Mixture Models

Hyunjoo Oh

12/5/2017

Imai & Tingley. 2012. “A Statitical Method for Empirical Testing of Competing Theories.” AJPS. 56(1): 218-236.

Why do we need this?

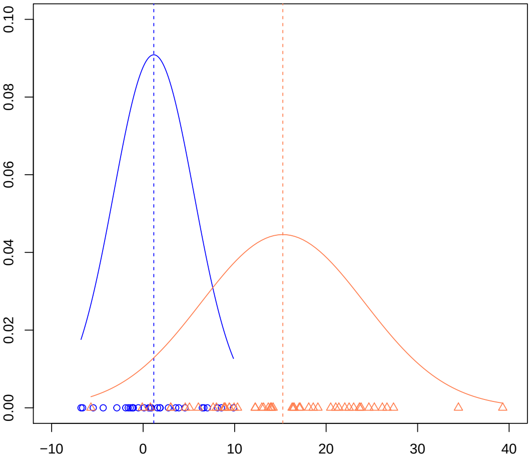
- We want to empirically test competiting theories
- How?
- Fitting a regression model with many explanatory variables that are derived from multiple theories?
- Achen (2005): atheoretical, “garbage-can regression”
- **Finite mixture models** provide a more effective method for comparative theory testing.

Basic idea

1. 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 statitical models under consideration.
2. In addition to the parameters of each model, researchers can estimate the probability that a specific observation is consistent with either of the competing theories.
3. These observation-specific probabilities can be averaged to serve as an overall performance measure for each model, thereby also achieving most of what standard model selection methods are designed to do.

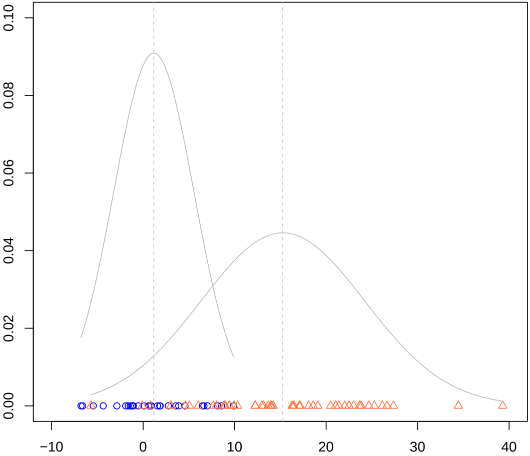
Make things simple: one dimension

- observations: $x_1, x_2, \cdots, x_n$
- $M \in \{\text{blue}, \text{orange}\}$, m=2, $\mathcal{N}(\mu_M, \sigma_M^2)$, unknown μ, σ^2
- Suppose we know which point came from which distribution.



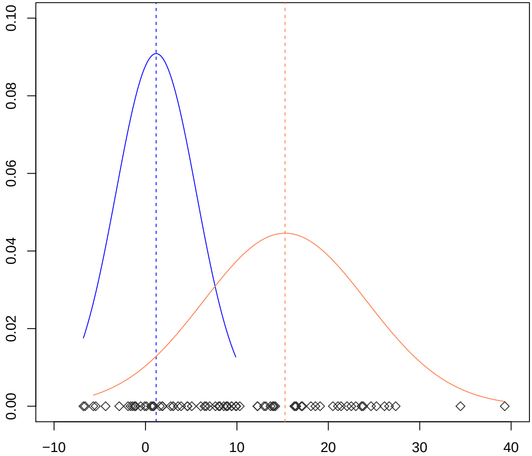
Make things simple: one dimension

- If we happen to know the parameters of the Gaussians
- Then, we can guess whether each point is more likely to be in orange or blue distribution.



Make things simple: one dimension

- What if we don't know the source?
- This is the situation when we apply mixture models



Example: Trade policy preferences (Hiscox, 2002)

- Legislative voting on trade bills in the US: factoral vs. sectoral cleavages.
- Stolper-Samuelson (SS) model: "owners of factors (land/capital) will favor trade liberalization."
- Richardo-Viner (RV) model: "those in export industries will favor liberalization whereas those in import industries should oppose it."

