Theory Testing Using Mixtures and EM with our Buddies Imai and Tingley

Machine Learning: Module 1

Sean Norton; Simon Hoellerbauer

October 1, 2018

Rationale

Remember the general form of the mixture distribution where distributional forms can be different?

$$p(x) = \sum_{k=1}^{K} \pi_k p(x|z_i = k, \theta) \tag{1}$$

Well, today, we are going to treat the different mixing distributions as different statistical models that could have produced our observations (designated as x here, but soon we will use x for covariates and the more traditional—for us—y for outcomes).

Mixtures of Models: Data Generating Process

We assume the following data generating process:

$$Y_i|X_i, Z_i \sim f_{Z_i}(Y_i|X_i, \theta_{Z_i}) \tag{2}$$

Where Y_i denotes the outcome for observation i, X_I denotes the covariates for observation i, and Z_i denotes the model $1 - M$ that produced observation i. In other words, if we knew Z_i, we could tell you from which model Y_I was drawn. Note that θ has a Z_i subscript—a different set of θ's can belong to each model (or mixing distribution m). This also implies, of course, that the covariate set X_i belong to observation i does not have to be the same for each model.

Observed-Data Likelihood

As usual, we assume conditional independence given covariates and the latent variable. This allows us to form the following observed-data likelihood function (where Z has been integrated out):

$$L_{obs}(\Theta, \Pi | \{X_i, Y_i\}_{i=1}^{N}) = \prod_{i=1}^{N} \left\{ \sum_{m=1}^{M} \pi_m f_m(Y_i | X_i, \theta_m) \right\} \tag{3}$$

Where the sum of all π_m is 1 (because it is a distribution), Θ is the set of all θ's for all M theories (distributions/models), and Π is the set of all π's for all M theories (distributions/models).
This equation is important because it shows us that we can think of this model as weighting the various theories for each observation i

Important Excursus: *Theory-Predicting Variables*

Being astute statisticians and wise mathematicians and amazing social scientists, it may come as no surprise, given the previous slide, that we can try to see if any variables influence the probability that an observation belongs to a certain model, which Imai and Tingley name a *theory-predicting variables* and term as W_i. We then make the probability that the observation i a function of W_i and some parameters ϕ:

$$Pr(Z_i = m|W_i) = \pi_m(W_i, \phi_m) \tag{4}$$

This is really, really cool, y'all.

Let's Do Some EM-ing: Complete-Data Log-Likelihood

We need the complete data log-likelihood to make applying the EM algorithm as straightforward as it can be:

$$l_{com}(\Theta, \Pi | \{X_i, Y_i, Z_i\}_{i=1}^{N}) = \sum_{i=1}^{N} \sum_{m=i}^{M} \left(\log \pi_m + \log f_m(Y_i | X_i, \theta_m) \right)^{z_m} \quad (5)$$

Note that Imai and Tingley put the indicator function out in front, but it is the same thing as putting it in the exponent.

Let's Do Some EM-ing: The E-Step

First, we must initialize all our parameters, so pick some Θ^{old} and some Π^{old}.

In this version of the E-step, we take the expectation of the complete-data log-likelihood w.r.t the posterior probabilities of the latent variable Z_i, such that

$$\mathbb{E}_{Z_i|\Theta^{old},\Pi^{old},\{X_i,Y_i,Z_i\}_{i=1}^N} = \sum_{i=1}^{N}\sum_{m=i}^{M} \zeta_{i,m}^{old}\Big(\log \pi_m + \log f_m(Y_i|X_i,\theta_m)\Big) \quad (6)$$

Where

$$\zeta_{i,m}^{old} = Pr(Z_i = m, |\Theta^{old}, \Pi^{old}, \{X_i, Y_i, Z_i\}_{i=1}^N) = \frac{\pi_m^{old} f_m(Y_i|X_i,\theta_m)}{\sum_{m'=1}^{M} \pi_{m'}^{old} f_{m'}(Y_i|X_i,\theta_{m'})} \quad (7)$$

What does this remind you of?

