A brief overview of Bayesian Model Averaging

Chris Sroka, Juhee Lee, Prasenjit Kapat, Xiuyun Zhang

Department of Statistics
The Ohio State University

Model Selection, Stat 882 AU 2006, Dec 6.

Jennifer A. Hoeting, David Madigan, Adrian E. Raftery and Chris T. Volinsky.
Bayesian Model Averaging: A Tutorial
Statistical Science, Vol. 14, No. 4. (Nov., 1999), pp. 382-401.[1]

[1] www.stat.washington.edu/www/research/online/hoeting1999.pdf

Part I

Christopher Sroka

Where are we?

Introduction: A Motivating Example

- Data concerning cancer of the esophagus
- Demographic, medical covariates
- Response is survival status
- Goal: predict survival time for future patients and plan interventions
- Standard statistical practice
  - Use data-driven search to find best model M^*
  - Check model fit
  - Use M^* to make estimate effects, make predictions

Introduction: A Motivating Example

- Unsatisfactory approach
- What do you do about competing model M^{**}?
- Too risky to base all of your inferences on M^{*} alone
- Inferences should reflect ambiguity about the model
- Solution: Bayesian model averaging (BMA)

Introduction: Notation

- Δ is quantity of interest
  - Effect size
  - Future observation
  - Utility of a course of action
- D is data
- $\mathcal{M} = \{M_k, k = 1, 2, ..., K\}$
- θ_k is vector of parameters in model M_k
- $\Pr(\theta_k|M_k)$ is prior density of θ_k underM_k
- $\Pr(D|\theta_k, M_k)$ is likelihood of data
- $\Pr(M_k)$ is prior probability that M_k is the true model

Introduction: Mathematical development

- Posterior distribution given data D is

$$\Pr(\Delta|D) = \sum_{k=1}^{K} \Pr(\Delta|M_k, D) \Pr(M_k|D)$$

- This is average of posterior distributions under each model considered, weighted by posterior model probability
- Posterior probability for model $M_k \in \mathcal{M}$ is

$$\Pr(M_k|D) = \frac{\Pr(D|M_k)\Pr(M_k)}{\sum_{l=1}^{K} \Pr(D|M_l)\Pr(M_l)}$$

where

$$\Pr(D|M_k) = \int \Pr(D|\theta_k, M_k) \Pr(\theta_k|M_k) d\theta_k$$

Introduction: Mathematical development

- Let $\hat{\Delta}_k = E[\Delta|D, M_k]$
- Posterior mean and variance of Δ:

$$
\begin{aligned}
E[\Delta|D] &= \int \Delta \left(\sum_{k=1}^{K} \Pr(\Delta|M_k, D) \Pr(M_k|D) \right) d\Delta \\
&= \sum_{k=1}^{K} \left(\int \Delta \Pr(\Delta|M_k, D) d\Delta \right) \Pr(M_k|D) \\
&= \sum_{k=1}^{K} \hat{\Delta}_k \Pr(M_k|D) \\
Var[\Delta|D] &= \sum_{k=1}^{K} (Var[\Delta|D, M_k] + \hat{\Delta}_k^2) \Pr(M_k|D) - E[\Delta|D]^2
\end{aligned}
$$

Introduction: Complications

- Previous research shows that averaging over *all* models provides better predictive ability than using single model
- Difficulties in implementation
  1. $\mathcal{M}$ can be enormous; infeasible to sum over all models
  2. Integrals can be hard to compute, even using MCMC methods
  3. How do you specify prior distribution on M_k?
  4. How to determine class $\mathcal{M}$ to average over?

Where are we?

Combining Models: Historical Perspective

- 1963: First mention of model combination
- 1965: Distribution to combine opinions of two experts
- 1969: Use of model combination for economic forecasting
- 1970s: Flurry of work in economics literature combining predictions from different forecasting models
- 1978: Basic paradigm for BMA, accounting for model uncertainty
- 1990s: Computational power and theoretical advances overcome difficulties of using BMA

Where are we?