Example: Trade policy preferences (Hiscox, 2002)

- Hiscox (2002): how specific (or mobile) factors of production are to particular industries.
- By using J test, he finds that SS model is preferred during eras where specificity was low, while RV model is preferred in eras where specificity was high.
- How can we group the votes? The continuous measure of the factor specificity viariable does not provide natural breakpoints → **Let's use a finite mixture model**

1. 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 statitical models under consideration.

Model specification

- Goal: to measure the *relative* explanatory power of the competing theories
- A finite number of M different statistical models
- Assumption: each observation is generated either from one of the M statistical models, but we do not know a priori which model generates a specific observation.

Model specification

The DGP:

$$Y_i|X_i, Z_i \sim f_{Z_i}(Y_i|X_i,\theta_{Z_i}), \quad \text{for } i = 1, 2, \cdots, N$$

- $Z_i \in \{1, 2, \cdots, M\}$: the theory with which obervation i is consistent.
- Y_i: outcome variable
- X_i: covariates
- θ_{Z_i}: parameters of model Z_i

2. In addition to the parameters of each model, researchers can estimate the probability that a specific observation is consistent with either of the competing theories.

Model specification

Observed-data likelihood function:

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

- assuming the conditional independence across observations given the covariates (X_i) and the latent variable (Z_i).
- population proportion of observations generated by theory m (one measure of the overall performance of theory): $\pi_m = \Pr(Z_i = m)$
- set of all model parameters: $\Theta = \{\theta_m\}_{m=1}^M$
- set of all model probabilities: $\Pi = \{\pi_m\}_{m=1}^M$

Model specification

Probability that an observation is consistent with theory m:

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

- W_i: theory-predicting variables (W_i may overlap with X_i)
- ψ_m: a vector of unknown parameters
- multinomial logistic regression, multinomial probit model, and semiparametric multinomial logit model can be used.

Estimation and inference

Complete-data log-likelihood function:

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

- assuming that Z_i is observed
- $\{\cdot\}$: indicator function

3. These observation-specific probabilities can be averaged to serve as an overall performance measure for each model, thereby also achieving most of what standard model selection methods are designed to do.

Estimation and inference

- To obtain the MLE, the EM (Expectation-Maximization) algorithm can be applied.
- E-step: computing the conditional expecation of Z_i given the observed data and the values of parameters at the previous iteration $(t-1)$.
- M-step: maximizing the function of E-step
- Iterate E-step and M-step until convergence.

Estimation and inference

- **E-step**: computing conditional expection of Z_i

$$Q_{com}\left(\Theta, \Pi | \Theta^{(t-1)}, \Pi^{(t-1)}, \{X_i, Y_i, Z_i\}_{i=1}^{N}\right)$$

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

where $\zeta_{i,m}^{(t-1)} = \Pr\left(Z_i = m | \Theta^{(t-1)}, \Pi^{(t-1)}, \{X_i, Y_i\}_{i=1}^{N}\right)$

- **M-step**: maximizing the weighted (weight $= \zeta_{i,m}^{(t-1)}$) log-likelihoodfor each model ($f_m(y|x, \theta_m)$.

Estimation and inference

1. By applying EM algorithm, we get $\zeta_{i,m}^{(t-1)}$. ($\zeta_{i,m}$: measure of consistency between i and m)
2. The updated estimate of π_m can be obtained by averaging $\zeta_{i,m}^{(t-1)}$ across all observations.

- π_m: measure of the overall performance of theory m:

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

- When π_m is modeled as a function of X_i, then maximizing the weighted log-likelihood function will give the updated estimate of model parameters ψ_m

Bayesian inference

- MCMC algorithm with data augmentation where Z_i is sampled along with model parameters

1. Sample Z_i given the current values of all parameters

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

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

- Given the subset of the data with $Z_i^{(t)} = m$, update θ_m using the standard MCMC algorithm for this particular model.
- Update π_m using the standard MCMC algorithm. Dirichlet distribution is often used as the prior distribution for π_m.

