theoretically, often viewed as choice of 0-1 loss in model selection

Goal 2. Selection of a model good for prediction of future observations or phenomena

- decision-theoretically, often viewed as choice of squared error loss for predictions, or Kullback-Liebler divergence for predictive distributions

Goal 3. Finding a simple approximate model that is close enough to the true model to be useful. (This is purely a utility issue.)

Goal 4. Finding important covariates, important relationships between covariates, and other features of model structure. (Often, the model is not of primary importance; it is the relationships of covariates in the model, or whether certain covariates should even be in the model, that are of interest.)

Caveat 1: Suppose the models under consideration do not contain the true model, often called the *open model* scenario.

Positive fact: Bayesian analysis will asymptotically give probability one to the model that is as close as possible to the true model (in Kullback Leibler divergence), among the models considered, so the Bayesian approach is still viable.

Issues of concern:

- Bayesian internal estimates of accuracy may be misleading (Barron, 2004).
- Model averaging may no longer be optimal.
- There is no obvious reason why models need even be ‘plausible’.

Caveat 2: There are two related but quite different statistical paradigms involving model uncertainty:

- Model criticism or model checking (Lecture 9)
- The process of model development

II. Formulation and Notation for Bayesian Model Uncertainty

Models (or hypotheses). Data, $\boldsymbol{x}$, is assumed to have arisen from one of several models:

M_1: $\boldsymbol{x}$ has density $f_1(\boldsymbol{x} \mid \boldsymbol{\theta}_1)$

M_2: $\boldsymbol{x}$ has density $f_2(\boldsymbol{x} \mid \boldsymbol{\theta}_2)$

$$\ddots$$

M_q: $\boldsymbol{x}$ has density $f_q(\boldsymbol{x} \mid \boldsymbol{\theta}_q)$

Assign prior probabilities, $P(M_i)$, to each model. Common is

$$P(M_i) = \frac{1}{q}\,,$$

but this is inappropriate in multiple testing scenarios.

Example: In variable selection, where each of m possible β_i could be in or out of the model, to control multiplicity

- let each variable, β_i, be independently in the model with unknown probability p (called the prior inclusion probability);
- assign p a uniform distribution.
- Then, if M_i has k_i variables,

  $$P(M_i) = \int_0^1 p^{k_i}(1-p)^{m-k_i}dp = Beta(1+k_i, 1+m-k_i) = \frac{k_i!(m-k_i)!}{(m+1)!} .$$

  (Using equal model probabilities would give $P(M_i) = 2^{-m}$ for all M_i.)
- This is equivalent to assigning probability $1/(m+1)$ to all models of a given size, with that mass divided equally among the models of the given size (as recommended by Jeffreys, 1961).

Under model M_i :

- Prior density of $\boldsymbol{\theta}_i$: $\boldsymbol{\theta}_i \sim \pi_i(\boldsymbol{\theta}_i)$
- Marginal density of $\boldsymbol{x}$:

$$m_i(\boldsymbol{x}) = \int f_i(\boldsymbol{x} \mid \boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i$$

  measures “how likely is $\boldsymbol{x}$ under M_i”

- Posterior density of $\boldsymbol{\theta}_i$:

$$\pi_i(\boldsymbol{\theta}_i \mid \boldsymbol{x}) = \pi(\boldsymbol{\theta}_i \mid \boldsymbol{x}, M_i) = \frac{f_i(\mathbf{x} \mid \boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)}{m_i(\boldsymbol{x})}$$

  quantifies (posterior) beliefs about $\boldsymbol{\theta}_i$ if the true model was M_i.

Bayes factor of M_j to M_i:

$$B_{ji} = \frac{m_j(\boldsymbol{x})}{m_i(\boldsymbol{x})}$$

Posterior probability of M_i:

$$P(M_i \mid \boldsymbol{x}) = \frac{P(M_i) m_i(\boldsymbol{x})}{\sum_{j=1}^{q} P(M_j) m_j(\boldsymbol{x})} = \left[\sum_{j=1}^{q} \frac{P(M_j)}{P(M_i)} B_{ji} \right]^{-1}$$

Particular case : $P(M_j) = 1/q$:

$$P(M_i \mid \boldsymbol{x}) = \overline{m}_i(\boldsymbol{x}) = \frac{m_i(\mathbf{x})}{\sum_{j=1}^{q} m_j(\boldsymbol{x})} = \frac{1}{\sum_{j=1}^{q} B_{ji}}$$

Reporting : It is useful to separately report $\{\overline{m}_i(\boldsymbol{x})\}$ and $\{P(M_i)\}$, along with

$$P(M_i \mid \boldsymbol{x}) = \frac{P(M_i)\, \overline{m}_i(\boldsymbol{x})}{\sum_{j=1}^{q} P(M_j)\, \overline{m}_j(\boldsymbol{x})} \,.$$

Example: location-scale

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d with density

$$f(x_i \mid \mu, \sigma) = \frac{1}{\sigma}\, g\left(\frac{x_i - \mu}{\sigma}\right) \qquad \text{(loc, scal)}$$

Several models are entertained:

- M_N: g is $N(0,1)$
- M_U: g is Uniform $(0,1)$
- M_C: g is Cauchy $(0,1)$
- M_L: g is Left Exponential $(\frac{1}{\sigma}\, e^{(x-\mu)/\sigma}\ ,\ x \le \mu)$
- M_R: g is Right Exponential $(\frac{1}{\sigma}\, e^{-(x-\mu)/\sigma}\ ,\ x \ge \mu)$

Difficulty: The models are not nested and have no common low dimensional sufficient statistics; classical model selection would thus typically rely on asymptotics.

Prior distribution: choose, for $i = 1, \ldots, 5$,

$P(M_i) = \frac{1}{q} = \frac{1}{5}$;

utilize the objective prior $\pi_i(\boldsymbol{\theta}_i) = \pi_i(\mu, \sigma) = \frac{1}{\sigma}$.

Objective improper priors (invariant) are o.k. for the *common* parameters in situations having the same invariance structure, if the right-Haar prior is used (Berger, Pericchi, and Varshavsky, 1998).

Marginal distributions, $m^N(\mathbf{x} \mid M)$, for these models can then be calculated in closed form.

