Finite Mixture Models

Jacob M. Montgomery

Department of Political Science, Washington University in St. Louis

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

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

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

Expectation-Maximization Algorithm

- Setting aside σ^2 for the moment, for problems such as this we need to find the values of $\mathbf{w}$. Maximize the log-likelihood.
- The EM algorithm works by breaking the problem into two simpler parts.
- We introduce a latent variable z_k^t, which is the probability that observation t comes from model k.

$$\mathbf{Z} = \begin{pmatrix} z_1^1 & z_2^1 & \dots & z_K^1 \\ z_1^2 & z_2^2 & \dots & z_K^2 \\ \vdots & & & \vdots \\ z_1^T & & \dots & z_K^T \end{pmatrix}$$

If we knew these values *for certain*, maximizing the log-likelihood would be easy, similar to how $\bar{x}$ is the MLE for a μ in a normal distribution.

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

$$\hat{w}_k = \frac{1}{n} \sum_t z_k^t$$

Now we can turn it around. Assume that we *knew* the true value of $\mathbf{w}$, what would be the expected value of our augmented data $\mathbf{Z}$?

$$\hat{z}_k^t = \frac{\hat{w}_k p(y|f_k^t)}{\sum_{k=1}^{K} \hat{w}_k p(y|f_k^t)}$$

In the same way that:

$$p(M_k|\mathbf{y}^t) = \frac{p(\mathbf{y}^t|M_k)\pi(M_k)}{\sum_k p(\mathbf{y}^t|M_k)\pi(M_k)}$$

E-M Algorithm for mixture models

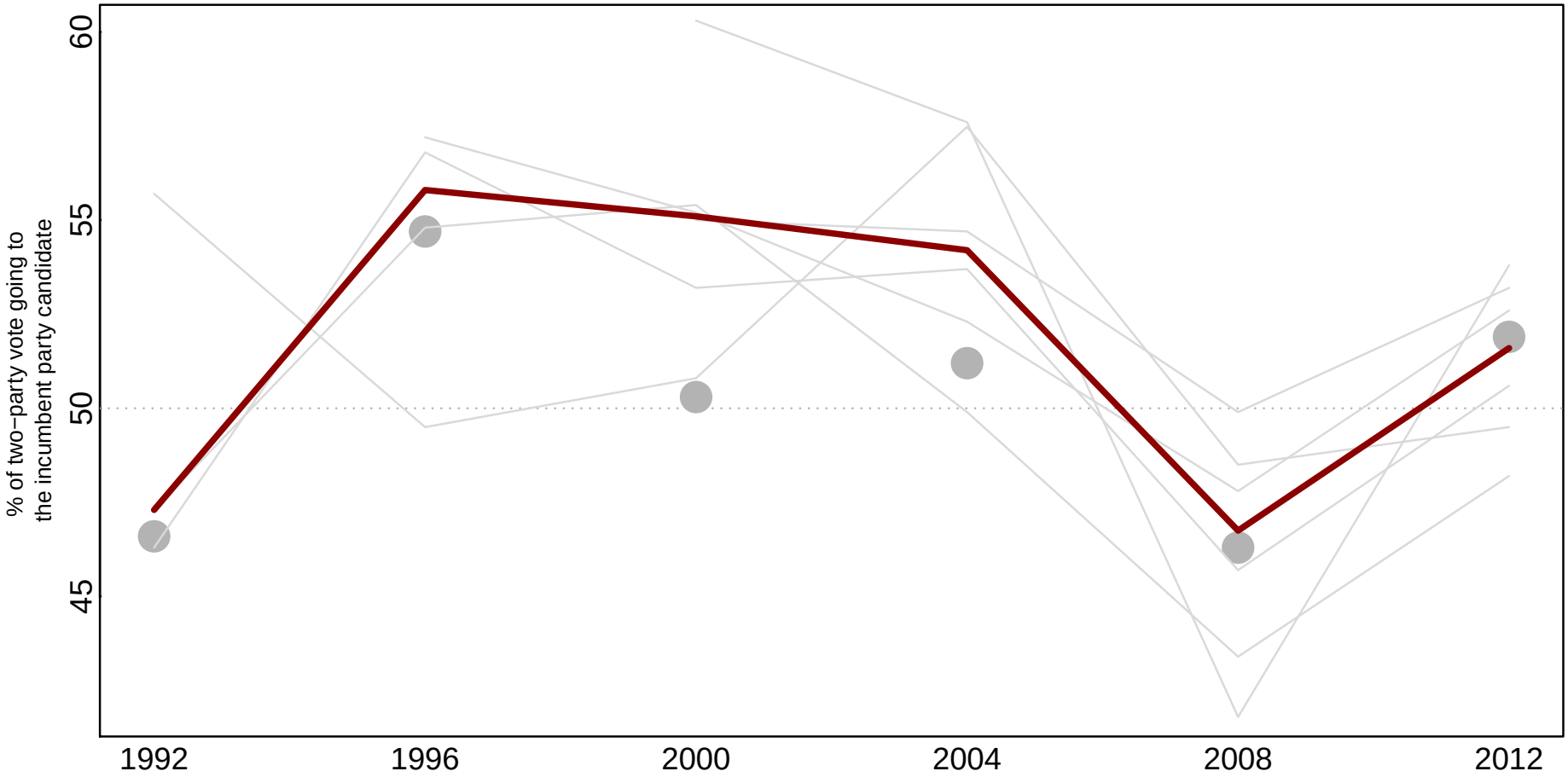
$$\hat{z}_k^{(j+1)t} = \frac{\hat{w}_k^{(j)} p^{(j)}(y|f_k^t)}{\sum_{k=1}^{K} \hat{w}_k^{(j)} p^{(j)}(y|f_k^t)},$$