Implementation: Managing the Summation

- Need practical way to compute the sum

$$\Pr(\Delta|D) = \sum_{k=1}^{K} \Pr(\Delta|M_k, D) \Pr(M_k|D)$$

- Approaches:
  1. Occam's window
  2. Markov chain Monte Carlo model composition (MC^3)

Implementation: Occam's Window

- Average over a subset of models supported by the data
- Principle 1: Disregard a model if it predicts the data far less well than model with best predictions
- Formally:

$$\mathcal{A}' = \left\{ M_k : \frac{\max_l \{\Pr(M_l|D)\}}{\Pr(M_k|D)} \leq C \right\}$$

Implementation: Occam's Window

- Exclude complex models if data support simpler models (Occam's razor)
- Formally:

$$\mathcal{B} = \left\{ M_k : \exists M_l \in \mathcal{A}', M_l \subset M_k, \frac{\Pr(M_l|D)}{\Pr(M_k|D)} > 1 \right\}$$

- Our subset of model to average over is $\mathcal{A} = \mathcal{A}' \setminus \mathcal{B}$

$$\Pr(\Delta|D) = \sum_{M_k \in \mathcal{A}} \Pr(\Delta|M_k, D) \Pr(M_k|D)$$

- All probabilities conditional on the set $\mathcal{A}$

Implementation: Occam's Window

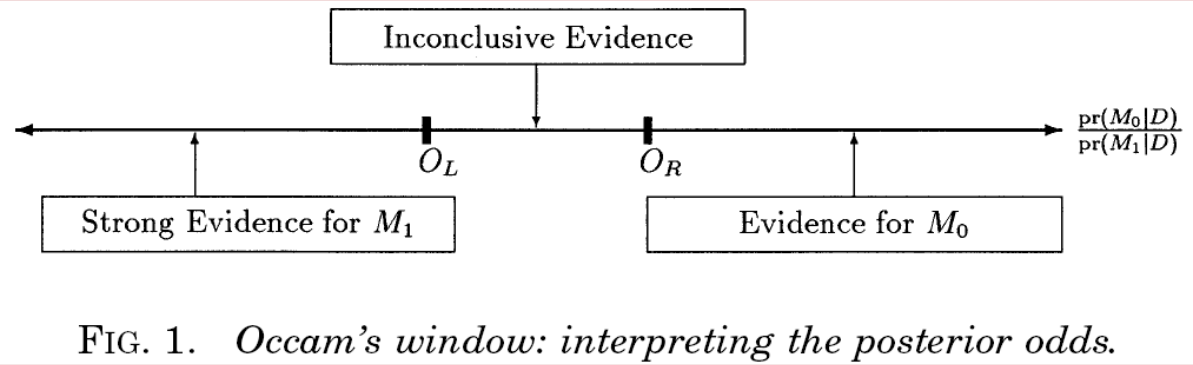


FIG. 1. *Occam's window: interpreting the posterior odds.*

- M_0 is smaller model
- Use $O_L = 1/20$, $O_R = 1$
- Need overwhelming evidence to accept larger model over smaller
- If $O_L = 1/20$, $O_R = 20$, using only first principle

Implementation: MC^3

- Use MCMC to directly approximate

$$\Pr(\Delta|D) = \sum_{k=1}^{K} \Pr(\Delta|M_k, D) \Pr(M_k|D)$$

- Construct a Markov chain $\{M(t)\}, t = 1, 2, ...$ with state space $\mathcal{M}$ and equilibrium distribution $\Pr(M_i|D)$
- Simulate chain to get observations $M(1), ..., M(N)$
- Then for any function $g(M_i)$ defined on $\mathcal{M}$, compute average

$$\hat{G} = \frac{1}{N} \sum_{t=1}^{N} g(M(t))$$

Implementation: MC^3

- Applying standard MCMC results,

$$\hat{G} \xrightarrow{a.s.} E(g(M)) \text{ as } N \to \infty$$

- For this application, set $g(M) = \Pr(\Delta|M, D)$
- Construct chain using Metropolis-Hastings, with transition probability

$$\min\left\{1, \frac{\Pr(M'|D)}{\Pr(M|D)}\right\}$$

Implementation: Computing integrals

- Integrals of the form

$$\Pr(D|M_k) = \int \Pr(D|\theta_k, M_k) \Pr(\theta_k|M_k) d\theta_k$$

  can be hard to compute
- Closed form integrals available for multiple regression and graphical models
- Laplace method (see literature) helps approximate $\Pr(D|M_k)$ and sometimes yields BIC approximation
- Approximate $\Pr(\Delta|M_k, D)$ with $\Pr(\Delta|M_k, \hat{\theta}, D)$, where $\hat{\theta}$ is MLE
- Some of these approximations discussed later

Part II

Xiuyun Zhang

Where are we?

4. Implementation Details for specific model classes
   - Linear Regression
   - GLM
   - Survival Analysis
   - Graphical Models
   - Softwares

Linear Regressions: Predictors, Outliers and Transformations

- Suppose dependent variable Y and predictors $X_1, \ldots, X_k$. Then variable selection methods try to find the "best" model with the form

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_{i_j} X_{i_j} + \varepsilon$$

- However, BMA tries to average over all possible sets of predictors. Raftery, Madigan and Hoeting (1997), and Fernandez, Ley and Steel (1997,1998) did lots of work on this.
- Hoeting, Raftery and Madigan (1996 and 1999) made extension to transformations and outliers.