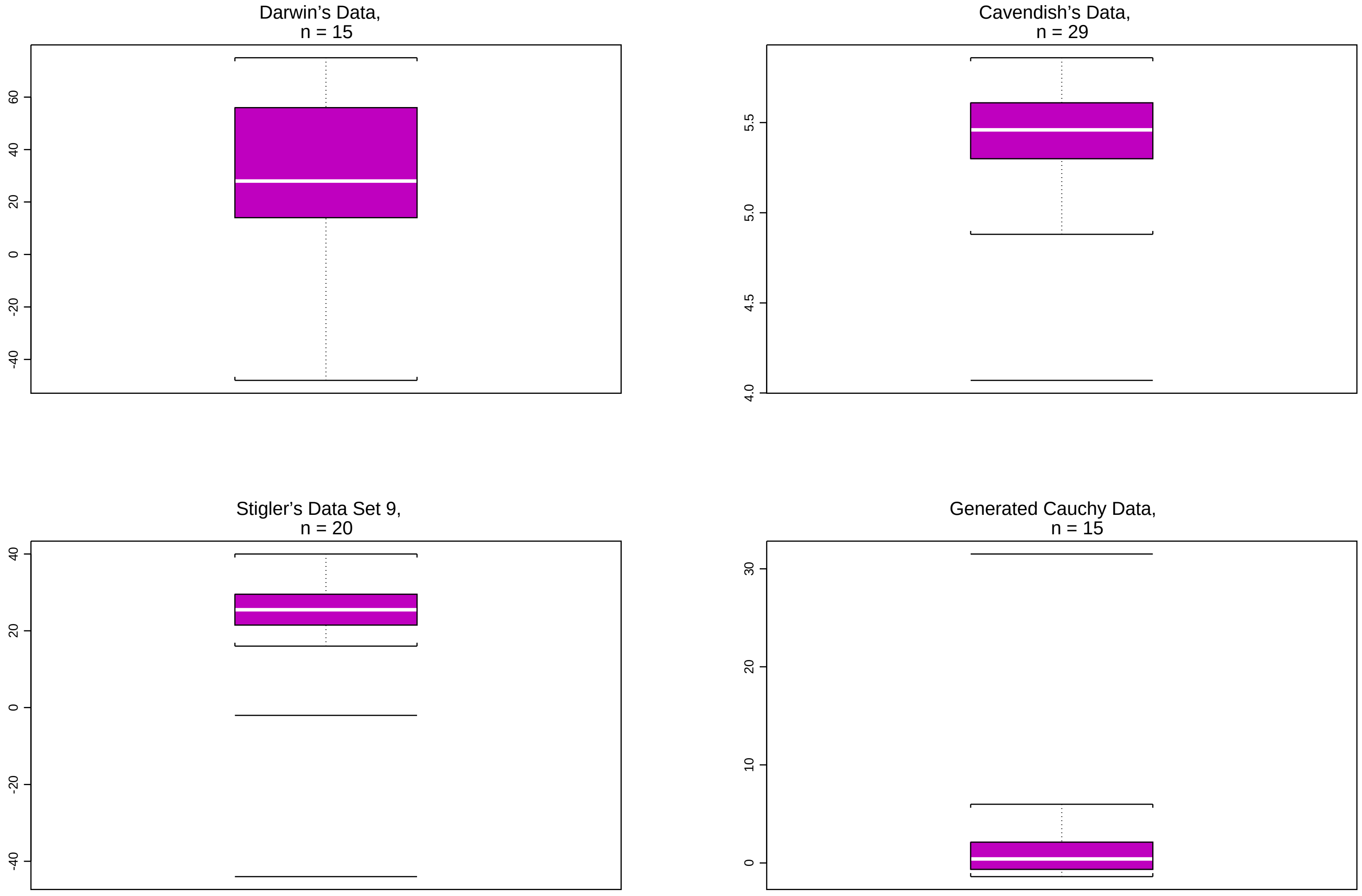
Here $m^N(\mathbf{x} \mid M) = \int \int \prod_{i=1}^n \left[\frac{1}{\sigma} g\left(\frac{x_i - \mu}{\sigma} \right) \right] \frac{1}{\sigma} \quad d\mu \, d\sigma.$

For the five models, the marginals are

1. Normal: $m^N(\mathbf{x} \mid M_N) = \frac{\Gamma((n-1)/2)}{(2\,\pi)^{(n-1)/2} \sqrt{n} \left(\sum_i (x_i - \bar{x})^2 \right)^{(n-1)/2}}$
2. Uniform: $m^N(\mathbf{x} \mid M_U) = \frac{1}{n\,(n-1)(x_{(n)} - x_{(1)})^{\,n-1}}$
3. Cauchy: $m^N(\mathbf{x} \mid M_C)$ is given in Spiegelhalter (1985).
4. Left Exponential: $m^N(\mathbf{x} \mid M_L) = \frac{(n-2)!}{n^n (x_{(n)} - \bar{x})^{\,n-1}}$
5. Right Exponential: $m^N(\mathbf{x} \mid M_R) = \frac{(n-2)!}{n^n (\bar{x} - x_{(1)})^{\,n-1}}$

Consider four classic data sets:

- Darwin's data ($n = 15$)
- Cavendish's data ($n = 29$)
- Stigler's Data Set 9 ($n = 20$)
- A randomly generated Cauchy sample ($n = 14$)


Darwin’s Data,
n = 15
60
40
20
0
-20
-40
Cavendish’s Data,
n = 29
5.5
5.0
4.5
4.0
Stigler’s Data Set 9,
n = 20
40
20
0
-20
-40
Generated Cauchy Data,
n = 15
30
20
10
0

The objective posterior probabilities of the five models, for each data set,

$$\overline{m}^N(\mathbf{x} \mid M) = \frac{m^N(\mathbf{x} \mid M)}{m^N(\mathbf{x} \mid M_N) + m^N(\mathbf{x}|M_U) + m^N(\mathbf{x}|M_C) + m^N(\mathbf{x}|M_L) + m^N(\mathbf{x}|M_R)},$$

are as follows:

DATA	MODELS				
SET	Normal	Uniform	Cauchy	L. Exp.	R. Exp.
Darwin	.390	.056	.430	.124	.0001
Cavendish	.986	.010	.004	4×10^{-8}	.0006
Stigler 9	7×10^{-8}	4×10^{-5}	.994	.006	2×10^{-13}
Cauchy	5×10^{-13}	9×10^{-12}	.9999	7×10^{-18}	1×10^{-4}

III. Model Averaging

Accounting for model uncertainty

- Selecting a single model and using it for inference ignores model uncertainty, resulting in inferior inferences, and considerable overstatements of accuracy.

