A Statistical Method for Empirical Testing of Competing Theories

Kosuke Imai and Dustin Tingley

American Journal of Political Science, 56(1)

2012

Outline

Goals

- Problem: "...most political scientists fit a regression model with many explanatory variables that are derived from multiple theories." (p. 218, the "garbage can model")
- Proposed solution: finite mixture models, in which each "observation is assumed to be generated either from a statistical model implied by one of the rival theories or more generally from a weighted combination of multiple statistical models under consideration."

Examples

- Is support for free trade motivated by factoral or sectoral differences?
- Are education outcomes driven by school quality or community characteristics?
- Are gender wage gaps driven by parenthood status or stereotypes about competency?
- ?

Advantages over Other Model Selection Approaches

(e.g., the J test)

- Allows model covariates to predict component membership
- Includes uncertainty of model membership (rather than discrete assignments)
- Can accommodate any likelihood-based model
- Allows competing theories to coexist in ways that are interpretable

Data Generating Process

$Y_i \mid X_i, Z_i \sim f_{Z_i}(Y_i \mid X_i, \theta_{Z_i})$ (Eq. 1, p. 220)

- Observations can come from some mixture (or discrete assignment) of M theories
- Z_i: latent variable indicating that observation i is consistent with model Z, where Z $\in \{1, 2, \ldots. \text{ M}\}$
- Also possible to group observations into groups j, e.g., for multiple observations of same individual, in same village, etc.

Likelihood

For a mixture model with m components:

$$L_{obs}(\theta, \Pi \mid \{X_i, Y_I\}_{i=1}^{N}) = \Pi_{i=i}^{N} \Big\{ \sum_{m=1}^{M} \pi_m f_m(Y_i \mid X_i, \theta_m) \Big\}$$

(Eq.2, p.221) where $\pi_m = \Pr(Z_i = m)$ is the population proportion of observations generated by theory m.

Log-Likelihood Function (Complete Data)

(Eq. 4, p. 222)

$$l_{com}(\theta, \Pi \mid \{X_i, Y_i, Z_i\}_{i=1}^N) =$$

$$\sum_{i=1}^N \sum_{m=1}^M \mathbf{1}\{Z_i = m\}\{\log \pi_m + \log f_m(Y_i \mid X_i, \theta_m)\}$$

Expectation-Maximization

(Eq. 5, p. 222)

$$Q(\theta, \Pi \mid \theta^(t-1), \Pi^(t-1), \{X_i, Y_i, Z_i\}_{i=1}^N) =$$

$$\sum_{i=1}^N \sum_{m=1}^M \zeta_{i,m}^{(t-1)} \{\log \pi_m + \log f_m(Y_i \mid X_i, \theta_m)\}$$

Where $\zeta_{i,m}^{(t-1)}$ is the posterior probability that observation i is associated with theory (model component) m.

Calculating ζ

(Eq. 6, p. 222)

The conditional expectation of $\zeta_{i,m}^{(t-1)} =$

$$\Pr(Z_i = m \mid \theta^{(t-1)}, \Pi^{(t-1)}, \{X_i, Y_i, Z_i\}_{i=1}^{N}) =$$

$$\frac{\pi_m^{(t-1)} f_m(Y_i|X_i,\theta_m)}{\sum_{m'}^{M} \pi_{m'}^{(t-1)} f_m'(Y_i|X_i,\theta_{m'})}$$

The Use of π_m

π_m is a function of theory-predicting variables W_i, which may overlap with model variables X_i:

$$\pi_m = \Pr(Z_i = m \mid W_i) = \pi_m(W_i, \psi_m)$$

(Eq. 3, p. 222)

Variables in W_i that are strong predictors can be used to make inferences about which theories apply to which circumstances.
Potential Issue: IIA Assumption

The Use of π_m

π_m is estimated by averaging the fitted values of ζ:

$$\pi_m = \frac{1}{N} \sum_{i=1}^{N} \zeta_{i,m}^{(t-1)}$$

(Eq. 7, p. 222)

"Statistically Significantly Consistent Observations"

- Observations can be identified as consistent with particular theories when their posterior probability $\zeta_{i,m}$ exceeds some predetermined threshold, λ_m.
- However, risk of multiple hypothesis testing and Type I error increases with m.
- An optimal λ_m can be computed by setting a false discovery threshold α.
- λ and α values can be set for each theory m separately or for the entire model.