Linear Regressions: Predictors, Outliers and Transformations (cont'd)

- HRM99 used the Box-Cox transformation for the response:

$$y^{(\rho)} = \begin{cases} \frac{y^{\rho}-1}{\rho} & \rho \neq 0 \\ log\,(y) & \rho = 0 \end{cases}$$

  And the model is $Y^{(\rho)} = X\beta + \varepsilon$ where $\varepsilon \sim N\left(0, \sigma^2 I\right)$
- HRM99 used "change point transformations" to transform the predictors:
  - Use the output from the alternating conditional expectation algorithm (ACE) to suggest the form of transformation.
  - Use Bayes factors to choose the precise transformation.

Linear Regressions: Predictors, Outliers and Transformations (cont'd)

- HRM96 averaged over sets of predictors and possible outliers. They used a variance-inflation model for outliers by assuming:

$$\varepsilon = \begin{cases} N\left(0, \sigma^2\right) & \text{w.p. } (1-\pi) \\ N\left(0, K^2\sigma^2\right) & \text{w.p. } \pi \end{cases}$$

- Simultaneous variable and outlier selection (SVO) method:
  - Use a highly robust technique to identify potential outliers.
  - Compute all possible posterior model probabilities or use MC^3, considering all possible subsets of potential outliers.
- SVO successfully identifies masked outliers.

Generalized Linear Models

- The Bayes factor for model M_1 against M_0:

$$B_{10} = pr(D|M_1)/pr(D|M_0)$$

- Consider $(M+1)$ models $M_0, M_1, \ldots, M_k$. Then the posterior probability of M_k is:

$$pr(M_k|D) = \alpha_k B_{k0} / \sum_{r=0}^{K} \alpha_r B_{r0}$$

where $\alpha_k = pr(M_k)/pr(M_0)$, $k = 0, \cdots, K$.

- Dependent variable: Y_i
  Independent variables: $X_i = (x_{i1}, \ldots, x_{ip})$, $i = 0, \ldots, n$
  where $x_{i1} = 1$
- The null model M_0 is defined by setting $\beta_j = 0 (j = 2, \ldots, p)$.

Generalized Linear Models(cont'd)

- Raftery (1996) used Laplace approximation:

$$pr\left(D|\ M_k\right) \approx (2\pi)^{p_k/2} \left|\Psi\right|^{1/2}\ pr\left(D|\ \hat{\beta}_k,\ M_k\right) pr\left(\hat{\beta}_k|\ M_k\right)$$

where p_k is the dimension of β_k, $\tilde{\beta}_k$ is the posterior mode of β_k and Ψ_k is minus the inverse Hessian of
$h\left(\beta_k\right) = log\{pr\left(D|\beta_k,\ M_k\right) pr\left(\beta_k|\ M_k\right)\}$ evaluated at $\beta_k = \tilde{\beta}_k$

Generalized Linear Models(cont'd)

- Suppose $E(\beta_k|\ M_k) = w_k$ and $var(\beta_k|\ M_k) = W_k$.
- Use one step of Newton's method to approximate $\tilde{\beta}_k$ starting from $\hat{\beta}$. Then we have the approximation

$$2\ log\ B_{10} \approx \chi^2 + (E_1 - E_0)$$

  - $\chi^2 = 2\{\ell_1\left(\hat{\beta}_1\right) - \ell_0\left(\hat{\beta}_0\right)\}$
  - $\ell_k(\beta_k) = log\{pr(D|\beta_k,\ M_k)\}$
  - $E_k = 2\lambda_k\left(\hat{\beta}_k\right) + \lambda_k'\left(\hat{\beta}_k\right)^T (F_k + G_k)^{-1}$
    $\cdot\{2 - F_k(F_k + G_k)^{-1}\}\lambda_k'\left(\hat{\beta}_k\right)$
    $-log|F_k + G_k| + p_k\ log(2\pi)$

    where F_k is the Fisher information matrix, $G_k = W_k^{-1}$, and $\lambda_k(\beta_k) = log\ pr(\beta_k|\ M_k)$

Survival Analysis

- Hazard rate: $\lambda(t) = f(t)/(1-F(t))$
- Cox proportional hazard model: $\lambda(t|X_i) = \lambda_0(t)\exp(X_i\beta)$ where $\lambda_0(t)$ is the baseline hazard rate at time t.
- The estimation of β is based on the partial likelihood:

$$PL(\beta) = \prod_{i=1}^{n}\left(\frac{exp(X_i\beta)}{\sum_{\ell\in R_i} exp(X_\ell^T\beta)}\right)^{w_i}$$

where R_i is the risk set at time t_i and w_i is an indicator for whether or not subject i is censored.