- The Bayesian approach incorporates this uncertainty by *model averaging*; if, say, inference concerning ξ is desired, it would be based on

  $$\pi(\xi \mid \boldsymbol{x}) = \sum_{i=1}^{q} P(M_i \mid \boldsymbol{x})\ \pi(\xi \mid \boldsymbol{x}\,, M_i).$$

  Note: ξ must have the same meaning across models.

- a useful link, with papers, software, URL's about BMA

  http://www.research.att.com/∼volinsky/bma.html

Example: Vehicle emissions data from McDonald et. al. (1995)

Goal: From i.i.d. vehicle emission data $\mathbf{X} = (X_1, \ldots, X_n)$, one desires to determine the probability that the vehicle type will meet regulatory standards.

Traditional models for this type of data are Weibull and lognormal distributions given, respectively, by

$$\begin{aligned} \mathcal{M}_1 : f_W(x \mid \beta, \gamma) &= \frac{\gamma}{\beta}\left(\frac{x}{\beta}\right)^{\gamma-1} \exp\left[-\left(\frac{x}{\beta}\right)^{\gamma}\right] \\ \mathcal{M}_2 : f_L(x \mid \mu, \sigma^2) &= \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(\log x - \mu)^2}{2\sigma^2}\right]. \end{aligned}$$

Note that both distributions are in the location-scale family after transforming to $y = \log x$.

Model Averaging Analysis:

- Assign each model prior probability 1/2.
- Because of the common location-scale invariance structures, assign the right-Haar prior densities $\pi(\mu, \sigma) = 1/(\sigma)$ to the models after the log-transformation (Berger, Pericchi and Varshavsky, 1998 Sankhyā).
- The posterior probabilities (and conditional frequentist error probabilities) of the two models are then

$$P(\mathcal{M}_1 \mid \boldsymbol{x}) = 1 - P(\mathcal{M}_2 \mid \boldsymbol{x}) = \frac{B(\boldsymbol{x})}{1 + B(\mathbf{x})} \,,$$

$$B(\mathbf{X}) = \frac{\Gamma(n) n^n \pi^{(n-1)/2}}{\Gamma(n - 1/2)} \int_0^\infty \left[\frac{v}{n} \sum_{i=1}^n exp\left(\frac{y_i - \bar{y}}{s_y v} \right) \right]^{-n} dv,$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ and $s_y^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2$.

- For the studied data set, $P(\mathcal{M}_1 \mid \boldsymbol{x}) = .712$. Hence,

$$\begin{aligned} P(\text{meeting standard}) \quad &= .712 \; P(\text{meeting standard} \mid \mathcal{M}_1) \\ &+ .288 \; P(\text{meeting standard} \mid \mathcal{M}_2) \,. \end{aligned}$$

Model averaging most used for prediction:

Goal: Predict a future Y, given X

Optimal prediction is based on model averaging (cf, Draper, 1994). If one of the M_i is indeed the true model (but unknown), the optimal predictor is based on

$$m(y \mid \mathbf{x}) = \sum_{i=1}^{k} P(M_i \mid \mathbf{x}) m_i(y \mid \mathbf{x}, M_i)\,,$$

where

$$m_i(y \mid \mathbf{x}, M_i) = \int f_i(y \mid \theta_i)\pi_i(\theta_i \mid \mathbf{x}) d\theta_i$$

is the predictive distribution when M_i is true.

Alternate expression if $Y = X_{n+1}$ and $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, marginal density of $(\boldsymbol{x}, x_{n+1})$ is

$$m(\boldsymbol{x}, x_{n+1}) = \sum_{i=1}^{q} P(M_i) m_i(\boldsymbol{x}, x_{n+1})$$

Thus

$$\begin{aligned} m(x_{n+1} \mid \boldsymbol{x}) &= \frac{\sum_{i=1}^{q} P(M_i) m_i(\boldsymbol{x}, x_{n+1})}{\int \sum_{i=1}^{q} P(M_i) m_i(\boldsymbol{x}, x_{n+1}) dx_{n+1}} \\ &= \frac{\sum_{i=1}^{q} P(M_i) m_i(\boldsymbol{x}, x_{n+1})}{\sum_{i=1}^{q} P(M_i) m_i(\boldsymbol{x})} \, . \end{aligned}$$

IV. Reasons for Adopting the Bayesian Approach to Model Uncertainty

Reason 1: *Ease of interpretation*

- Bayes factors $\leadsto$ odds

  Posterior model probabilities $\leadsto$ direct interpretation

- In the location example and Darwin's data, the posterior model probabilities are $\overline{m}_N(\boldsymbol{x}) = .390,\ \overline{m}_U(\boldsymbol{x}) = .056,$ $\overline{m}_C(\boldsymbol{x}) = .430,\ \bar{m}_L(\boldsymbol{x}) = .124,\ \overline{m}_R(\boldsymbol{x}) = .0001,$

  Bayes factors are $B_{NU} = 6.96,\ B_{NC} = .91,\ B_{NL} = 3.15,\ B_{NR} = 3900.$

- Interpreting four (or 20) p-values is much harder.

Reason 2: *Prior information can be incorporated, if desired*

- If the default report is $\overline{m}_i(\boldsymbol{x}),\ i = 1, \ldots, q$, then for any prior probabilities
$$P(M_i \mid \boldsymbol{x}) = \frac{P(M_i)\,\overline{m}_i(\boldsymbol{x})}{\sum_{j=1}^{q} P(M_j)\,\overline{m}_j(\boldsymbol{x})}$$

Example: In the Darwin's data, if symmetric models are twice as likely as the non-symmetric, then

$\Pr(M_N) = \Pr(M_U) = \Pr(M_C) = 1/4$ and $\Pr(M_L) = \Pr(M_R) = 1/8$, so that

$\Pr(M_N \mid \boldsymbol{x}) = .416,\ \Pr(M_U \mid \boldsymbol{x}) = .06,\ \Pr(M_C \mid \boldsymbol{x}) = .458,$
$\Pr(M_L \mid \boldsymbol{x}) = .066,\ \Pr(M_R \mid \boldsymbol{x}) = .00005.$

Reason 3: *Consistency*

A. If one of the M_i is true, then $\bar{m}_i \to 1$ as $n \to \infty$

B. If some other model, M^*, is true, $\bar{m}_i \to 1$ for that model closest to M^* in Kullback-Leibler divergence

(Berk, 1966, Dmochowski, 1994)

Reason 4: *Ockham's razor*

Bayes Factors automatically seek parsimony; no adhoc penalties for model complexity are needed. (Illustration and explanation in lecture 10.)