Grouped observations

- What if we want want to assume that all observations of one group arise from the same statistical model implied by a particular theory?
- ex. Because they all share the same level of factor mobility (group i), all votes on a particular bill (multiple observations j collected from a particular bill) are assume to be generated by a single theory (consistent with one of the competing theories, $Z_i = m$).

$$Y_{ij}|X_{ij}, Z_i \sim f_{Z_i}(Y_{ij}|X_{ij}, \theta_{Z_i}), \quad \text{for } i = 1, \cdots, N, \;\; j = 1, \cdots, J_i$$

How to do it in R

- **flexmix**: A general framework for finite mixtures that uses the EM algorithm to obtain the MLE for a wide range of mixtures of regression models.
- Alternatively, **mixtools** (analyzing a variety of finite mixture models)

Example: Hiscox (2002)

- Hiscox. 2002. "Commerce, Coalitions, and Factor Mobility: Evidence from Congressional Votes on Trade Legislation." *APSR* 96(3): 593-608.
- Data: Roll-call voting (29 trade bills) in the US Senate
- Outcome: $Y_i \in \{0, 1\}$, $1 =$ voting against liberalization
- Theory-predicting variable W_i: national-level measure of factor specificity
- Covariates X_i:
- **SS theory**: profit (state-level measures of profit), employ (employment in manufacturing), farm (agricultural production)
- **RV theory**: import, export (import and export orientation of a state)

Example: Hiscox (2002)

- Estimating a mixture model with two logistic models
- Instead of using fixed effects for each logistic regression, we use a mixture model with clustering where all votes for a particular trade bill are assumed to be consistent with the same theory.
- model the mixing probability π (the population proportion of observations consistent with the **RV model**)

$$\begin{aligned} f_{SS}(Y_{ij}|X_{ij},\theta_{SS}) &= logit^{-1}(\beta_0 + \beta_1 \text{profit}_{ij} + \beta_2 \text{employ}_{ij} + \beta_3 \text{farm}_{ij}) \\ f_{RV}(Y_{ij}|X_{ij},\theta_{RV}) &= logit^{-1}(\gamma_0 + \gamma_1 \text{export}_{ij} + \gamma_2 \text{import}_{ij}) \\ \pi_{RV}(W_j,\phi_{RV}) &= logit^{-1}(\delta_0 + \delta \text{factor}_j) \\ &\text{where } i \text{ index votes, } j \text{ index bills} \end{aligned}$$

Example: Hiscox (2002)

```
library(flexmix)
Hiscox <- read.dta("HiscoxAPSR-Senate.dta",
                    convert.underscore=F)

# specify the explanatory variables for each model
# & specify nesting structure
model <- FLXMRglmfix(family = "binomial",
                # sets up 2 different models
                # non-nested; each has one component
                nested = list(k = c(1, 1),
                # models: SS vs. RV
                formula = c(~ profits + employ + farm,
                            ~ import + export)))
```

Example: Hiscox (2002)

```
# specify the outcome variable (vote),
# whether all votes for the same bill should be clustered,
# & the model for how specificity affects the mixture prob.
                                # clustering for each bill
result <- stepFlexmix(cbind(voted, 1 - voted) ~  1|bill,
            # theory predicting (concomitant) variable
            # factor = the sole covariate to model the
            #          mixing probabilities
          concomitant = FLXPmultinom(factor),
            # k = number of competing models
          k = 2, model = model,
            # the total number of EM algorithm runs
            # with different starting values
          data = Hiscox, nrep=20)
```

- **stepFlexmix()** estimates the model using the EM algorithm with different random starting values to avoid local maxima

Example: Hiscox (2002)

- Stolper-Samuelson Model

	Estimate	Std. Error	z value	Pr(>|z|)
profits	-5.69	1.19	-4.78	0.00
employ	19.79	2.59	7.65	0.00
farm	-1.27	0.43	-2.97	0.00
(Intercept)	0.02	0.21	0.09	0.93

