Bayesian Model Averaging

Jacob M. Montgomery

Department of Political Science, Washington University in St. Louis

Too many variables, not enough theory

Fearon and Laitin (2003) wish to model civil conflict

Constant	-6.731	-7.019
	(0.736)	(0.751)
Prior war	-0.954	-0.916
	(0.314)	(0.312)
Per capita income	-0.344	-0.318
	(0.072)	(0.071)
log (population)	0.263	0.272
	(0.073)	(0.074)
log (% mountainous)	0.219	0.199
	(0.085)	(0.085)
Noncontiguous state	0.443	0.426
	(0.274)	(0.272)
Oil exporter	0.858	0.751
	(0.279)	(0.278)
New state	1.709	1.658
	(0.339)	(0.342)
Instability	0.618	0.513
	(0.235)	(0.242)
Democracy (Polity IV)	0.021	
	(0.017)	
Ethnic fractionalization	0.166	0.164
	(0.373)	(0.368)
Religious fractionalization	0.285	0.326
	(0.509)	(0.506)
Anocracy		0.521
		(0.237)
Democracy		0.127
		(0.304)
Wars in neighboring countries		

Too many variables, not enough theory

> *"When we add dummy variables for countries that have an ethnic or religious majority and a minority of at least 8% of the countrys poulation, both are incorrectly signed and neither comes close to statistical significance. This finding does not depend on which other variables are included in the model."*

We often want to show that our results are "robust" to modeling choices:

Too many variables, not enough theory

> *"When we add dummy variables for countries that have an ethnic or religious majority and a minority of at least 8% of the countrys poulation, both are incorrectly signed and neither comes close to statistical significance. This finding does not depend on which other variables are included in the model."*

We often want to show that our results are "robust" to modeling choices:

- Online appendix includes 18 additional tables

Too many variables, not enough theory

> *"When we add dummy variables for countries that have an ethnic or religious majority and a minority of at least 8% of the countrys poulation, both are incorrectly signed and neither comes close to statistical significance. This finding does not depend on which other variables are included in the model."*

We often want to show that our results are "robust" to modeling choices:

- Online appendix includes 18 additional tables
- At least 74 possible explnatory variables discussed

Too many variables, not enough theory

"When we add dummy variables for countries that have an ethnic or religious majority and a minority of at least 8% of the countrys poulation, both are incorrectly signed and neither comes close to statistical significance. This finding does not depend on which other variables are included in the model."

We often want to show that our results are "robust" to modeling choices:

- Online appendix includes 18 additional tables
- At least 74 possible explnatory variables discussed
- $2^{74} = 2 \times 10^{22} = 20 sextillion$

Too many variables, not enough theory

> *"When we add dummy variables for countries that have an ethnic or religious majority and a minority of at least 8% of the countrys poulation, both are incorrectly signed and neither comes close to statistical significance. This finding does not depend on which other variables are included in the model."*

We often want to show that our results are "robust" to modeling choices:

- Online appendix includes 18 additional tables
- At least 74 possible explnatory variables discussed
- $2^{74} = 2 \times 10^{22} = 20 sextillion$
- Our actual uncertainty vastly understated

Bayesian model averaging

The setup:

- Y is a $n \times 1$ vector of outcomes
- X is an $n \times p$ matrix
- $Y = X\beta + \epsilon$
- $\epsilon \sim N(0, \sigma^2 I)$
- $q = 2^p$
- $\mathcal{M} = [\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_q]$

Bayesian model averaging

Priors and likelihood:

- $\mathcal{M}_k \sim \pi(\mathcal{M}_k)$
- $\sigma^2|\mathcal{M}_k \sim \pi(\sigma^2|\mathcal{M}_k)$
- $\beta_\omega|\sigma^2, \mathcal{M}_k \sim \pi(\beta_\omega|\sigma^2, \mathcal{M}_k)$
- $\Omega = [\omega_1, \ldots, \omega_p]$ is a binary vector indicating inclusion.