Reason 5: *Frequentist interpretation*

In many important situations involving testing of M_1 versus M_2, $B_{12}/(1+B_{12})$ and $1/(1+B_{12})$ are 'optimal' conditional frequentist error probabilities of Type I and Type II, respectively. Indeed, $Pr(H_0 \mid \text{data } \boldsymbol{x})$ is the *frequentist type I error probability*, conditional on observing data of the same "strength of evidence" as the actual data $\boldsymbol{x}$. (Classical unconditional error probabilities make the mistake of reporting the error averaged over data of very different strengths.)

See Sellke, Bayarri and Berger (2001) for discussion and earlier references.

Reason 6: *The Ability to Account for Model Uncertainty through Model Averaging*

- For non-Bayesians, it is highly problematical to perform inference from a model based on the same data used to select the model.
  - It is not legitimate to do so in the frequentist paradigm.
  - It is very hard to determine how to ‘correct’ the inferences (and known methods are often extremely conservative).
- Bayesian inference based on model averaging overcomes this problem.
- There is some controversy about Bayesian inference from a single selected model.
  - Inference internal to the model – e.g., parameter inference – is okay (though not all agree).
  - External inference, e.g. prediction, is on shakier grounds.

Reason 7: *Generality of application*

- Bayesian approach is viable for *any* number of models, regular or irregular, nested or not, large or small sample sizes.

- Classical approach is most developed for only two models (hypotheses), and typically requires at least one of: *(i)* nested models, *(ii)* standard distributions, *(iii)* regular asymptotics.

- *Example 1:* In the location-scale example, there were 5 non-nested models, with small sample sizes, irregular asymptotics, and no reduced sufficient statistics.

- *Example 2:* Suppose the X_i are i.i.d from the $N(\theta, 1)$ density, where $\theta =$ "mean effect of T1 - mean effect of T2."

  Standard Testing Formulation:

  $H_0 : \theta = 0$ (no difference in treatments) vs. $H_1 : \theta \neq 0$ (a difference exists)

  A More Revealing Formulation:

  $H_0 : \theta = 0$ (no difference) vs. $H_1 : \theta < 0$ (T2 is better) vs. $H_2 : \theta > 0$ (T1 is better)

Reason 8: *Standard Bayesian 'ease of application' motivations*

1. In sequential scenarios, there is no need to 'spend α' for looks at the data; posterior probabilities are not affected by the reason for stopping experimentation.
2. The distributions of known censoring variables are not needed for Bayesian analysis.
3. Multiplicity problems can be addressed in a straightforward manner. (Lecture 4)

V. Challenges with the Bayesian Approach to Model Uncertainty

Difficulty 1: *Inadequacy of common priors, such as (improper) objective priors, vague proper priors, and conjugate priors.*

As in hypothesis testing,

- improper priors can only be used for "common parameters" in models;
- vague proper priors are usually horrible!

The problem with conjugate priors:

Normal Example: Suppose $X_1, \ldots, X_n$ are i.i.d. $N(x \mid \theta, \sigma^2)$.

- The natural conjugate priors are
  $\pi_1(\sigma^2) = 1/\sigma^2$; $\pi_2(\theta \mid \sigma^2) = N(\theta \mid 0, \sigma^2)$ and $\pi_2(\sigma^2) = 1/\sigma^2$.
- *Bayes factor:* $B_{12} = \sqrt{n+1}\left(\frac{1+t^2/(n+1)}{1+t^2}\right)^{n/2}$, where $t = \sqrt{n}\bar{x}/s$.
- *Silly behavior, called 'information inconsistency':*
  as $|t| \to \infty$, $B_{12} \to (n+1)^{-(n-1)/2} > 0$

For instance: if $n = 3$, $B_{12} \to 1/4$

if $n = 5$, $B_{12} \to 1/36$

Difficulty 2: *Meaning of parameters*

Parameters in different models typically have different meanings, so that separate (proper) priors must be constructed for each model

The difficulty applies to both objective priors for model selection *and* subjective priors.

Lindley: “Beware of the fallacy of Greek letters.”

Example:

WRONG WAY:

$$\begin{aligned} M_1 : Y &= \beta_0 + \beta_1 X_1 + \sigma\,\epsilon \\ M_2 : Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \sigma\,\epsilon\,, \qquad \epsilon \sim N(0,1) \end{aligned}$$

Use priors $\pi(\beta_0)$, $\pi(\beta_1)$, $\pi(\beta_2)$, and $\pi(\sigma)$, independent of the model.

RIGHT WAY:

$$\begin{aligned} M_1 : Y &= \beta_0^{(1)} + \beta_1^{(1)} X_1 + \sigma^{(1)}\,\epsilon \\ M_2 : Y &= \beta_0^{(2)} + \beta_1^{(2)} X_1 + \beta_2^{(2)} X_2 + \sigma^{(2)}\,\epsilon\,. \end{aligned}$$

Distinct Priors: $\pi_1(\beta_0^{(1)}, \beta_1^{(1)}, \sigma^{(1)})$ and
$\pi_2(\beta_0^{(2)}, \beta_1^{(2)}, \sigma^{(2)})\pi_2(\beta_2^{(2)} \mid \beta_0^{(2)}, \beta_1^{(2)}, \sigma^{(2)})$.

Example

- Predict fuel consumption Y from weight X_1 and engine size X_2. Clearly β_1 has a very different meaning under M_1 and M_2. (Regressing Y on X_1 alone produces larger β_1 than regressing on both, due to the high correlation between X_1 and X_2).

Difficulty 3: *Utilization of Subjective Priors for Model Uncertainty*

Utilization of subjective priors faces the following challenges:

- There is usually a multitude of models, each having potentially many unknown parameters.
- Parameters in different models that are denoted by the same Greek letter usually have different meanings, so that separate elicitations would need to be done in each model.
- Conjugate priors (for which most elicitation tools have been built) are problematical.

Scenarios in which subjective elicitation can succeed include:

- Orthogonal problems where the variables have meaning independent of the model.
- Problems with extensive variable exchangeability, where hierarchical modeling can be utilized.

We will also see interesting possibilities involving inducing model priors from a single elicited distribution.

Difficulty 4: *Computation*

Computation of the needed marginal distributions (or their ratios) can be very challenging in high dimensions.

The total number of models under consideration can be enormous (e.g., in variable selection), so good search strategies are needed.

See Lecture 8.

Difficulty 5: *Posterior Inference*

When thousands (or millions) of models all have similarly high posterior probability (as is common), how does one summarize the information contained therein?

Model averaging summaries are needed, but summaries of what quantities?

See Lectures 8 and 10.