Let's Do Some EM-ing: E-Step—Responsibilities and the Lower Bound

That's right those are the responsibilities! As Imai and Tingley point out, this "represents the posterior probability ... that observation i arises from statistical model implied by theory m" (pg. 222).

If we remember back to our discussion last week, this is just the $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old})$ that represented the function $q(Z)$ that maximized the $\mathcal{L}(q, \theta)$, where in our case, $\mathbf{X}$ includes both covariates and outcome. Similarly we can see that the equation of the previous page results when we maximize the lower bound and drop the *entropy* portion of this new function (because there are only θ^{old}'s)

Let's Do Some EM-ing: The M-Step

Using these responsibilities values, we then maximize the expectation (which remember, is just a version of the new lower bound for the distribution we are trying to maximize) shown previously w.r.t to the parameters. This allows us to generate Θ^{new} and Π^{new}, which are the free parameters given our set-up. Because we have not specified the form of the mixing distribution (that is, the model we assume/use for the various theories we are comparing), we can not define most of these new values yet. Because we do know, however, that we are working with mixtures, we can specify the update for the mixing parameters:

$$\pi_m^{new} = \frac{1}{N} \sum_{i=1}^{N} \zeta_{i,m}^{old} \tag{8}$$

This shows us that π_m can be interpreted as how well theory m performs overall, while $\zeta_{i,m}$ represents how well observation i fits with theory m.

We then repeat the E- and M-steps until convergence.

Important Excursus: *Theory-Predicting Variables, Pt. 2*

If we model π_m with covariates, then updating π_m will not look like it did in the previous equation. However, because of the relationship between π_m and $\zeta_{i,m}$ across the E- and M-steps, the gist and interpretation of both π_m and $\zeta_{i,m}$ is maintained.

The Really Cool Stuff Begins: How to Identify Observations Consistent with a Theory, Pt. 1

- If we want to decide to which theory an observation belongs (which, to be clear, we do not have to in order for the approach to make sense), we can use $\zeta_{i,m}$ — Why?

The Really Cool Stuff Begins: How to Identify Observations Consistent with a Theory, Pt. 1

- If we want to decide to which theory an observation belongs (which, to be clear, we do not have to in order for the approach to make sense), we can use $\zeta_{i,m}$ — Why?
- This is the posterior probability that observation i fits with theory m.
- We get this value as a byproduct of the EM algorithm
- But how to decide what threshold $lamba_m$, greater than which we accept observation i as belonging to theory m to use as a cut-off?
- If we use naive $lamba_m$, we run into problem of "multiple testing problem," where the probability of false positives increases with m.

The Really Cool Stuff Begins: How to Identify Observations Consistent with a Theory, Pt. 2

- So what to do?

The Really Cool Stuff Begins: How to Identify Observations Consistent with a Theory, Pt. 2

- So what to do?
- We want to limit number of false positives; therefore, we will want to "choose the smallest value of λ_m ... while ensuring that the posterior expected value of false discovery rate on the resulting list does no exceed a certain threshold α_m"(pg. 224).

$$\lambda_m^* = \inf\left\{ \lambda_m : \frac{\sum_{i=1}^{N}(1-\zeta_{i,m})\mathbb{I}(\zeta_{i,m} \geq \lambda_m)}{\sum_{i=1}^{N}\mathbb{I}(\zeta_{i,m} \geq \lambda_m) + \prod_{i=1}^{N}\mathbb{I}(\zeta_{i,m} < \lambda_m)} \leq \alpha_m \right\} \quad (9)$$

- Only useful if you think an observation is produced by one model and only one model.

What This Is All About: Comparing Rival Theories

- π_m allows to estimate to what extent the observations in a data set are consistent with one theory or another
- Can also be seen as the average individual observation membership in each theory m.
- Finally, if we do believe that every observation fits only with one theory, we can use the previously specified λ_m's to determine which observations are statistically significantly identified with one theory or another, and then use the number of observations that fit in this way with each theory to determine which theory performs better overall.