Survival Analysis (cont'd)

- Volinsky, Madigan, Raftery and Kronmal (1997) (VMRK) adopted the MLE approximation:

$$pr\left(\Delta|M_k, D\right) \approx pr\left(\Delta|M_k, \hat{\beta}_k, D\right)$$

and the Laplace approximation:

$$log\ pr\left(D|M_k\right) \approx log\ pr\left(D|M_k, \hat{\beta}_k\right) - d_k\ log\ n$$

where d_k is the dimension of β_k.

Survival Analysis (cont'd)

- Procedures to choose a subset of models in VMRK (1997):
  - Apply leaps and bounds algorithm to choose top q models.
  - Use the approximate likelihood ratio test to reduce the subset of models.
  - Calculate BIC values and eliminate the models not in $\mathcal{A}$.
- Posterior effect probability of a variable is computed by

$$P(\beta \neq 0|D) = \sum_{k \in \{i:\ \beta \neq 0 \text{ in } M_i\}} P(M_k|D)$$

- VMRK showed that these posterior effect probabilities can lead to substantive interpretations that are at odds with the usual P-values.

Graphical Models: Missing Data and Auxiliary

- A graphical model is a statistical model with a set of conditional independence relationships being described by means of a graph.
- Acyclic directed graph (ADG):

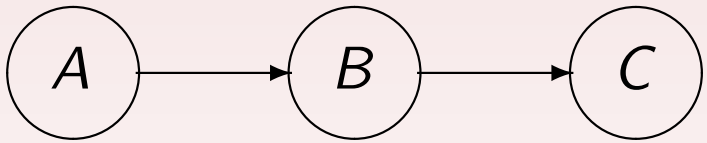


Figure: A simple discrete graphical model.

Graphical Models: Missing Data and Auxiliary(cont'd)

- The above graph tells: C and A are conditionally independent given B. Thus we have

$$pr(A, B, C) = pr(A)\, pr(B|A)\, pr(C|B)$$

- Use either analytical or numerical approximations when we apply BMA and Bayesian graphical models to solve problems with missing data. Please see Madigan and York (1995) and York et al. (1995) for details.
-

$$\frac{pr(D \mid M_0)}{pr(D \mid M_1)} = E\left(\frac{pr(D, Z| M_0)}{pr(D, Z| M_1)} | D, M_1\right)$$

Software for BMA

The programs can be obtained at
http://www.research.att.com/~volinsky/bma.html

- bic.glm
- bic.logit
- bicreg
- bic.surv
- BMA
- glib

Part III

Prasenjit Kapat

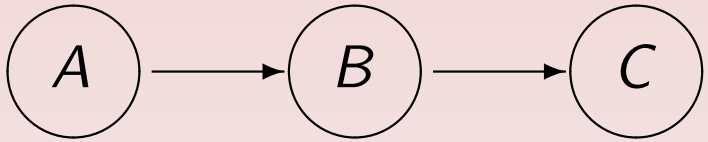
Where are we?

How important is β_j?

- Informative priors provide improved predictive performance, than "neutral" priors.
- Consider the following setup:
  - M_i : Linear model with p covariates.
  - π_j : Prior $P(\beta_j \neq 0)$ (inclusion probability).
  - δ_{ij} : Indicator whether X_j is included in M_i or not.
- The prior for model M_i:

$$pr(M_i) = \prod_{j=1}^{p} \pi_j^{\delta_{ij}} (1 - \pi_j)^{1-\delta_{ij}}$$

In the context of graphical models...



- "link priors": prior probability on existence of each potential link.
- "full prior": product of link priors.
- Hierarchical modelling?

Assumption: presence/absence of each component is apriori independent of the presence/absence of other components.

Eliciting informative prior...

- Start with a uniform prior on the model space.
- Update it using "imaginary data" provided by the domain expert.
- Use the updated prior (posterior based on imaginary data) as the new informative prior for the actual analysis.
- Imaginary data: pilot survey?

Where are we?

How to assess the success of a model?

- An approach: how well a model predicts future observations.
- Split data into two halves: training and testing set.
- Predictive Log Score (P.L.S):

$$\text{Single model: } \sum_{d \in D^{test}} -\log pr(d|M, D^{train})$$

$$\text{BMA: } \sum_{d \in D^{test}} -\log \left\{ \sum_{M \in \mathcal{A}} pr(d|M, D^{train}) \cdot pr(M|D^{train}) \right\}$$

- Smaller P.L.S indicates better predictive performance.

