

Regression III

Lecture 13: Mixture Models

Dave Armstrong

University of Wisconsin – Milwaukee
Department of Political Science

e: armstrod@uwm.edu
w: www.quantoid.net/ICPSR.php

1 / 24

Goals of the lecture

- Explain the problem of testing multiple theories
- Introduce finite mixture models as a tool
- Example of mixture modeling using R
- When can mixture modeling fail?

2 / 24

Testing Multiple Theories

- The methods we talked about earlier are good at testing competing theories if
 1. One theory explains the relationship between x and y for all observations equally well
 2. All models are simplified, parametric approximations of the same truth $f(x)$
- AIC and BIC can help us choose models in this sense (multiple competing working hypotheses), but the non-nested *tests* can only accommodate one null and one alternative.
- If multiple theories are required to explain the relationships between x and y , then none of the methods we have proposed are particularly useful.

The discussion below builds on the discussion in Kosuke Imai and Dustin Tingley's paper, "A Statistical Method for Empirical Testing of Competing Theories." (currently forthcoming at the *American Journal of Political Science* and available from: <http://imai.princeton.edu/research/files/mixture.pdf>)

3 / 24

Finite Mixture Model (1)

- The finite mixture model proposes that there are M different statistical models.
 - Each model is implied by one of a set of competing theories of the same phenomenon (think of these as "working hypotheses" from information theory)
- Each observation is generated by one of the M competing models (or more likely a weighted combination of models).
- The goal is to measure the relative explanatory power of competing theories under consideration.

4 / 24

Formalization

Let $f_m(y|x, \theta_m)$ denote a model of outcome y implied by theory m

- x is a set of covariates from X
- θ_m a vector of parameters

Since we don't know m_i , the theory to which each observation (or set of observations) belongs, we instead try to estimate the following:

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

where Z_i is a latent variable representing the theory with which observation i is consistent.

So, the likelihood is:

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

- Here, $\pi_m = \Pr(Z_i = m)$ with $\sum_{m=1}^M \pi_m = 1$. π_m can be provide a measure of overall performance of theory m .

5 / 24

Theory-predicting Variables

It is possible to include variables (say W_i) that predict the theory to which observations belong.

- Thus, we can model the probability directly.

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

- ψ_m is a set of coefficients relating W_i to the probability of being in category m through a multinomial logit (or similar) model.
- W_i (if shown to be a powerful predictor of $\Pr(Z_i = m)$) can provide a sense of the circumstances under which various theories apply.

6 / 24

Estimation

Estimation can either proceed through an E-M algorithm to maximize the observed data likelihood or through Bayesian MCMC Simulation to explore the posterior distributions.

- In E-M, the E step computes the conditional expectation of the latent variable Z_i and the M step maximizes the likelihood conditional on the E-step values of Z_i .
- The Bayesian context proceeds in a similar way with block sampling of the Z_i given the other model parameters and then updating the other model parameters conditional on the last drawn values of Z_i .

7 / 24

Grouped Observations

It is possible (sometimes necessary) to think about observations as grouped (such as individuals over time) such that each observation for that individual is the same weighted mix of the theories under investigation:

$$Y_{ij}|X_{ij}, Z_i \sim f_{z_i}(Y_{ij}|X_{ij}, \theta_{z_i})$$

- Grouping has implications for the estimation of $\Pr(Z_i = m|W_i)$, specifically that with a small number of groups, ψ_m may not be precisely estimated.
- Grouping is required if you have a binary dependent variable as the model is not identified otherwise.

8 / 24

Calculating Quantities of Interest

Quantities of interest can be calculated easily for these types of models.

- The quantity of interest should be calculated for each model under consideration.
- Weighted combinations (with weights of π_m) of each quantity of interest can be used for plots, etc...
- Sampling variance can be incorporated with parametric bootstrap or bayesian posterior simulation .
 - Sampling variability in π_m should also be incorporated in these quantities.

9 / 24

Identifying the Observations Consistent with Theories

Not all observations will be strongly predicted by a single theory (i.e., have a high $\Pr(Z_i = m)$ for one m and a low value for the others). However, it is possible to figure out which observations are *statistically significantly consistent* with each theory in the following way.