Bayesian model averaging

Priors and likelihood:

- $\mathcal{M}_k \sim \pi(\mathcal{M}_k)$
- $\sigma^2|\mathcal{M}_k \sim \pi(\sigma^2|\mathcal{M}_k)$
- $\beta_\omega|\sigma^2, \mathcal{M}_k \sim \pi(\beta_\omega|\sigma^2, \mathcal{M}_k)$
- $\Omega = [\omega_1, \ldots, \omega_p]$ is a binary vector indicating inclusion.

$$p(Y|\beta_\omega, \sigma^2, \mathcal{M}_k) \sim N(X_\omega \beta_\omega, \sigma^2 I)$$

Bayesian model averaging

Priors and likelihood:

- $\mathcal{M}_k \sim \pi(\mathcal{M}_k)$
- $\sigma^2|\mathcal{M}_k \sim \pi(\sigma^2|\mathcal{M}_k)$
- $\beta_\omega|\sigma^2, \mathcal{M}_k \sim \pi(\beta_\omega|\sigma^2, \mathcal{M}_k)$
- $\Omega = [\omega_1, \ldots, \omega_p]$ is a binary vector indicating inclusion.

$$p(Y|\beta_\omega, \sigma^2, \mathcal{M}_k) \sim N(X_\omega \beta_\omega, \sigma^2 I)$$

$$p(Y|\mathcal{M}_k) \sim \int \int p(Y|\beta_\omega, \sigma^2, \mathcal{M}_k) \pi(\beta_\omega|\sigma^2, \mathcal{M}_k) \pi(\sigma^2|\mathcal{M}_k) d\beta_\omega d\sigma^2$$

Bayesian model averaging

$$P(\mathcal{M}_k|Y) = \frac{p(Y|\mathcal{M}_k)\pi(\mathcal{M}_k)}{\sum_k p(Y|\mathcal{M}_k)\pi(\mathcal{M}_k)}$$

Bayesian model averaging

$$P(\mathcal{M}_k|Y) = \frac{p(Y|\mathcal{M}_k)\pi(\mathcal{M}_k)}{\sum_k p(Y|\mathcal{M}_k)\pi(\mathcal{M}_k)}$$

With this, we can easily create other quantities of interest as weighted sums. For example,

$$E(\beta_k|Y) = \sum_{k=0}^{q} P(\mathcal{M}_k|Y)E(\beta|M_k, Y)$$

Bayesian model averaging

$$P(\mathcal{M}_k|Y) = \frac{p(Y|\mathcal{M}_k)\pi(\mathcal{M}_k)}{\sum_k p(Y|\mathcal{M}_k)\pi(\mathcal{M}_k)}$$

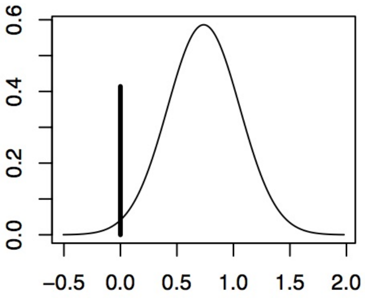
Bayesian model averaging

$$P(\mathcal{M}_k|Y) = \frac{p(Y|\mathcal{M}_k)\pi(\mathcal{M}_k)}{\sum_k p(Y|\mathcal{M}_k)\pi(\mathcal{M}_k)}$$

With this, we can easily create other quantities of interest as weighted sums. For example,

$$E(\beta_k|Y) = \sum_{k=0}^{q} P(\mathcal{M}_k|Y)E(\beta|M_k, Y)$$

What do we get?



1. Does the variable contribute to the models explanatory power? (i.e. what is the posterior probability of all models that include this variable?)
2. Is it correlated with unexplained variance when it is included? (i.e. what is the conditional posterior distribution assuming that the variable is included?)

Practical and technical notes

- Can constrain model space in motivated ways

Practical and technical notes

- Can constrain model space in motivated ways
- Can reformulate the posterior for each model probability as:

$$P(\mathcal{M}_k|Y) = \frac{p(Y|\mathcal{M}_k)\pi(\mathcal{M}_k)/p(\mathcal{M}_0|Y)\pi(\mathcal{M}_0)}{\sum_k p(Y|\mathcal{M}_k)\pi(\mathcal{M}_k)\pi(\mathcal{M}_k)/p(\mathcal{M}_0|Y)\pi(\mathcal{M}_0)}$$

Practical and technical notes

- Can constrain model space in motivated ways
- Can reformulate the posterior for each model probability as:

$$P(\mathcal{M}_k|Y) = \frac{p(Y|\mathcal{M}_k)\pi(\mathcal{M}_k)/p(\mathcal{M}_0|Y)\pi(\mathcal{M}_0)}{\sum_k p(Y|\mathcal{M}_k)\pi(\mathcal{M}_k)\pi(\mathcal{M}_k)/p(\mathcal{M}_0|Y)\pi(\mathcal{M}_0)}$$