- Richardo-Viner Model

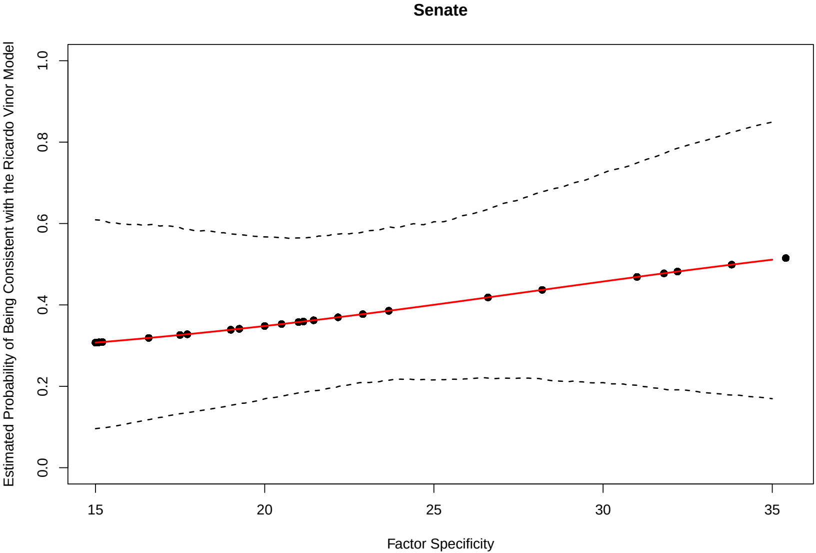
	Estimate	Std. Error	z value	Pr(>|z|)
import	2.53	0.80	3.18	0.00
export	-2.80	0.77	-3.65	0.00
(Intercept)	-0.83	0.13	-6.61	0.00

- Mixture probability

	Coefficient	Std. Error
(Intercept)	-1.60	1.62
factor	0.05	0.07

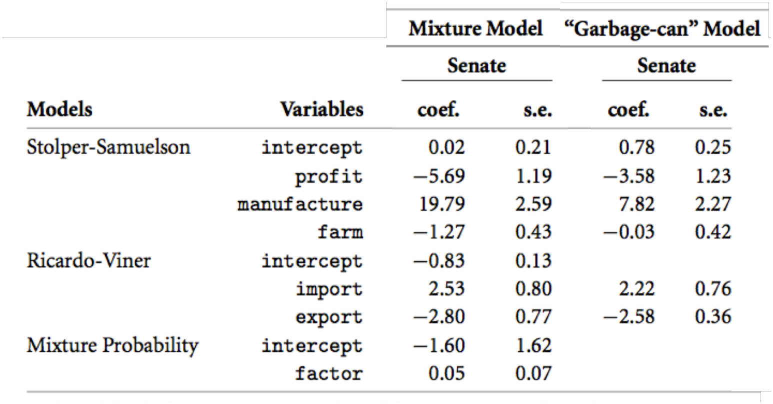
Example: Hiscox (2002): applicability of each theory

- The estimated population proportion of observations consistent with the RV model, $\hat{\pi}$, across factor specificity



Example: Hiscox (2002)

- Parameter estimates and their standard errors from the mixture model for the Senate



		Mixture Model		“Garbage-can” Model	
		Senate		Senate	
Models	**Variables**	**coef.**	**s.e.**	**coef.**	**s.e.**
Stolper-Samuelson	`intercept`	0.02	0.21	0.78	0.25
	`profit`	−5.69	1.19	−3.58	1.23
	`manufacture`	19.79	2.59	7.82	2.27
	`farm`	−1.27	0.43	−0.03	0.42
Ricardo-Viner	`intercept`	−0.83	0.13		
	`import`	2.53	0.80	2.22	0.76
	`export`	−2.80	0.77	−2.58	0.36
Mixture Probability	`intercept`	−1.60	1.62		
	`factor`	0.05	0.07		

- logistic regression with model intercepts omitted in order to ease presentation.

Potential pitfalls of finite mixture model

1. The method in itself does not solve endogeneity and other fundamental problems of causal inference.
2. Don't test too many competing theories at once. Overfitting can be a problem.
3. High correlations across predictors may reduce the statistical power of mixture models.
4. The conditions under which different theories are applicable must be directly derived from the underlying assumptions of each rival theory.