Simulation Performance

FIGURE 1 Estimated Population Proportion of Observations Consistent with Model 1 (four left plots) and Classification Success Rates (four right plots) in the Two-Theory Mixture Model Simulation Study

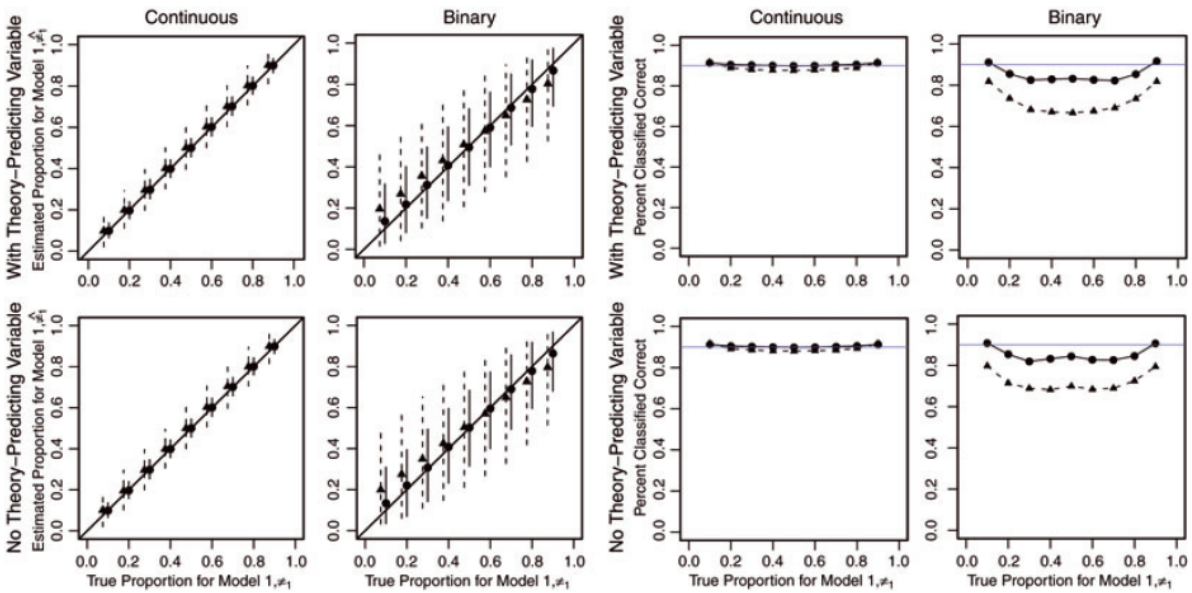


Figure: Testing Theory Testing with Simulations

Source: Imai and Tingley (2012), pg. 228

Empirical Tests: Ricardo-Viner vs Stolper-Samuelson

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				

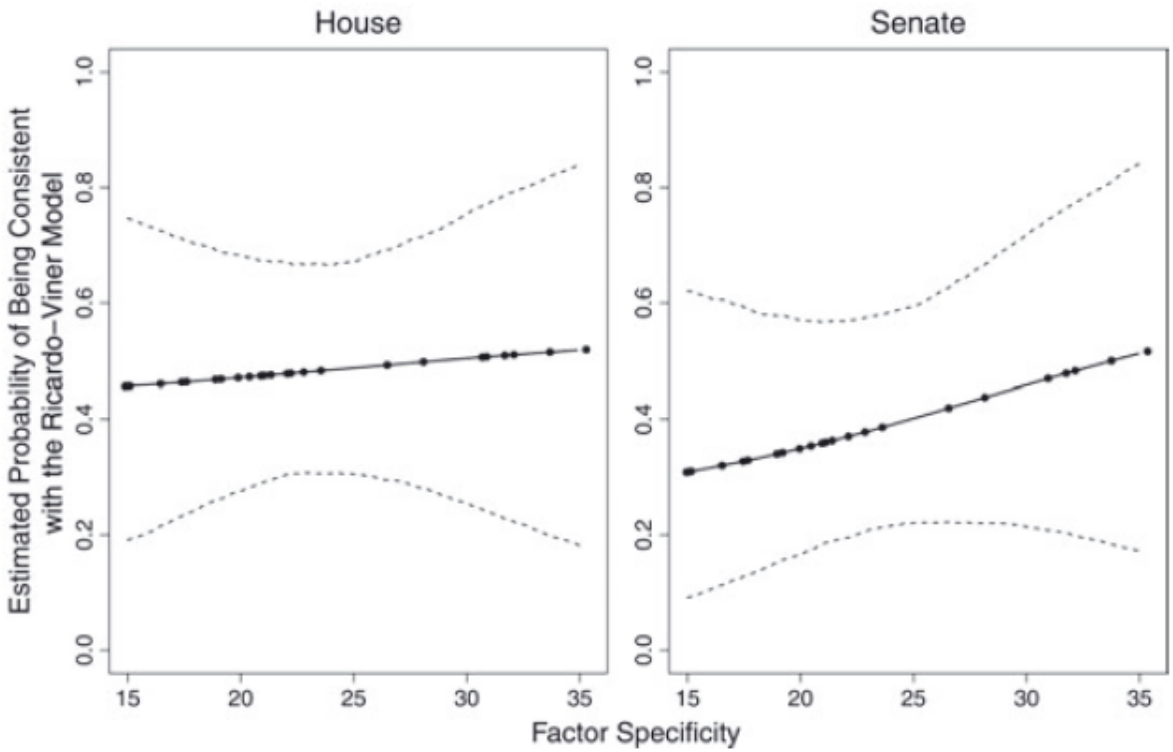
Each model is the logistic regression with model intercepts omitted in order to ease presentation. The first set of models uses the proposed mixture model approach with bill clustering. The second set is based on a "garbage-can" regression that uses all variables from both Stolper-Samuelson and Ricardo-Viner models as well as bill fixed effects.

Figure: Coefficients for Trade Preference Models

Source: Imai and Tingley (2012), pg. 232

Empirical Tests: Ricardo-Viner vs Stolper-Samuelson

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



Note: Solid line is the estimated probability with actual observations indicated by sold circles, and dashed lines represent 95% confidence intervals based on the Monte Carlo approximation. Although there is a considerable