- Also need a prior on the model space: $\pi(\mathcal{M}_k) = \gamma^{p_\omega}(1-\gamma)^{p-p_\omega}$

Practical and technical notes

- Can constrain model space in motivated ways
- Can reformulate the posterior for each model probability as:

$$P(\mathcal{M}_k|Y) = \frac{p(Y|\mathcal{M}_k)\pi(\mathcal{M}_k)/p(\mathcal{M}_0|Y)\pi(\mathcal{M}_0)}{\sum_k p(Y|\mathcal{M}_k)\pi(\mathcal{M}_k)\pi(\mathcal{M}_k)/p(\mathcal{M}_0|Y)\pi(\mathcal{M}_0)}$$

- Also need a prior on the model space: $\pi(\mathcal{M}_k) = \gamma^{p_\omega}(1-\gamma)^{p-p_\omega}$
- Can even put a hyperprior on γ

Prior Selection

The limiting factor in BMA approaches has always been a combination of:

- an *extremely* high dimensional space
- intractable integrals.

Early work used “approximations” of the Bayes factors as a solution:

- $BIC_k = -2\log(L_k - L_0) + p_k \log n$
- $AIC_k = -2\log(L_k - L_0) + 2p$

Newer alternatives

Clyde (2003) and Clyde and George(2004) summarize a more comprehensive approach.

$$\pi(\mathcal{M}_k) = \gamma^{p_\omega}(1-\gamma)^{p-p_\omega} \tag{1}$$

Priors for posterior calculations

Prior	*Formulation*
g-prior	$\pi(\beta_\omega \mid \mathcal{M}_k, \sigma^2) \sim N_{p_\omega}(0, g\sigma^2(X'_\omega X_\omega)^{-1})$ $\pi(\beta_0, \sigma^2 \mid \mathcal{M}_k) \propto 1/\sigma^2$

Hyper-priors for g

Prior	*Formulation*
Hyper-g	$\pi(g) = \frac{a-2}{2}(1+g)^{\frac{a}{2}} \text{if} g > 0$
ZS	$\pi(g) = \frac{(n/2)^{1/2}}{\Gamma(1/2)} g^{-3/2} e^{-\frac{n}{2g}}$

Mean cand.	Voter-specific	Mean cand.		Voter-specific	
Mean	Mean	Cond. mean	$P(\beta \neq 0)$	Cond. mean	$P(\beta \neq 0)$
(SD)	(SD)	(SD)		(SD)	
-3.053	-2.007	-0.222	0.067	0.745	0.033
(1.315)	(1.056)	(0.954)		(0.677)	
7.953	4.177	3.576	0.299	2.363	0.787
(2.854)	(1.356)	(2.015)		(0.840)	
1.060	1.139	1.607	0.194	1.295	0.159
(1.201)	(1.189)	(1.242)		(1.237)	
3.117	2.378	2.964	0.599	2.740	0.541
(1.240)	(1.216)	(1.246)		(1.223)	
0.270	0.265	0.3238	1.00	0.314	1.00
(0.041)	(0.040)	(0.041)		(0.037)	
0.055	0.060	0.0749	0.201	0.0661	0.181
(0.054)	(0.054)	(0.057)		(0.055)	
53.309	52.028	51.381	1.00	51.556	1.00
(1.155)	(0.758)	(1.032)		(0.764)	
94	94	94		94	

Improving forecasting with BMA

- We often have many forecasting models for specific outcomes
- Not all of them are equally valuable, and not all provide unique insight
- Can we combine forecasts to reduce model dependency and improve our out-of-sample performance?

Improving forecasting with BMA

The setup:

- $\mathbf{y}^{t*}$ are outcomes in the future we want to predict.

Improving forecasting with BMA

The setup:

- $\mathbf{y}^{t*}$ are outcomes in the future we want to predict.
- $\mathbf{y}^{t}$ are outcomes in the past that we previously tried to predict (out of sample)

Improving forecasting with BMA

The setup:

- $\mathbf{y}^{t*}$ are outcomes in the future we want to predict.
- $\mathbf{y}^{t}$ are outcomes in the past that we previously tried to predict (out of sample)
- We have K forecasting models or teams, $M_1, M_2, \ldots, M_K$.