Difficulty 6: *Evaluating procedures*

- There are are variety of criteria for evaluating approaches to model uncertainty (see Lecture 5), but these are not widely appreciated.
  - For instance, DIC is very widely used, but violates the most basic criterion of model selection, namely consistency.
- Many procedures that are thought of as Bayesian (e.g., DIC) do not behave at all like true Bayesian procedures
  - We call such procedures *pseudo-Bayesian* (and do not discuss them); they require external validation, having no Bayesian guarantees.
  - In lecture 6 – dealing with data-driven methods for deriving priors – the most important criterion for evaluation will be if the procedure behaves, in some sense, like a true Bayesian procedure.
- Simulations are of limited usefulness for true Bayesian procedures, as each will necessarily perform optimally for problems that reflect its Bayesian assumptions.

VI. Approaches to Prior Choice for Model Uncertainty

- Use of ‘conventional priors’ (Lecture 5)
- Use of data-driven priors (Lecture 6)
- Prior from notions of ‘predictive matching’
- Inducing model priors from a single prior.

Priors arising from Predictive Matching

Idea: If $\mathbf{X}^*$ are observables of interest, the pair $\{M_i, \pi_i\}$ would yield the marginal density:

$$m_i(\mathbf{x}^*) = \int f_i(\mathbf{x}^* \mid \theta_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i \,.$$

$\{M_i, \pi_i\}$ is "predictively matched" to $\{M_j, \pi_j\}$ if $m_i(\mathbf{x}^*) \approx m_j(\mathbf{x}^*)$.

Note: $\mathbf{X}^*$ is typically some small set of possible observations.

Approach 1. Predictive moment matching

choose $\pi_i(\boldsymbol{\theta}_i)$ and $\pi_j(\boldsymbol{\theta}_j)$ so $m_i(\mathbf{x}^*)$ and $m_j(\mathbf{x}^*)$ have equal moments (Ibrahim and Laud,93,94,95, others . . .)

Approach 2. Use (improper) objective priors with $m_i(\mathbf{x}^*_\mathbf{0}) = m_j(\mathbf{x}^*_\mathbf{0})$ at some specified point $\mathbf{x}^*_\mathbf{0}$ (Spiegelhalter and Smith, 1982, Ghosh, 1997)

Approach 3. Inducing priors by posterior mixing based on a common marginal ('EP' priors). (Lecture 6.)

Approach 4. Subjective Matching

- Subjectively elicit a predictive $m(\mathbf{x}^*)$, where $\mathbf{x}^*$ is a *minimal training sample* (MTS), typically the dimension of the largest model.
- Goal: for each M_i, find $\pi_i(\boldsymbol{\theta}_i)$ so that

$$m_i^*(\boldsymbol{x}^*) = \int f_i(\boldsymbol{x}^* \mid \boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i$$

is close to $m(\boldsymbol{x}^*)$, $\forall i$.

COOL FACT: Define $\pi_i^{(0)}(\boldsymbol{\theta}_i \mid \mathbf{x}^*) = \pi_i^N(\boldsymbol{\theta}_i \mid \mathbf{x}^*)$ for any objective prior π^N (such that the posterior exists) and

$$\pi_i^{(l)}(\boldsymbol{\theta}_i) = \int \pi_i^{(l-1)}(\boldsymbol{\theta}_i \mid \mathbf{x}^*)\, m(\mathbf{x}^*)d\mathbf{x}^* \,.$$

Then, as $l \to \infty$, $\pi_i^{(l)}$ converges to a π^* such that the resulting $m_i^*(\mathbf{x}^*)$ is closest to $m^*(\mathbf{x}^*)$ in terms of Kullback-Leibler divergence (Shyamalkumar, 1996).

- The first step is the ‘biggest’ and is the EPPrior (*Expected Posterior Prior*) with respect to $m(\boldsymbol{x}^*)$

  (Berger and Pérez, 2002), to be covered later.

Inducing Model Priors from a Single Prior

- Specify a prior $\pi_L(\beta_L)$ for the 'largest' model.
- Use this prior to induce priors on the other models. Possibilities include
  - In variable selection, conditioning by setting variables to be removed to zero. (*Logically sound if variables are orthogonal*)
  - Marginalizing out the variables to be removed. (*Not logically sound*)
  - Matching model parameters for other models with the parameters of the largest model by, say, minimizing Kullback-Leibler divergence between models, and then projecting $\pi_L(\beta_L)$. (*Probably best, but hard*)

Example - Dirichlet: Suppose the largest model has prior $\pi_L(p_1, \ldots, p_m) \sim Dirichlet(1, \ldots, 1)$ (i.e., the uniform distribution on the simplex). If other models have parameters $(p_{i_1}, \ldots, p_{i_l})$,

- conditioning yields $\pi(p_{i_1}, \ldots, p_{i_l}, p^* \mid \text{other}\, p_j = 0) = Dirichlet(1, \ldots, 1)$, where $p^* = 1 - \sum_{j=1}^{l} p_{i_j}$;
- marginalizing yields $\pi(p_{i_1}, \ldots, p_{i_l}, p^*) = Dirichlet(1, \ldots, 1, m - l)$ (too concentrated at zero for $(p_{i_1}, \ldots, p_{i_l})$).

Example. Posterior Model Probabilities for Variable Selection in Probit Regression

(from CMU Case Studies VII, Viele et. al. case study)

Motivating example: Prosopagnosia (face blindness), is a condition (usually developing after brain trauma) under which the individual cannot easily distinguish between faces. A psychological study was conducted by Tarr, Behrmann and Gauthier to address whether this extended to a difficulty of distinguishing other objects, or was particular to faces.

The study considered 30 control subjects (C) and 2 subjects (S) diagnosed as having prosopagnosia, and had them try to differentiate between similar faces, similar 'Greebles' and similar 'objects' at varying levels of difficulty and varying comparison time.

Data:

$$\left.\begin{array}{l} C = Subject\ C \\ S = Subject\ S \\ G = Greebles \\ O = Object \\ D = Difficulty \\ B = Brief\ time \\ A = Images\ match\ or\ not \end{array}\right\} \Rightarrow \ \ R = Response\ (answer\ correct\ or\ not).$$

All variables are binary. Sample size was $n = 20,083$. {C=S=1} and {G=O=1} are not possible combinations, so there are $3 \times 3 \times 2 \times 2 \times 2 = 72$ possible covariates.