degree of uncertainty due to the small number of bills, the positive slopes in the House (the left panel) and Senate (the right panel) are consistent with the hypothesis that the Ricardo-Viner model rather than the Stolper-Samuelson model is supported when the level of factor specificity is high.

Empirical Tests: Democratic Peace

TABLE 2 Parameter Estimates and Their Standard Errors for the Affinity and Norms Models from the Mixture Model and Standard Bivariate Probit Model Used in the Original Analysis by Huth and Allee (2002)

	Mixture Model		Huth and Allee	
	coef.	s.e.	coef.	s.e.
Affinity Model				
Challenger				
Political Similarity	0.005	0.005	0.005	0.005
Change in Political Similarity	0.003	0.009	0.003	0.007
Defender				
Political Similarity	−0.204	0.265	−0.233	0.260
Change in Political Similarity	0.784	1.419	0.929	1.330
Norms Model				
Challenger				
Nonviolent Norms	0.004	0.002	0.004	0.002
Stalemate	0.015	0.027	0.014	0.029
Nonviolent Norms × Stalemate	−0.003	0.003	−0.002	0.003
Nonviolent Norms × Military Advantage	−0.003	0.002	−0.003	0.002
Defender				
Nonviolent Norms	0.047	0.028	0.073	0.023
Stalemate	0.216	0.656	0.283	0.531
Nonviolent Norms × Stalemate	−0.004	0.098	−0.001	0.051
Nonviolent Norms × Military Advantage	−0.016	0.026	−0.025	0.021