Where are we?

Data description

- Clinical trial of 312 patients from 1974 to 1984.
- Drug: DPCA from Mayo Clinic.
- 14 covariates & one treatment factor (DPCA).
- 8 patients with various missing observations; removed.
- 123 uncensored data (subjects died!).
- 181 censored data (subjects survived!).

Data summary

Variable	Range	Mean	Mean $\beta \mid D$	SD $\beta \mid D$	$P(\beta \neq 0 \mid D)$
Bilirubin (log)	-1.20 - 3.33	0.60	0.784	0.129	100
Albumen (log)	0.67 - 1.54	1.25	-2.799	0.796	100
Age (years)	26 - 78	49.80	0.032	0.010	100
Edema	0 = no edema	n = 263	0.736	0.432	84
	0.5 = edema but no diuretics	n = 29			
	1 = edema despite diuretics	n = 20			
Prothrombin time	2.20 - 2.84	2.37	2.456	1.644	78
Urine copper (log)	1.39 - 6.38	4.27	0.249	0.195	72
Histologic stage	1 - 4	3.05	0.096	0.158	34
SGOT	3.27 - 6.13	4.71	0.103	0.231	22
Platelets	62 - 563	262.30	-0.000	0.000	5
Sex	0 = male	0.88	-0.014	0.088	4
Hepatomegaly	1 = present	0.51	0.006	0.051	3
Alkaline phosphates	5.67 - 9.54	7.27	-0.003	0.028	3
Ascites	1 = present	0.08	0.003	0.047	2
Treatment (DPCA)	1 = DPCA	0.49	0.002	0.028	2
Spiders	1 = present	0.29	0.000	0.027	2
Time observed (days)	41 - 4556	2001			
Status	0 = censored 1 = died	0.40			

Table: PBC example: summary statistics and BMA estimates

Two typical approaches

- Classical Approach:
  - "FH": Fleming and Harrington, 1991.
  - Cox regression model.
  - Multistage variable selection to choose the "best" variables.
  - Chosen variables: age, edema, bilirubin, albumin and prothrombin time.
- Stepwise backward elimination:
  - Final model: age, edema, bilirubin, albumin, prothrombin time and *urine copper*.

New approach: BMA using leaps-and-bounds algo.

Model#	Age	Ede	Bil	Alb	UCop	SGOT	Pro	His	PMP	logLik
1	*	*	*	*	*		*		0.17	-174.4
2	*	*	*	*	*		*	*	0.07	-172.6
3	*	*	*	*	*			*	0.07	-172.5
4	*		*	*	*		*		0.06	-172.2
5	*	*	*	*			*		0.05	-172.0
6	*	*	*	*	*				0.05	-172.0
7	*	*	*	*	*	*	*		0.04	-171.7
8	*	*	*	*		*	*		0.04	-171.4
9	*	*	*	*		*	*	*	0.04	-171.3
10	*	*	*	*	*	*	*	*	0.03	-170.9
$P(\beta \neq 0 \mid D)$	1.00	0.84	1.00	1.00	0.72	0.22	0.78	0.34		

Table: PBC example: results for the full data set

- PMP denotes the posterior model probability . Only the 10 models with highest PMP values are shown.
- Model 5 corresponds to the one selected by FH.

What did we see from the tables?

- Stepwise model: highest approximate posterior probability.
- But, represents only 17% of total posterior probability.
- Fair amount of model uncertainty!
- FH model represents only 5% of total posterior probability.
- $P(\beta \neq 0|D)$: Averaged posterior distribution associated with the variable Edema has 16% of its mass at zero.
- In this process of accounting for the model uncertainty, the standard deviation of the estimates increases.

p-values versus $P(\beta \neq 0|D)$...

Var	p-value	$P(\beta \neq 0|D)$
Bilirubin	$< 0.001^{**}$	> 99 %
Albumen	$< 0.001^{**}$	> 99 %
Age	$< 0.001^{**}$	> 99 %
Edema	0.007^{**}	84 %
Prothrombin	0.006^{**}	78 %
Urine copper	0.009^{**}	72 %
Histology	0.09^{*}	34 %
SGOT	0.08^{*}	22 %

Table: A comparison of some p-values from the stepwise selection model to the posterior effect probabilities from BMA.

p-values versus $P(\beta \neq 0|D)$...