Statistical modeling: For a specified covariate vector $\boldsymbol{X}_i$, let y_i and $n_i - y_i$ be the numbers of successes and failures among the responses with that covariate vector, with probability of success p_i assumed to follow the probit regression model

$$p_i = \Phi(\beta_1 + \textstyle\sum_{j=2}^{72} \boldsymbol{X}_{ij}\beta_j).$$

The full model likelihood (up to a fixed proportionality constant) is then

$$f(\boldsymbol{y} \mid \boldsymbol{\beta}) = \textstyle\prod_{i=1}^{72} p_i^{y_i}(1-p_i)^{n_i - y_i}.$$

Goal: Select from among the 2^{72} submodels which have some of the β_j set equal to zero. (Actually, only models with graphical structure were considered, i.e., if an interaction term is in the model, all the lower order effects must also be there.)

Prior Choice: *Conditionally inducing submodel priors from a full-model prior*

- A standard noninformative prior for $\boldsymbol{p} = (p_1, p_2, \ldots, p_{72})$ is the uniform prior, usable here since it is proper.
- Change of variables yields $\pi_L(\boldsymbol{\beta})$ is $\mathcal{N}_{72}(0, (\boldsymbol{X}'\boldsymbol{X})^{-1})$, where $\boldsymbol{X}' = (\boldsymbol{X}'_1, \boldsymbol{X}'_2, \ldots, \boldsymbol{X}'_{72})$.
- Then $\pi_j(\boldsymbol{\beta}_{(j)} \mid \boldsymbol{\beta}_{(-j)} = 0)$ is $\mathcal{N}_{k_j}(0, (\boldsymbol{X}'_{(j)}\boldsymbol{X}_{(j)})^{-1})$, where $\boldsymbol{\beta}_{(j)}$ is a subvector of parameters of dimension k_j and $\boldsymbol{X}_{(j)}$ is the corresponding design matrix.

VII. Selecting a Single Model for Prediction

Context: Prediction with Normal linear models

- Observe the $n \times 1$ vector

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\,;$$

where $\mathbf{X}$ is the $n \times k$ design matrix, $\boldsymbol{\beta}$ is the $k \times 1$ vector of unknown coefficients, and $\boldsymbol{\epsilon}$ is $\mathcal{N}(0, \sigma^2 I)$.

- Choose from among submodels

$$M_{\mathbf{l}} : \boldsymbol{y} = \mathbf{X}_{\mathbf{l}}\,\boldsymbol{\beta}_{\mathbf{l}} + \boldsymbol{\epsilon}\,,$$

where $\mathbf{l} = (l_1, l_2, \ldots, l_k)$ is the model index, l_i being either 1 or 0 as covariate x_i is in or out of the model.

Basics of Bayesian prediction

- The goal is to predict a future $y^* = \mathbf{x}^* \boldsymbol{\beta} + \epsilon$, using squared error loss $(y^* - \widehat{y}^*)^2$.
- Combining the data and prior yields, for all $\mathbf{l}$,
  - $P(M_{\mathbf{l}} \mid \boldsymbol{y})$, the posterior probability of model $M_{\mathbf{l}}$;
  - $\pi^*_{\mathbf{l}}(\boldsymbol{\beta}_{\mathbf{l}}, \sigma \mid \boldsymbol{y})$, the posterior distribution of $(\boldsymbol{\beta}_{\mathbf{l}}, \sigma)$.
- The best predictor of y^* is, via *model averaging*,

$$\bar{\boldsymbol{y}}^* = \mathbf{x}^* \, \bar{\boldsymbol{\beta}} \equiv \mathbf{x}^* \sum_{\mathbf{l}} P(M_{\mathbf{l}} \mid \boldsymbol{y}) \, \widetilde{\boldsymbol{\beta}}_{\mathbf{l}} \, ,$$

where $\widetilde{\boldsymbol{\beta}}_{\mathbf{l}}$ is the posterior mean for $\boldsymbol{\beta}$ under $M_{\mathbf{l}}$.

Selecting a single model

- Often a single model is desired for prediction.
- A common misperception is that the best single model is that with the largest $P(M_{\mathbf{l}} \mid \boldsymbol{y})$;
  - however, this is true if there are only two models;
  - and it is true if $\mathbf{X}^{'}\mathbf{X}$ is diagonal, σ^2 is known, and suitable priors are used (Clyde and Parmigiani, 1996).
- The best single model will typically depend on $\mathbf{x}^*$.
- An important case is when the future covariates are like the past covariates, i.e., when $E[(\mathbf{x}^*)'\mathbf{x}^*] = \mathbf{X}^{'}\mathbf{X}$.

Posterior inclusion probabilities

The *posterior inclusion probability* for variable i is

$$p_i \equiv \sum_{\boldsymbol{l}\,:\,l_i=1} P(M_{\boldsymbol{l}} \,|\, \boldsymbol{y}),$$

i.e., the overall posterior probability that variable i is in the model.

These are of considerable independent interest

- as basic quantities of interest,
- as aids in searches of model space,
- in defining the median probability model.

The (posterior) median probability model

If it exists, the *median probability model*, $M_{\boldsymbol{l}^*}$, is defined to be the model consisting of those variables whose posterior inclusion probability is at least 1/2. Formally, $\boldsymbol{l}^*$ is defined, coordinatewise, by

$$l_i^* = \begin{cases} 1 & \text{if} \quad p_i \geq \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Note: If computation is done by a model-jumping MCMC, the median probability model consists of those coordinates that were present in over half the iterations.

Existence of the median probability model

The median probability model exists when the models under consideration follow a *graphical model structure*, including

- when any subset of variables is allowed;
- the situation in which the allowed variables consist of main effects and interactions, but a higher order interaction is allowed only if lower order interactions are included;
- a sequence of nested models, such as arises in polynomial regression and autoregressive time series.

Example (Polynomial Regression): Model j is

$$y = \sum_{i=0}^{j} \beta_i \, x^i + \epsilon.$$

(Model) j	0	1	2	3	4	5	6
$P(M_j \mid \boldsymbol{y})$	~ 0	0.06	0.22	0.29	0.38	0.05	~ 0

(Covariate) i	0	1	2	3	4	5	6
$P(x^i$ is in model $\mid \boldsymbol{y})$	1	1	0.94	0.72	0.43	0.05	0

Thus M_3 is the median probability (optimal predictive) model, while M_4 is the maximum probability model.