"Statistically Significantly Consistent Observations"

$$\lambda : \frac{\sum_{i=1}^{N}\sum_{m=1}^{M}(1-\hat{\zeta}_{i,m})\mathbf{1}\{\hat{\zeta}_{i,m}\geq\lambda\}}{\sum_{i=1}^{N}\sum_{m=1}^{M}\mathbf{1}\{\hat{\zeta}_{i,m}\geq\lambda\}+\prod_{i=1}^{N}\prod_{m=1}^{M}\mathbf{1}\{\hat{\zeta}_{i,m}<\lambda\}} \leq \alpha$$

(Eq. 13, p. 224)

- Numerator: Posterior expected number of false positives
- Denominator: Total positives (plus an indicator variable that equals 1 if the first term equals 0)
- Both π_m and proportion of SSOs can be used to evaluate theory performance.

Empirical Example

What determines votes on trade bills, sectoral cleavage or factor specificity? p. 230

$$f_{SS}(Y_{ij} \mid X_{ij}, \theta_{SS}) = \text{logit}^{-1}(\beta_0 + \beta_1 \text{profit}_{ij} + \beta_2 \text{manufacture}_{ij} + \beta_3 \text{farm}_{ij}) \quad (1)$$

$$f_{RV}(Y_{ij} \mid X_{ij}, \theta_{RV}) = \text{logit}^{-1}(\gamma_0 + \gamma_1 \text{export}_{ij} + \gamma_2 \text{import}_{ij}) \quad (2)$$

$$\pi_{RV}(W_j, \phi_R V) = \text{logit}^{-1}(\delta_0 + \delta_1 \text{factor}_j) \quad (3)$$