- if $\zeta_{i,m}$ is The probability that observation i is a member of group m , then we want to pick λ_m , a threshold such that observation i is *ssc* with theory m if $\zeta_{i,m} > \lambda_m$.
- Since there are m hypotheses being tested simultaneously for each observation, we want to set the Type I error rate for the set of hypotheses (rather than for a single one) to $\alpha_m = 0.05$.

The following is the optimal value of λ_m

$$\lambda_m^* = \inf \left\{ \lambda_m : \frac{\sum_{i=1}^N (1 - \hat{\zeta}_{im}) \mathbf{1}\{\zeta_{i,m} \geq \lambda_m\}}{\sum_{i=1}^N \mathbf{1}\{\hat{\zeta}_{i,m} \geq \lambda_m\} + \prod_{i=1}^N \mathbf{1}\{\hat{\zeta}_{i,m} < \lambda_m\}} \leq \alpha_m \right\}$$

10 / 24

Identifying the Observations Consistent with Theories (2)

An omnibus threshold constant across all theories (λ) can also be obtained in a similar way:

$$\lambda^* = \inf \left\{ \lambda : \frac{\sum_{i=1}^N \sum_{m=1}^M (1 - \hat{\zeta}_{im}) \mathbf{1}\{\zeta_{i,m} \geq \lambda\}}{\sum_{i=1}^N \sum_{m=1}^M \mathbf{1}\{\hat{\zeta}_{i,m} \geq \lambda\} + \prod_{i=1}^N \prod_{m=1}^M \mathbf{1}\{\hat{\zeta}_{i,m} < \lambda\}} \leq \alpha_m \right\}$$

- The numerator gives the rate of false discovery across all of the lists
- The denominator gives the total number of observations across all of the lists

11 / 24

Measure of Overall Performance

There are two measures of overall performance of each theory that might be worth considering

1. The average probability that observations belong to category m : $\frac{1}{N} \sum_{i=1}^N \zeta_{i,m}$
2. The number of statistically significantly consistent observations for each theory.

12 / 24

Example Re: Democracy

There are two different sets of theories about how democracy relates to repression.

- One theory suggests that it is the electoral shadow of the future that keeps leaders in line.
- Another theory suggests that it is the set of checks and balances other institutions have on the executive that prohibit unilateral repressive action.

Often times these theories are estimated simultaneously, but the variables representing these are quite highly collinear. Here is the R-code to estimate the mixture model:

13 / 24

```
> options(useFancyQuotes=F)
> library(flexmix)
> library(foreign)
> dat <- na.omit(read.dta("~/Desktop/ICPSR_Slides/Collinearity/mixture_data.dta"))
> dat$gdppc10k <- dat$gdppc/10000
> model <- FLXMRglmfix(family = "gaussian", nested=list(k=c(1,1),
+ formula = c(~lgates + aclp + bnr + xrcomp + parcomp,
+ ~log_checks + polconiii + xconst + laworder + subfed)),
+ fixed = ~ gdppc10k + logpop + cwar + iwar)
> out <- stepFlexmix(rep1 ~ 1 | ccode, k=2, model=model, data=dat, nrep=20)

2 : * * * * *
```

14 / 24

Probability of Group Membership

The probability of being a member of each group can be calculated by taking the mean of the individual probabilities:

```
> probs <- out@posterior$scaled
> apply(probs, 2, mean)
```

```
[1] 0.417957 0.582043
```

```
> summary(out)
```

```
Call:
stepFlexmix(rep1 ~ 1 | ccode, model = model, data = dat, k =
  nrep = 20)
```

	prior	size	post>0	ratio
Comp.1	0.429	688	850	0.809
Comp.2	0.571	980	1117	0.877

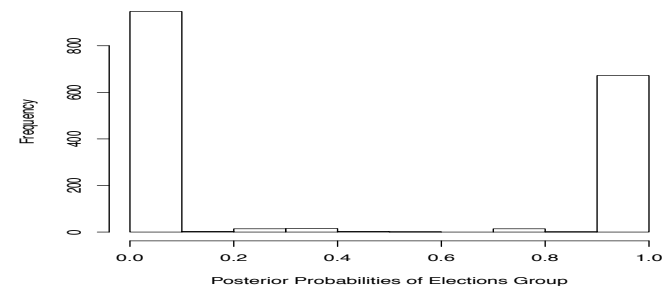
```
'log Lik.' -1897.65 (df=19)
AIC: 3833.3   BIC: 3936.269
```