- Qualitatively different conclusions!
- p-values "overstates" the evidence for an effect.
- Distinction between:
  - p-value: not enough evidence to reject (the null) "no-effect".
  - $P(\beta \neq 0)$: evidence in favor of accepting (the null) "no-effect".
- Example:
  - SGOT: 22%; status: indecisive.
  - DPCA: 2%; status: evidence for "no-effect".

Predictive Performance

- Split data randomly in two parts (s.t. 61 deaths in each set).
- Use Partial Predictive Scores (PPS, approximation to PLS)
- BMA predicts who is at risk 6% more effectively than the stepwise model.

Method	PPS
Top PMP Model	221.6
Stepwise	220.7
FH model	22.8
BMA	217.1

Table: PBC example: partial predictive scores for model selection techniques and BMA

BMA by categorizing the patients

Method	Risk Categ.	Survived	Died	% Died
BMA	Low	34	3	8%
	Med	47	15	24%
	High	10	43	81%
Stepwise	Low	41	3	7%
	Med	36	15	29%
	High	14	43	75%
Top PMP	Low	42	4	9%
	Med	31	11	26%
	High	18	46	72%

Table: PBC example: classification for predictive discrimination.

- Risk Scores $(i) = \sum_{k=1}^{K} (x_i' \hat{\beta}^{(k)}) \cdot pr(M_k | D^{train})$; $M_k \in \mathcal{A}$, $\hat{\beta}^{(k)}$ from M_k.
- "A method is better if it consistently assigns higher risks to the peoples who actually die."
- People assigned to higher risk group by BMA had higher death rate than those assigned high risk by the other methods.

Part IV

Juhee Lee

Where are we?

Predicting Percent Body Fat

Overview: Predicting Percent Body Fat

- The goal is to predict percent body fat using 13 simple body mearusements in a multiple regression model.
- Compare BMA to single models selected using several standard variable selection techniques.
- Determine whether there are advantages to accounting for model uncertainty for these data.

Predicting Percent Body Fat

TABLE 6
Body fat example: summary statistics for full data set[1]

Predictor number	Predictor	mean	s.d.	min	max
X_1	Age (years)	45	13	21	81
X_2	Weight (pounds)	179	29	118	363
X_3	Height (inches)	70	3	64	78
X_4	Neck circumference (cm)	38	2	31	51
X_5	Chest circumference (cm)	101	8	79	136
X_6	Abdomen circumference (cm)	93	11	69	148
X_7	Hip circumference (cm)	100	7	85	148
X_8	Thigh circumference (cm)	59	5	47	87
X_9	Knee circumference (cm)	39	2	33	49
X_{10}	Ankle circumference (cm)	23	2	19	34
X_{11}	Extended biceps circumference	32	3	25	45
X_{12}	Forearm circumference (cm)	29	2	21	35
X_{13}	Wrist circumference (cm)	18	1	16	21

[1]Abdomen circumference was measured at the umbilicus and level with the iliac crest. Wrist circumference (cm) was measured distal to the styloid processes.

Predicting Percent Body Fat

- Anlayze the full data set
- Split the data set into two parts, using one portion of the data to do BMA and select models using standard techniques and the other portion to assess performance.
- Compare the predictive performance of BMA to that of individual models selected using standard techniques.
- For Bayesian approach, compute the posterior model probability for all possible models using the diffuse (but proper) prior (Raftery, Madigan and Hoeting, 1997).
- Three choosen techniques: Efroymson's stepwise method, minimum MIlow's C_p, and maximum adjusted R^2

Predicting Percent Body Fat

TABLE 7
Body fat example: least squares regression results from the full model[1]

Predictor		Coef	Std error	*t*-statistic	*p*-value
Intercept		−17.80	20.60	−0.86	0.39
X_1	age	0.06	0.03	1.89	0.06
X_2	weight	−0.09	0.06	−1.50	0.14
X_3	height	−0.04	0.17	−0.23	0.82
X_4	neck	−0.43	0.22	−1.96	0.05
X_5	chest	−0.02	0.10	−0.19	0.85
X_6	abdomen	0.89	0.08	10.62	<0.01
X_7	hip	−0.20	0.14	−1.44	0.15
X_8	thigh	0.24	0.14	1.74	0.08
X_9	knee	−0.02	0.23	−0.09	0.93
X_{10}	ankle	0.17	0.21	0.81	0.42
X_{11}	biceps	0.16	0.16	0.98	0.33
X_{12}	forearm	0.43	0.18	2.32	0.02
X_{13}	wrist	−1.47	0.50	−2.97	<0.01

[1]Residual standard error = 4, $R^2 = 0.75$, $N = 251$, F-statistic = 53.62 on 13 and 237 df, p-value <0.0001.