Improving forecasting with BMA

The punchline

- $M_k \sim \pi(M_k)$

Improving forecasting with BMA

The punchline

- $M_k \sim \pi(M_k)$
- Pdf for the forecast is $p(\mathbf{y}^t | M_k)$

Improving forecasting with BMA

The punchline

- $M_k \sim \pi(M_k)$
- Pdf for the forecast is $p(\mathbf{y}^t|M_k)$
- $$p(M_k|\mathbf{y}^t) = \frac{p(\mathbf{y}^t|M_k)\pi(M_k)}{\sum_k p(\mathbf{y}^t|M_k)\pi(M_k)}$$

Improving forecasting with BMA

The punchline

- $M_k \sim \pi(M_k)$
- Pdf for the forecast is $p(\mathbf{y}^t|M_k)$
- $$p(M_k|\mathbf{y}^t) = \frac{p(\mathbf{y}^t|M_k)\pi(M_k)}{\sum_k p(\mathbf{y}^t|M_k)\pi(M_k)}$$
- $$p(\mathbf{y}^{t*}) = \sum p(\mathbf{y}^{t*}|M_k)p(M_k|\mathbf{y}^t)$$

Improving forecasting with BMA

The punchline

- $M_k \sim \pi(M_k)$
- Pdf for the forecast is $p(\mathbf{y}^t|M_k)$
- $$p(M_k|\mathbf{y}^t) = \frac{p(\mathbf{y}^t|M_k)\pi(M_k)}{\sum_k p(\mathbf{y}^t|M_k)\pi(M_k)}$$
- $$p(\mathbf{y}^{t*}) = \sum p(\mathbf{y}^{t*}|M_k)p(M_k|\mathbf{y}^t)$$
- $$E(\mathbf{y}^{t*}) = \sum E(\mathbf{y}^{t*}|M_k)p(M_k|\mathbf{y}^t)$$

EBMA as a finite mixture model

- Denote $w_k = p(M_k|\mathbf{y}^t)$
- Let $p(\mathbf{y}^{t^*}|M_k) = N(f_k^{t^*}, \sigma^2)$

$$p(y|f_1^{s|t^*}, \ldots, f_K^{s|t^*}) = \sum_{k=1}^{K} w_k N(f_k^{t^*}, \sigma^2). \tag{2}$$

$$\mathcal{L}(\mathbf{w}, \sigma^2) = \sum_t \log\left(\sum_{k=1}^{K} w_k N(f_k^t, \sigma^2)\right), \tag{3}$$

E-M Algorithm

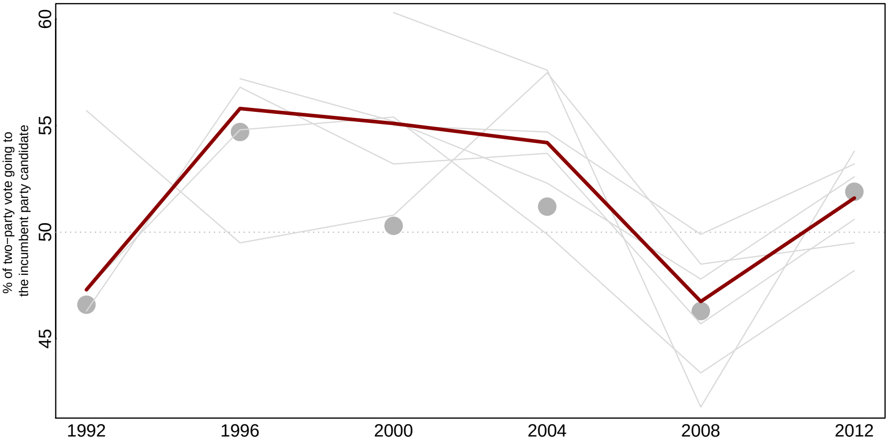
$$\hat{z}_k^{(j+1)t} = \frac{\hat{w}_k^{(j)} p^{(j)}(y|f_k^t)}{\sum\limits_{k=1}^{K} \hat{w}_k^{(j)} p^{(j)}(y|f_k^t)}, \tag{4}$$

$$\hat{w}_k^{(j+1)} = \frac{1}{n} \sum_t \hat{z}_k^{(j+1)t}, \tag{5}$$

Example: Predicting presidential elections

- **Campbell**: Campbells Trial-Heat and Economy Model
- **Abramowitz**: The Time-for-Change Model created by ?
- **Fair**: Fairs presidential vote-share model16
- **Lewis-Beck/Tien**: Lewis-Beck and Tiens Jobs Model Forecast
- **EW**: Erikson & Wlezien,