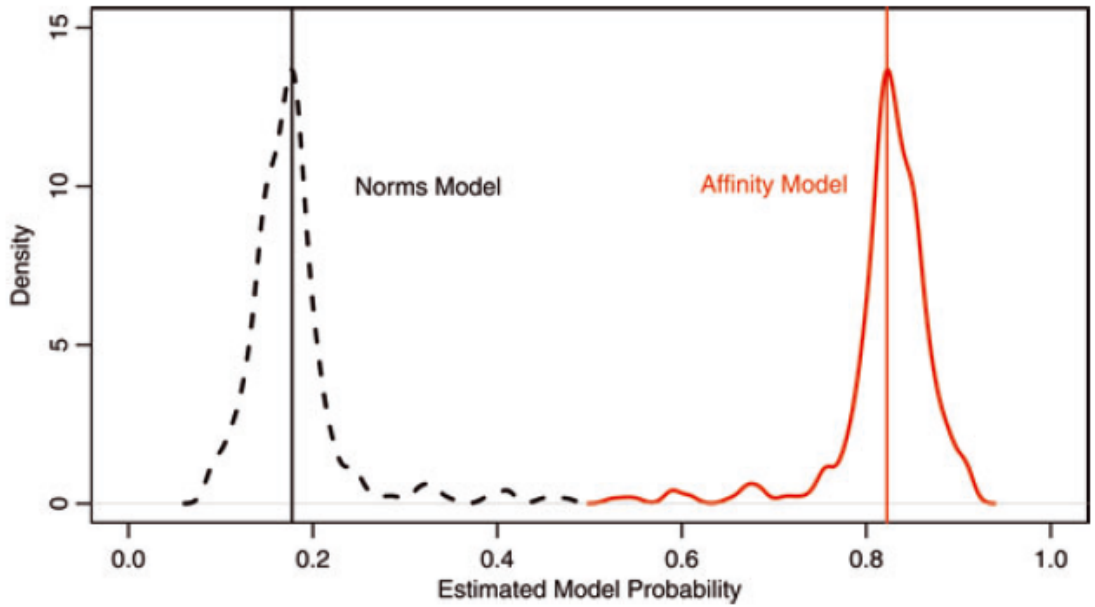
Note: Each model also contains a set of control variables, which are omitted from this table. Standard errors are based upon the nonparametric bootstrap. The two methods give similar results for both models.

Figure: Democratic Peace Theories Coefficients

Source: Imai and Tingley (2012), pg. 234

Empirical Tests: Democratic Peace

FIGURE 4 Smoothed Histograms of Estimated Probabilities That Each Observation Is Consistent with Each Competing Theory of Democratic Peace, $\hat{\zeta}_{i,m}$



Note: Solid vertical lines represent the estimated overall probability that observations are consistent with each model, $\hat{\pi}_m$. The affinity model receives the greatest support. The estimated probability for the accountability model is essentially zero for all observations.

Figure

Source: Imai and Tingley (2012), pg. 233

Discussion Questions

1. What is the difference between predictive and causal inference and why is it important for this paper? Why is it important for machine learning?
2. What is one of the key differences in terms of assumptions between the finite mixture model approach and other, more traditional model selection approaches?
3. How is this different from Bayesian model averaging?
4. What are some of the limitations of this approach?
5. Under what conditions do you think it would be appropriate to think a theory produces one observation deterministically?
6. What are some alternatives to model selection that Imai and Tingley did not discuss?
7. How convincing do you find their empirical tests—the Ricardo-Viner vs Stolper-Samuelson theories of trade policy preferences and the three theories of the democratic peace?

Limitations

- Only helps with predictive inference, not causal inference
- Approach breaks down when you try to compare too many theories (Imai and Tingley recommend using two to three)
- Could be identification issues across the full model that aren't there in submodels
- High correlations across predictors are possible, which can impact power
- It is possible for a theory with more predictors to be selected purely because it has more predictors—it is very important to build and specify carefully, based in theory.
- It can be difficult to find *theory-predicting variables*.

J Test

$$Y_i = (1 - \pi) f(X_i, \beta) + \pi g(X_i, \gamma) + \epsilon_i \tag{10}$$