Predicting Percent Body Fat

Results

Posterior effect probabilities (PEP) $P(\beta_i \neq 0|D)$

- Obtained by summing the posterior model probabilities across models for each predictor
- Abdomen and Weight appear in the models that account for a very high percentage of the total model probability.
- Age, Height, Chest, Ankle, and Knee: smaller than 10%.
- Top three predictors: Weight, Abdomen, and Wrist

BMA results indcate considerable model uncertainty

- The model with the highest posterior model probability (PMP) accounts for only 14% of the total posterior prob.
- The top 10 models account for 57%.

Predicting Percent Body Fat

Comparison of BMA with models selected using standard techniques

- All three standard model selection methods selected the same eight predictor model.
- Agreement: Abdomen, Weight, Wrist.
- Small p values as compared to PEP: Age, Forearm, Neck and Thigh.

Posterior distribution of the coefficient of predictor13 (Wrist) based on the BMA results

- a mixture distribution of non-central Student's t distributions
- spike $P(\beta_{13} = 0|D) = 0.38$

Predicting Percent Body Fat

TABLE 8

Body fat example: comparison of BMA results to model selected using standard model selection methods[1]

Predictor		Bayesian model averaging: Mean $\beta \mid D$	Bayesian model averaging: SD $\beta \mid D$	Bayesian model averaging: $P(\beta \neq 0 \mid D)$	Stepwise model p-value
X_6	abdomen	1.2687	0.08	100	<0.01
X_2	weight	−0.4642	0.15	97	0.03
X_{13}	wrist	−0.0924	0.08	62	<0.01
X_{12}	forearm	0.0390	0.06	35	0.01
X_4	neck	−0.0231	0.06	19	0.05
X_{11}	biceps	0.0179	0.05	17	
X_8	thigh	0.0176	0.05	15	0.02
X_7	hip	−0.0196	0.07	13	0.12
X_5	chest	0.0004	0.02	6	
X_1	age	0.0029	0.02	5	0.05
X_9	knee	0.0020	0.02	5	
X_3	height	−0.0015	0.01	4	
X_{10}	ankle	0.0011	0.01	4	

[1]Stepwise, minimum Mallow's C_p, and maximum adjusted R^2 all selected the same model. The predictors are sorted by $P(\beta_i \neq 0 \mid D)$ which is expressed as a percentage. The results given here are based on standardized data (columns have means equal to 0 and variances equal to 1).

Predicting Percent Body Fat

TABLE 9

Body fat example: Ten models with highest posterior model probability (PMP)

X_2	X_4	X_6	X_8	X_{11}	X_{12}	X_{13}	*PMP*
•		•				•	0.14
•		•			•	•	0.14
•		•					0.12
•		•		•		•	0.05
•		•	•				0.03
•	•	•					0.03
•		•			•		0.02
•		•		•			0.02
•	•	•			•	•	0.02
•	•	•			•		0.02