Three optimality theorems

Theorem 1. If (i) the models under consideration have graphical structure; (ii) $\mathbf{X}^{'}\mathbf{X}$ is diagonal, and (iii) the posterior mean of $\boldsymbol{\beta_l}$ is simply the relevant coordinates of $\widetilde{\boldsymbol{\beta}}$ (the posterior mean in the full model), then the best predictive model is the median probability model. Condition (iii) is satisfied under any mix of

- noninformative priors for the β_i;
- independent $\mathcal{N}(0, \sigma^2\lambda_i)$ priors for the β_i, with the λ_i given (objectively or subjectively specified, or estimated via empirical Bayes) and any prior for σ^2.

Corollary (Clyde and Parmigiani, 1996). If any submodel of the full model is allowed, $\mathbf{X}^{'}\mathbf{X}$ is diagonal, $\mathcal{N}(0, \sigma^2\lambda_i)$ priors are used for the β_i, with the λ_i given and σ^2 known, and the prior probabilities of the models satisfy

$$P(M_{\boldsymbol{l}}) = \prod_{i=1}^{k} (p_i^0)^{l_i} (1 - p_i^0)^{(1-l_i)} \, ,$$

where p_i^0 is the prior probability that variable x_i is in the model, then the optimal predictive model is that with highest posterior probability (which is also the median probability model).

Theorem 2. Suppose a sequence of nested linear models is under consideration. If (i) prediction is desired at 'future covariates like the past' and (ii) the posterior mean under $M_{\boldsymbol{l}}$ satisfies $\widetilde{\boldsymbol{\beta}}_{\boldsymbol{l}} = b\,\widehat{\boldsymbol{\beta}}_{\boldsymbol{l}}$, where $\widehat{\boldsymbol{\beta}}_{\boldsymbol{l}}$ is the least squares estimate, then the best predictive model is the median probability model.
Condition (ii) is satisfied if we use either

- noninformative priors for model parameters; or
- g-type $\mathcal{N}_{k_{\boldsymbol{l}}}(\mathbf{0}, c\,\sigma^2\,(\mathbf{X}_{\boldsymbol{l}}'\mathbf{X}_{\boldsymbol{l}})^{-1})$ priors, with the same constant $c > 0$ for each model, and any prior for σ^2.

Theorem 3. Theorems 1 and 2 essentially remain true even if there are non-orthogonal nuisance parameters (i.e., parameters common to all models) that are assigned the usual noninformative priors.

Nonparametric Regression

- $y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n, \quad \epsilon_i \sim N(0, \sigma^2).$
- Represent f via a (orthonormal) series expansion

$$f(x) = \sum_{j=1}^{\infty} \beta_j \, \phi_j(x).$$

- *Base* prior distribution: $\beta_i \sim N(0, v_i)$, with $v_i = \frac{c}{i^a}$, where c is unknown and a is specified.
- The model M_j, for $j = 1, 2, \ldots, n$, is given by:

$$y_i = \sum_{k=1}^{j} \beta_k \, \phi_k(x_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2).$$

- Choose equal prior probabilities for the models $M_j, \;\; j = 1, 2, \ldots n$. Within M_j, use the base prior to induce the prior distributions for $\boldsymbol{\beta}_j = (\beta_1, \ldots, \beta_j)$.

- For the data $\boldsymbol{y} = (y_1, \ldots, y_n)$, compute $P(M_j \mid \boldsymbol{y})$, the posterior probability of model M_j, for $j \leq n$.
- Within M_j, predict $\boldsymbol{\beta}_j$ by its posterior mean, $\tilde{\boldsymbol{\beta}}_j$.

Example 1. The Shibata Example

- $f(x) = -\log(1-x)$ for $-1 < x < 1$.
- Choose $\{\phi_1(x),\, \phi_2(x), \ldots\}$ to be the Chebyshev polynomials.
- Then $\beta_i = 2/i$, so the 'optimal' choice of the prior variances would be $v_i = 4/i^2$, i.e., $c = 4$ and $A = 2$.
- Measure the predictive capability of a model by expected squared error loss relative to the true function (here known) – thus we use a frequentist evaluation, as did Shibata.

	MaxPr	MedPr	ModAv	BIC	AIC
$a = 1$	0.99 [8]	0.89 [10]	0.84	1.14 [8]	1.09 [7]
$a = 2$	0.88 [10]	0.80 [16]	0.81	1.14 [8]	1.09 [7]
$a = 3$	0.88 [9]	0.84 [17]	0.85	1.14 [8]	1.09 [7]

Table 1: For $n = 30$ and $\sigma^2 = 1$, the expected loss and average model size for the maximum probability model (MaxPr), the Median Probability Model (MedPr), Model Averaging (ModAv), and BIC and AIC, in the Shibata example.

	MaxPr	MedPr	ModAv	BIC	AIC
$a = 1$	0.54 [14]	0.51 [19]	0.47	0.59 [11]	0.59 [13]
$a = 2$	0.47 [23]	0.43 [43]	0.44	0.59 [11]	0.59 [13]
$a = 3$	0.47 [22]	0.46 [45]	0.46	0.59 [11]	0.59 [13]

Table 2: For $n = 100$ and $\sigma^2 = 1$, the expected loss and average model size for the maximum probability model (MaxPr), the Median Probability Model (MedPr), Model Averaging (ModAv), and BIC and AIC, in the Shibata example.

	MaxPr	MedPr	ModAv	BIC	AIC
$a = 1$	0.34 [23]	0.33 [26]	0.30	0.41 [12]	0.38 [21]
$a = 2$	0.26 [42]	0.25 [51]	0.25	0.41 [12]	0.38 [21]
$a = 3$	0.29 [38]	0.29 [50]	0.29	0.41 [12]	0.38 [21]

Table 3: For $n = 2000$ and $\sigma^2 = 3$, the expected loss and average model size for the maximum probability model (MaxPr), the Median Probability Model (MedPr), Model Averaging (ModAv), and BIC and AIC, in the Shibata example.

Comments

- AIC is better than BIC (as Shibata showed), but the true Bayesian procedures are best.
- Model averaging is generally best (not obvious), followed closely by the median probability model. The maximum probability model can be considerably inferior.
- BIC is a very poor approximation to the Bayesian answers.
- The true Bayesian answers choose substantially *larger* models than AIC (and then shrink towards 0).

ANOVA

Many ANOVA problems, when written in linear model form, yield diagonal $\mathbf{X}'\mathbf{X}$ and any such problems will naturally fit under our theory. In particular, this is true for any balanced ANOVA in which each factor has only two levels. As an example, consider the full two-way ANOVA model with interactions:

$$y_{ijk} = \mu + a_i + b_j + ab_{ij} + \epsilon_{ijk}$$

with $i = 1, 2$, $j = 1, 2$, $k = 1, 2, \ldots, K$ and ϵ_{ijk} independent $N(0, \sigma^2)$, with σ^2 unknown. In linear model form, this leads to $\mathbf{X}'\mathbf{X} = 4K\mathbf{I}_4$.