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

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 $\tau_t \in [1, 2, \ldots, K] \forall t \in [1, 2, \ldots, T]$

$$p(y_t|\boldsymbol{\tau}, \boldsymbol{\sigma}^2, \mathbf{X}) \sim \sum_{k}^{K} N(x_{kt}, \sigma)\mathcal{I}(\tau_t = k), \tag{1}$$

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

$$\pi(\boldsymbol{\tau}) \sim \text{Multinomial}(\boldsymbol{\omega}) \tag{2}$$

$$\pi(\boldsymbol{\omega}) \sim \text{Dirichlet}(\boldsymbol{\alpha}) \tag{3}$$

$$\pi(\sigma^2) \sim (\sigma^2)^{-1} \tag{4}$$

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,

$$p(\theta_{tk}|\mathbf{X}, \mathbf{y}, \boldsymbol{\omega}) = \frac{\omega_k N(y_t|x_{tk}, \sigma)}{\sum_k^K \left(\omega_k N(y_t|x_{tk}, \sigma)\right)} \tag{5}$$

We then draw:

$$\tau_t|\Theta \sim \text{Multinomial}(\boldsymbol{\theta}_t) \tag{6}$$

We then draw:

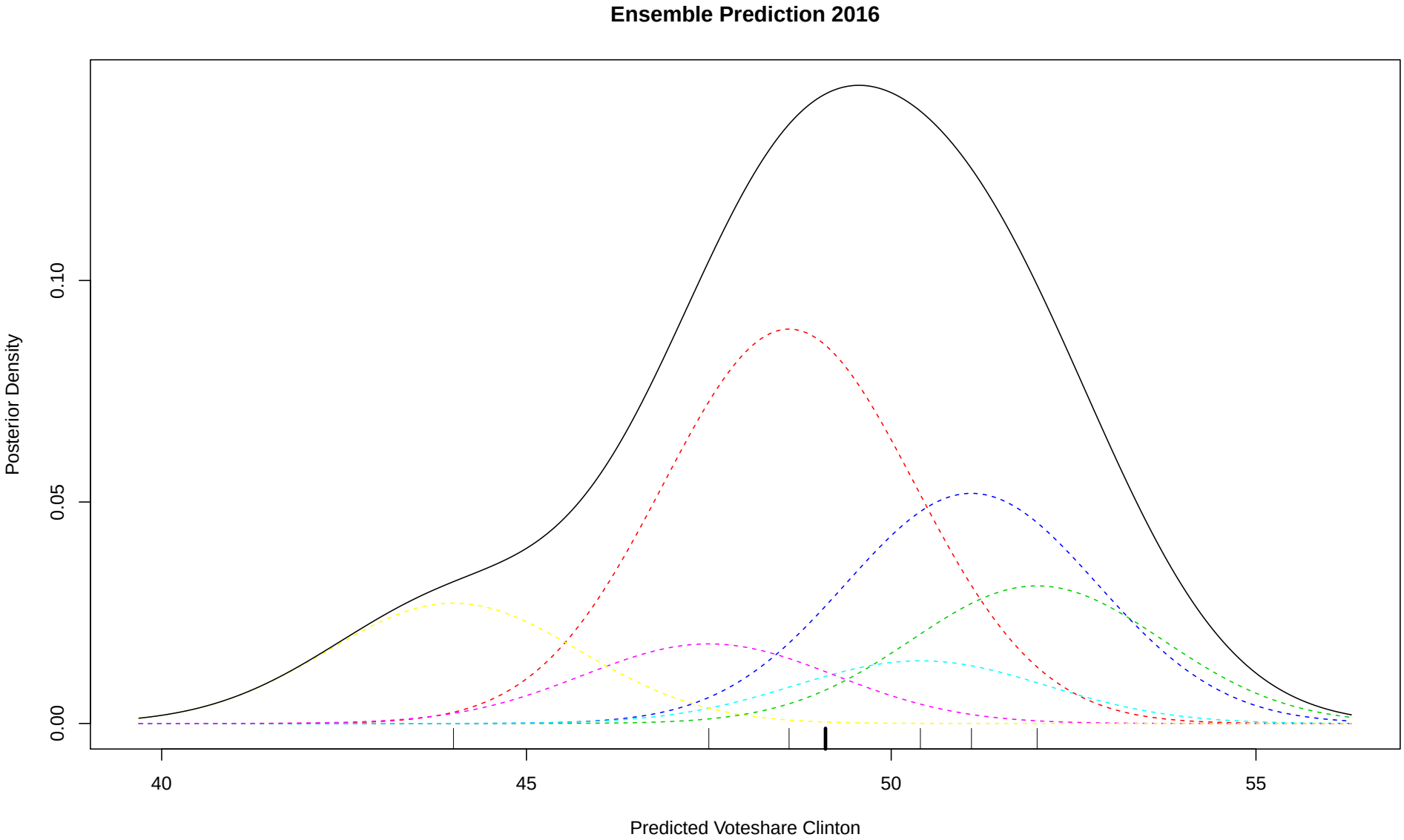
$$\boldsymbol{\omega}|\boldsymbol{\phi} \sim \text{Dirichlet}(\boldsymbol{\eta}), \tag{7}$$