R Example

```
library(flexmix)

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

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

Quantity of Interest: Coefficients

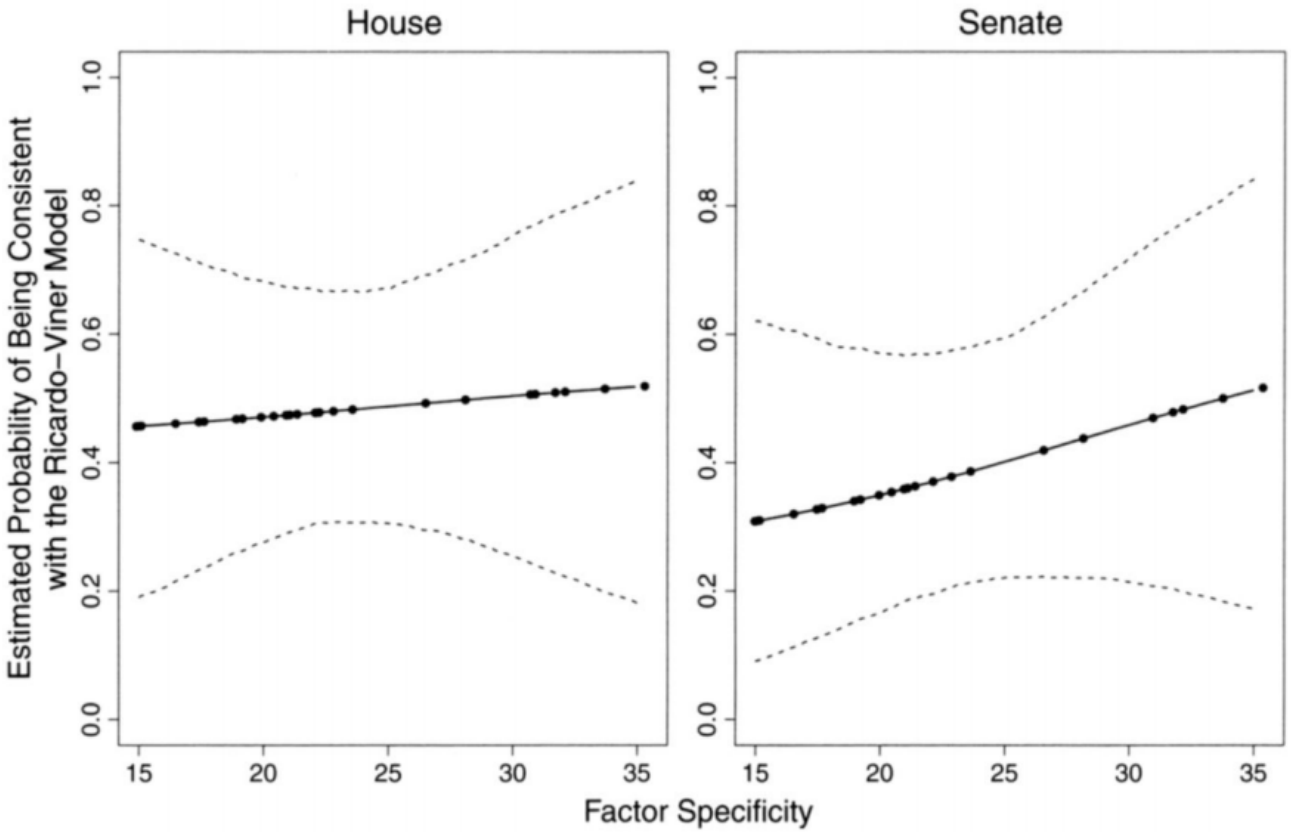
TABLE 1 Parameter Estimates and Their Standard Errors from the Mixture Model for the House and Senate

		Mixture Model				"Garbage-can" Model			
		House		Senate		House		Senate	
Models	**Variables**	**coef.**	**s.e.**	**coef.**	**s.e.**	**coef.**	**s.e.**	**coef.**	**s.e.**
Stolper-Samuelson	intercept	−0.23	0.14	0.02	0.21	0.47	0.12	0.78	0.25
	profit	−1.60	0.53	−5.69	1.19	−0.93	0.56	−3.58	1.23
	manufacture	17.60	1.54	19.79	2.59	10.01	1.11	7.82	2.27
	farm	−1.33	0.29	−1.27	0.43	−0.14	0.24	−0.03	0.42
Ricardo-Viner	intercept	−0.61	0.05	−0.83	0.13				
	import	3.09	0.33	2.53	0.80	1.03	0.34	2.22	0.76
	export	−0.85	0.16	−2.80	0.77	−1.45	0.14	−2.58	0.36
Mixture Probability	intercept	−0.39	1.48	−1.60	1.62				
	factor	0.01	0.06	0.05	0.07				

Each model is the logistic regression with model intercepts omitted in order to ease presentation. The first set of models uses the proposed

Quantity of Interest: Relationship between ζ and W

FIGURE 3 Estimated Probability of Votes for a Bill Being Consistent with the Ricardo-Viner Model as a Function of Factor Specificity



Another Empirical Example

Does construction of long distance bus stations incentives labor migration or encourage potential migrants to seek opportunities at home?

R Example

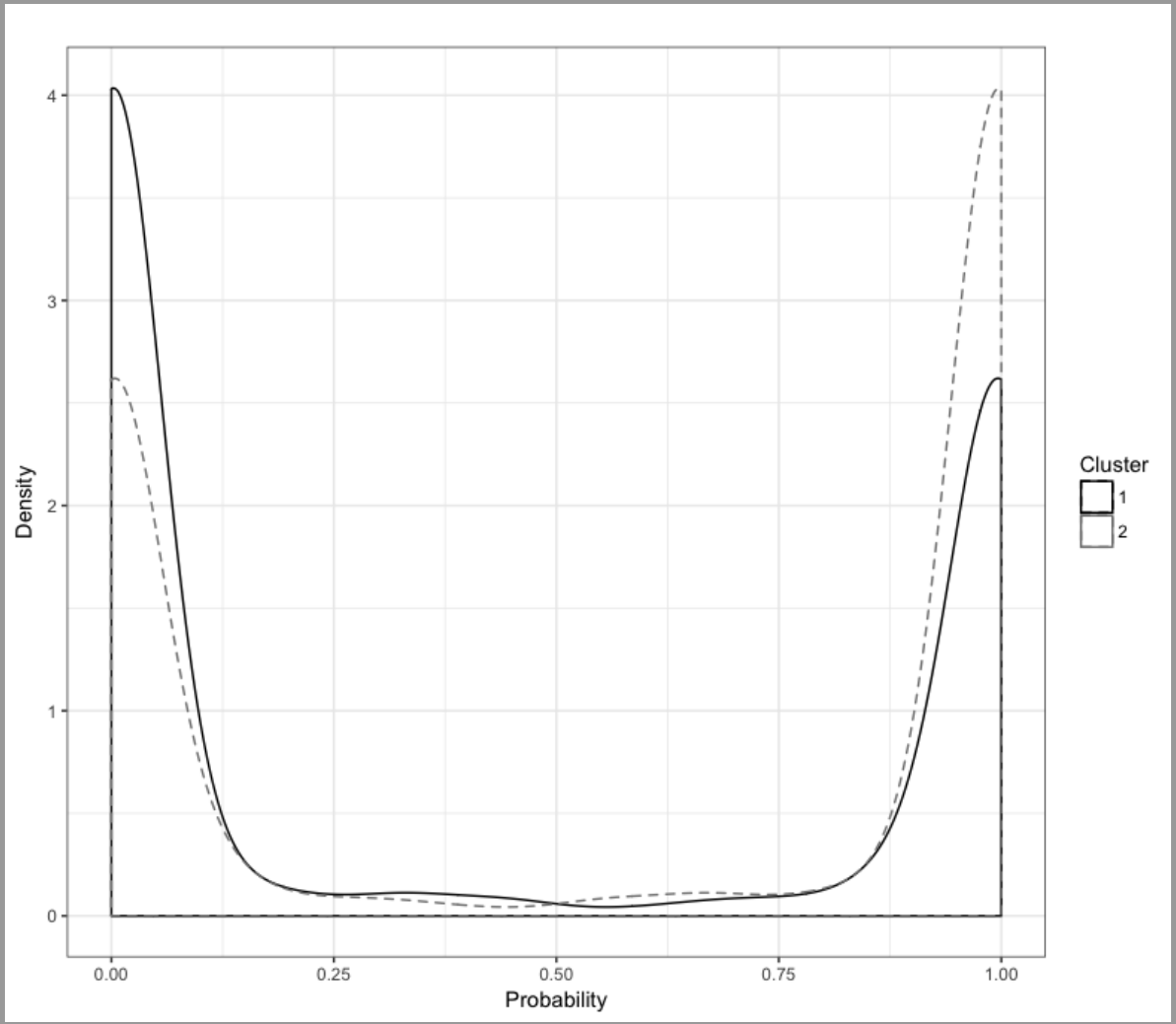
```
mixmod <- stepFlexmix(cbind(migrate, 1 - migrate)
~ busshorter + dirt + modernpower + modernwater [...] +
education + familysize + gender | village,
 model = FLXMRglmfix(family = "binomial"),
 k = 2, nrep = 20, data = cfps)
```

Coefficients

Table 4: Pooled and Clustered Regressions

	Migration after 2012			
	Bivariate Association	Pooled Data	Cluster 0	Cluster 1
	(1)	(2)	(3)	(4)
Bus Closer	.208***	.194***	−.004	1.052***
	(.025)	(.063)	(.076)	(.141)

Posterior ζ Distributions



Other Examples

- Weidmann (2011). "Violence "from above" or "from below"? The Role of Ethnicity in Bosnia's Civil War." *The Journal of Politics* 73(4):1178–1190.