Possible modeling scenarios

We use the simplified notation M_{1011} instead of $M_{(1,0,1,1)}$, representing the model having all parameters except a_1.

Scenario 1 - *All models with the constant μ:* Thus the set of models under consideration is $\{M_{1000}, M_{1100}, M_{1010}, M_{1001}, M_{1101}, M_{1011}, M_{1110}, M_{1111}\}$.

Scenario 2 - *Interactions present only with main effects, and μ included:* The set of models under consideration here is $\{M_{1000}, M_{1100}, M_{1010}, M_{1110}, M_{1111}\}$. Note that this set of models has graphical structure.

Scenario 3 - *An analogue of an unusual classical test:* In classical ANOVA testing, it is sometimes argued that one might be interested in testing for no interaction effect followed by testing for the main effects, even if the no-interaction test rejected. The four models that are under consideration in this process, including the constant μ in all, are $\{M_{1101}, M_{1011}, M_{1110}, M_{1111}\}$. This class of models does *not* have graphical model structure and yet the median probability model is guaranteed to be in the class.

Example 2. Montgomery (1991, pp.271–274) considers the effects of the concentration of a reactant and the amount of a catalyst on the yield in a chemical process. The reactant concentration is factor A and has two levels, 15% and 25%. The catalyst is factor B, with the two levels 'one bag' and 'two bags' of catalyst. The experiment was replicated three times and the data are

treatment	replicates		
combination	I	II	III
A low, B low	28	25	27
A high, B low	36	32	32
A low, B high	18	19	23
A high, B high	31	30	29

For each modeling scenario, two Bayesian analyses were carried out, both satisfying the earlier conditions so that the median probability model is known to be the optimal predictive model.

- I. The reference prior $\pi(\mu, \sigma) \propto \frac{1}{\sigma}$ was used for the common parameters, while the standard $N(0, \sigma^2)$ g-prior was used for a_1, b_1 and ab_{11}. In each scenario, the models under consideration were given equal prior probabilities of being true.

- II. The g-prior was also used for the common μ.

model	posterior probability	posterior expected loss
M_{1000}	0.0009	237.21
M_{1100}	0.0347	60.33
M_{1010}	0.0009	177.85
M_{1110}	0.6103	0.97
M_{1111}	0.3532	3.05

Table 4: Scenario 2 – graphical models, prior I. The posterior inclusion probabilities are $p_2 = 0.9982$, $p_3 = 0.9644$, and $p_4 = 0.3532$; thus M_{1110} is the median probability model.

model	posterior probability	posterior expected loss
M_{1011}	0.124	143.03
M_{1101}	0.286	36.78
M_{1110}	0.456	10.03
M_{1111}	0.134	9.41

Table 5: Scenario 3 – unusual classical models, prior II. The posterior inclusion probabilities are $p_2 = 0.876$, $p_3 = 0.714$, and $p_4 = 0.544$; thus M_{1111} is the median probability model.

When does the median probability model fail? (Merlise Clyde)

- Suppose
  - under consideration are the model with only a constant term, and the models with a constant term and a single covariate x_i, $i = 1, \ldots, k$, with $k \geq 3$;
  - the models have equal prior probability of $\frac{1}{k+1}$;
  - all covariates are nearly perfectly correlated, with each other and with y.
- Then
  - the posterior probability of the constant model will be near zero, and that of each of the other models will be approximately $1/k$;
  - thus the posterior inclusion probabilities will also be approximately $1/k < 1/2$;
  - so the median probability model is the constant model, which will have poor predictive performance compared to any other model.

A high correlation example where the theory does not apply

Example: Consider Hald's regression data (Draper and Smith, 1981), consisting of $n = 13$ observations on a dependent variable y, with four potential regressors: x_1, x_2, x_3, x_4. The full model is thus

$$y = c + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2),$$

with σ^2 unknown.

- All models that include the constant term are considered. This example does not formally satisfy the theory, since the models are not nested and the conditions of Theorem 3 do not apply.
- Least squares estimates are used for parameters.
- Default posterior model probabilities, $P(M_{\boldsymbol{l}}|\boldsymbol{y})$, are computed using the Encompassing Arithmetic Intrinsic Bayes Factor (Berger and Pericchi, 1996), together with equal prior model probabilities.
- Predictive risks, $R(M_{\boldsymbol{l}})$, are computed.

Model	$P(M_{\boldsymbol{l}}\|\boldsymbol{y})$	$R(M_{\boldsymbol{l}})$
c	0.000003	2652.44
c,1	0.000012	1207.04
c,2	0.000026	854.85
c,3	0.000002	1864.41
c,4	0.000058	838.20
c,1,2	0.275484	8.19
c,1,3	0.000006	1174.14
c,1,4	0.107798	29.73

Model	$P(M_{\boldsymbol{l}}\|\boldsymbol{y})$	$R(M_{\boldsymbol{l}})$
c,2,3	0.000229	353.72
c,2,4	0.000018	821.15
c,3,4	0.003785	118.59
c,1,2,3	0.170990	1.21
c,1,2,4	0.190720	0.18
c,1,3,4	0.159959	1.71
c,2,3,4	0.041323	20.42
c,1,2,3,4	0.049587	0.47

- The posterior inclusion probabilities are

$$p_1 = \sum_{\boldsymbol{l}:l_1=1} P(M_{\boldsymbol{l}}|\boldsymbol{y}) = 0.95, \;\; p_2 = \sum_{\boldsymbol{l}:l_2=1} P(M_{\boldsymbol{l}}|\boldsymbol{y}) = 0.73$$

$$p_3 = \sum_{\boldsymbol{l}:l_3=1} P(M_{\boldsymbol{l}}|\boldsymbol{y}) = 0.43, \;\; p_4 = \sum_{\boldsymbol{l}:l_4=1} P(M_{\boldsymbol{l}}|\boldsymbol{y}) = 0.55.$$

- Thus the median probability model is $\{c, x_1, x_2, x_4\}$ which clearly coincides with the optimal predictive model.

- Note that the risk of the maximum probability model $\{c, x_1, x_2\}$ is considerably higher than that of the median probability model.

Conclusions

- The (posterior) median probability model will typically be the optimal predictive model.
- The median probability model is typically easy to compute, requiring only rough estimates of the posterior inclusion probabilities.
- The posterior inclusion probabilities are themselves quantities of basic interest in model selection and searches of model space.