Predicting Percent Body Fat

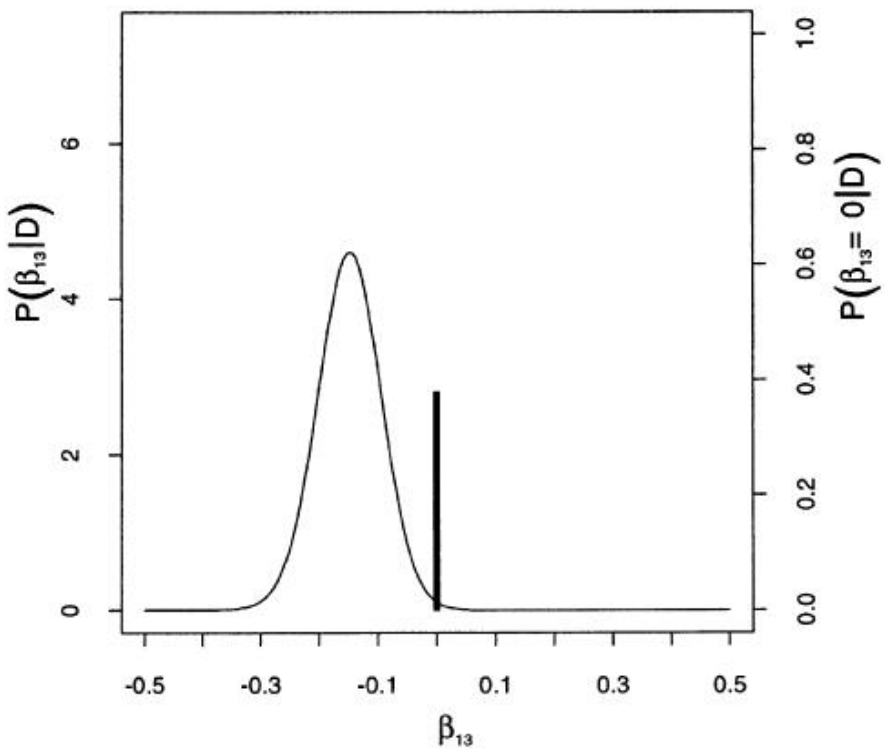


FIG. 4. *Body fat example: BMA posterior distribution for* β_{13}*, the coefficient for wrist circumference. The spike corresponds to* $P(\beta_{13} = 0 \mid D) = 0.38$*. The vertical axis on the left corresponds to the posterior distribution for* β_{13} *and the vertical axis on the right corresponds to the posterior distribution for* β_{13} *equal to 0. The density is scaled so that the maximum of the density is equal to* $P(\beta_{13} \neq 0 \mid D)$ *on the right axis.*

Predicting Percent Body Fat

Predictive Performance

- Predictive coverage was measured using proportion of observations in the performance set that fall in the corresponding 90% prediction interval.
- Prediction interval: the posterior predictive distribution for individual models and a mixture of these posterior predictive distributions for BMA
- Conditioning on a single selected model ingnores model uncertainty.
- Underestimation of uncertainty when making inferences about quantities of interest.
- Predictive coverage is less than the stated coverage level.

Predicting Percent Body Fat

TABLE 10

Body fat example: performance comparison[1]

Method	Model	Predictive coverage %
BMA	Model averaging	90.8
Stepwise and C_p	$X_1\ X_2$ X_6 $X_{10}\ X_{12}\ X_{13}$	84.4
Adjusted R^2	$X_1\ X_2\ X_4\ X_6\ X_7\ X_8$ $X_{10}\ X_{12}\ X_{13}$	83.5

[1]Predictive coverage % is the percentage of observations in the performance set that fall in the 90% prediction interval. For BMA, the top 2500 models, accounting for 99.99% of the posterior model probability, were used to estimate predictive coverage.

Where are we?

8. Examples
   - Predicting Percent Body Fat

9. Discussion
   - Choosing the class of models for BMA
   - Other Approaches to Model Averaging
   - Perspectives on Modeling
   - Conclusion

Choosing the class of models for BMA

In the examples,

- Chose the model structure (e.g., linear regression).
- Averaged either over a **reduced set** of models supported by the data or over the **entire class** of models.

Alternative Approaches (Draper, 1995)

- Finding a good model and then averaging over an expanded class of models 'near' the good model.
- Averaging over models with different error structure.

Other Approaches to Model Averaging

- Frequentist solution to model uncertainty problem: Boostrap the entire data analysis, including model selection
- A minmax multiple shrinkage Stein estimator (George, 1986, a, b, c): When the prior model is finite normal mixtures, these minimax multiple shrinkage estimates are emperical Bayes and formal Bayes estimates.
- Several ad hoc non-Bayesian approaches (Buckland, Burnham and Augustin, 1997): Use AIC, BIC, bootstrapping methods to approximate the model weights
- Computational learning theory (COLT) provides a large body of theoretical work on predictive performance of non-Bayesian model mixing.

Perspectives on Modeling

Two perspecives

- M-closed perspective: the entire class of models is known
- M-open perspective: the model class is not fully known

In the M-open situation, with its open and less constrained search for better models, model uncertainty may be even greater than in the M-closed case, so it may be more important for well-calibrated inference to take account of it.

- The basic principles of BMA apply to the M-open situation.
- The Occam's window approach can be viewed as an implementation of the M open perspective.

Perspectives on Modeling

- Two perspecives
  M-closed perspective: the entire class of models is known
  M-open perspective: the model class is not fully known
- In the M-open situation, with its open and less constrained search for better models, model uncertainty may be even greater than in the M-closed case, so it may be more important for well-calibrated inference to take account of it.
- The basic principles of BMA apply to the M-open situation.
- The Occam's window approach can be viewed as an implementation of the M open perspective.

Conclusion

Taking accout of model uncertainty or uncertainty about statistical structure can be very important, when making inference.

In theory, BMA provides better average predictive performance than any single model.

Common Criticism

- Too complicate to present easily
  focus on the poesterior effect probalbilities
  avoid the problem of having to defend the choice of model
- Higher estimates of variance
  model averaging is more correct

Thank you for the patience! Have a nice break.