  - Uses MMs to identify when different explanations for ethnic violence are relevant (macro-plans for ethnostates vs. local ethnic frictions)
  - "Though rarely applied to social science problems, finite mixture models have been extensively researched in other disciplines such as finance or computational biology" (1187).

Other Examples

- Krook and O'Brien (2012). "All the President's Men? The Appointment of Female Cabinet Ministers Worldwide." *The Journal of Politics* 74(3):840–855.
  - Uses MMs to test different theories explaining gender make-up of cabinet (focus is on posterior probabilities ζ_i)

Other Examples

- Heinrich (2013). "When is Foreign Aid Selfish, When is it Selfless?." *The Journal of Politics* 75(2):422–435.
  - Uses "mixture-of-experts" model to identify conditions when different explanations of foreign aid behavior are relevant.

Potential Pitfalls

- Not, in itself, a causal identification strategy
- Model complexity vs. data limitation trade-off
- Bias towards more complex component models
- Component models still need to make theoretical sense

Other Thoughts

- How strong is the link between theory predicting variables and actual reality?
  - MHT: With enough covariates, even a random partition might be "explained" by something
- Which quantity of interest?
- When is this useful?

The Bayesian Approach

1. Sample Z_i given the current values of all parameters with the following probability,

$$\Pr\left(Z_i^{(t)} = m \mid \Theta^{(t-1)}, \Pi^{(t-1)}, \{Y_i, X_i\}_{i=1}^N\right) = \zeta_{i,m}^{(t-1)} \propto \pi_m^{(t-1)} f_m\left(Y_i \mid X_i, \theta_m^{(t-1)}\right), \quad (8)$$

for $i = 1, \ldots, N$ and $m = 1, \ldots, M$.

2. Given $Z_i^{(t)}$, sample all parameters.

(a) Given the subset of the data with $Z_i^{(t)} = m$, update θ_m using the standard MCMC algorithm for this particular model.

(b) Update π_m using the standard MCMC algorithm. For example, if the Dirichlet distribution is used as the prior distribution, then we have,

$$\left(\pi_1^{(t)}, \ldots, \pi_M^{(t)}\right) \sim \text{Dirichlet}\left(s_1 + \sum_{i=1}^N \mathbf{1}\{Z_i = 1\}, \ldots, s_M + \sum_{i=1}^N \mathbf{1}\{Z_i = M\}\right), \quad (9)$$