Previous performance



	2004 Election				2008 Election			
	Weights	RMSE	MAE	Pred. Error	Weights	RMSE	MAE	Pred. Error
Campbell	0.40	1.71	1.33	0.53	0.36	1.65	1.28	6.33
Abramowitz	0.00	1.50	1.18	2.20	0.06	1.53	1.26	−2.37
Hibbs	0.12	1.95	1.38	1.54	0.25	1.92	1.38	−1.39
Fair	0.48	2.07	1.47	4.82	0.00	2.22	1.80	−2.02
Lewis-Beck/Tien	0.00	1.67	1.42	−0.41	0.17	1.61	1.33	−2.65
Erikson/Wlezien	0.00	2.67	2.06	4.76	0.17	2.81	2.18	−0.14
EBMA		1.29	1.01	2.08		1.30	1.01	−0.53

Table 1

Ensemble Weights and Fit Statistics for Calibration-Period Performance (1948–2008)

	ENSEMBLE WEIGHT	RMSE	MAE
Ensemble		0.859	0.696
Abramowitz	0.674	0.981	0.769
Berry	0.006	0.808	0.750
Campbell (Trial Heat)	0.047	1.610	1.252
Cuzán (FPRIME short)	0.178	1.800	1.357
Erikson/Wlezien	0.012	1.775	1.549
Hibbs	0.004	2.806	2.240
Holbrook	0.015	2.144	1.734
Lewis-Beck/Tien (Jobs)	0.039	1.264	1.050
Lockerbie	0.009	3.943	3.329
Norpoth/Bednarczuk	0.015	2.411	2.129

The second column contains the weight assigned each component model in the final ensemble. The other columns show two fit statistics to evaluate the relative performance of each component model and the ensemble across the calibration period. EBMA tends to place higher weight on better performing models, but the relationship is not montonic.

- Forecast 50.2 [46.4, 52.5]
- Outcome 51.3%

Truly Bayesian EBMA

- $t = [1, \ldots, T]$ is the number of predictions being made.
- $k = [1, \ldots, K]$ is the number of models making predictions.
- y_t is the observed outcome for period t.
- $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K]$, where $\mathbf{x}_k = (x_{k1}, x_{k2}, \ldots, x_{kT})$ is the vector of predictions made by model k.
- $\boldsymbol{\tau} = [\tau_1, \tau_2, \tau_T]$ indexes which model actually generated observation t such that $\tau_t \in [1, 2, \ldots, K] \forall t \in [1, 2, \ldots, T]$

$$p(y_t|\boldsymbol{\tau}, \boldsymbol{\sigma}^2, \mathbf{X}) \sim \sum_k^K N(x_{kt}, \sigma)\mathcal{I}(\tau_t = k), \tag{6}$$

where $\mathcal{I}(\cdot)$ is the standard indicator function. The model is complete by specifying the following priors/hyperpriors.

$$\pi(\boldsymbol{\tau}) \sim \text{Multinomial}(\boldsymbol{\omega}) \tag{7}$$

$$\pi(\boldsymbol{\omega}) \sim \text{Dirichlet}(\boldsymbol{\alpha}) \tag{8}$$

$$\pi(\sigma^2) \sim (\sigma^2)^{-1} \tag{9}$$

Let Θ be a T by K matrix holding a parameter indicating the latent probability such that θ_{tk} represents the that observation t comes from model k. We calculate that,

$$p(\theta_{tk}|\mathbf{X}, \mathbf{y}, \boldsymbol{\omega}) = \frac{\omega_k N(y_t|x_{tk}, \sigma)}{\sum_k^K \left(\omega_k N(y_t|x_{tk}, \sigma)\right)} \tag{10}$$

We then draw:

$$\tau_t|\Theta \sim \text{Multinomial}(\boldsymbol{\theta}_t) \tag{11}$$

We then draw:

$$\boldsymbol{\omega}|\boldsymbol{\phi} \sim \text{Dirichlet}(\boldsymbol{\eta}), \tag{12}$$

where $\eta_k = \alpha_k + \sum_{t=1}^T \mathcal{I}(\tau_t = k)$

Finally, we need to calculate the conditional distribution for the posterior for the common variance term σ^2, which is

$$\sigma^2|\boldsymbol{\tau} \sim \text{Inv.}\chi^2\Big(\frac{T-1}{2}, \frac{\sum_{t=1}^{T}\left(y_t - \sum_{k=1}^{K} x_{tk}\mathcal{I}(\tau_t = k)\right)}{2}\Big) \qquad (13)$$

Forecasting the 2016 election