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

$$\sigma^2|\boldsymbol{\tau} \sim \text{Inv.}\chi^2\left(\frac{T-1}{2}, \frac{\sum_{t=1}^{T}\left(y_t - \sum_{k=1}^{K} x_{tk}\mathcal{I}(\tau_t = k)\right)}{2}\right) \tag{8}$$

Forecasting the 2016 election



Stepping back: Mixture models

The framework above gives us the ability to:

- Set up multiple models
- Get posterior estimates for which observation belongs to which model
- Get overall posterior probabilities for which model is most correct most often

Stepping back: Mixture models

The framework above gives us the ability to:

- Set up multiple theories
- Get posterior estimates for which observation belongs to which theory
- Get overall posterior probabilities for which theory is most correct most often

We can also add in a step where we directly model the probability for each observation to be assigned to each model.

Imai and Tingley 2012 - AJPS

What drives legislative voting on trade policy?:

- Stolper-Samuelson (SS) argue that senators will vote for trade liberalization when their districts are rich in factors scarce in the rest of the world
- Ricardo-Viner (RV) model argues that export-oriented firms will favor liberalization, while import competing industries will oppose.

Imai and Tingley 2012 - AJPS

- Dependent variable is a Senators vote on a trade bill.
- SS:
  - `profit`: State-level measure of profits
  - `manufacture`: Employment in manufacturing
  - `farm`: Employment in farm sector
- RV:
  - `export`:Export orientation of the state
  - `import`: Import competing nature of the state economy

$$f_{ss}(Y_{ij}|X_{ij}) = logit^{-1}(\beta_0 + \beta_1 profit + \beta_2 manufacture + \beta_3 farm)$$

$$f_{rv}(Y_{ij}|X_{ij}) = logit^{-1}(\beta_0 + \beta_i import + \beta_2 export)$$

Taking it a bit further

- Note that the models are not nested, except for the β_0
- Hiscox (2002) argues that which model is most appropriate will be a function of a national measure of factor specificity (immobility of factors).

$$\pi_{rv} = logit^{-1}(\delta_0 - \delta_1 factor_i)$$

Fitting the model

`flexmix` package in R allows you to fit the model using the same EM algorithm as above:

```
model <- FLXMRglmfix(family = "binomial",
         nested = list (k = c(1, 1),
         formula = c(~ profit +manufacture +farm,
                     ~ export + import )
```

Fitting the model

```
result <- stepFlexmix(cbind(vote, 1 - vote)~ 1|bill,
                      k = 2, model = model,
                      concomitant = FLXPmultinom(factor),
                      data = Hiscox, nrep = 20)
```

Let's do this thing!

`https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/16378`

`http://tinyurl.com/gtkq3uv`

We want to work with the file:
`ImaiTingleyAJPSReplication_Hiscox.R`