15 / 24

Statistically Significant Consistency with Theory

We want to see which observations are significantly consistent with each theory. First, it's probably worth looking at the histogram of posterior probabilities.

```
> pdf("hsitprob.pdf", heigh=6, width=6)
> hist(probs[,1], xlab="Posterior Probabilities of Elections
> invisible(dev.off())
```



16 / 24

Identifying Observations

```
> source("mixtureTools.R")
> id.list <- IdentifyList(out@posterior$scaled, dat, dat$ccode, cluster=T, alpha=.05)
> id.list$caseID.1

 [1]  2  20  94 110 200 205 210 211 225 230 235 290 310 315 316 349 359 366 368
[20] 369 375 380 385 390 404 420 432 433 436 437 439 452 475 481 500 510 520 530
[39] 540 553 571 600 615 651 698 705 710 712 713 740 771 775 850 920

> id.list$caseID.2

 [1]  40  41  42  51  52  70  90  91  92  93  95 100 101 130 135 140 145 150 155
[20] 160 165 220 305 325 339 345 350 355 359 360 366 367 370 371 373 438 450 451
[39] 461 471 484 490 501 541 551 552 560 565 580 616 620 625 630 640 645 652 663
[58] 666 670 690 692 696 705 731 732 750 770 780 800 820 830 840 900 910
```

17 / 24

Looking at the Components

First, we have to refit the model and then summarizing the refitted object will show us the component coefficients:

```
> out.refit <- refit(out)
> summary(out.refit)

$Comp.1
      Estimate Std. Error z value Pr(>|z|)
gdppc10k -0.938394  0.040374 -23.2427 < 2.2e-16 ***
logpop   0.308759  0.016581  18.6213 < 2.2e-16 ***
cwar     0.242832  0.103318   2.3503  0.018756 *
iwar     1.276012  0.087635  14.5605 < 2.2e-16 ***
lgates   -0.109272  0.034207  -3.1944  0.001401 **
aclp     -0.235784  0.134458  -1.7536  0.079500 .
bnr      -0.614736  0.076156  -8.0721  6.912e-16 ***
xrcomp   0.048588  0.054043   0.8991  0.368620
parcomp  -0.141693  0.056904  -2.4901  0.012772 *
(Intercept) -3.034754  0.230865 -13.1451 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$Comp.2
      Estimate Std. Error z value Pr(>|z|)
gdppc10k -0.938394  0.040374 -23.2427 < 2.2e-16 ***
logpop   0.308759  0.016581  18.6213 < 2.2e-16 ***
cwar     0.242832  0.103318   2.3503  0.018756 *
iwar     1.276012  0.087635  14.5605 < 2.2e-16 ***
log_checks -0.084712  0.064988  -1.3035  0.192407
polconiii -0.568057  0.214240  -2.6515  0.008014 **
xconst   -0.122256  0.020699  -5.9064  3.496e-09 ***
laworder -0.054209  0.022098  -2.4531  0.014162 *
subfed   -0.064832  0.104152  -0.6225  0.533626
(Intercept) -1.150481  0.160850  -7.1525  8.520e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

18 / 24

Adding Concomitant Variables

We can add variables in to predict the theory to which observations belong.

- If we use a grouping variable (as we did above), the concomitant variables need to be constant within each group.
- These should be guided by theory as well

```
> out.a <- stepFlexmix(rep1 ~ 1 | ccode , k=2,
+ model=model, data=dat, nrep=20,
+ concomitant = FLXPmultinom( ~ parlresp_c))
2 : * * * * *
> out.a.refit <- refit(out.a)
```

19 / 24

Summarizing the Results

```
> summary(out.a.refit)

$Comp.1
      Estimate Std. Error z value Pr(>|z|)
gdppc10k -0.940451  0.038622 -24.3499 < 2.2e-16 ***
logpop   0.313097  0.014995  20.8798 < 2.2e-16 ***
cwar     0.223818  0.095751   2.3375  0.0194126 *
iwar     1.275715  0.087299  14.6132 < 2.2e-16 ***
lgates   -0.102122  0.031894  -3.2019  0.0013653 **
aclp     -0.241432  0.140270  -1.7212  0.0852161 .
bnr      -0.620276  0.072950  -8.5027 < 2.2e-16 ***
xrcomp   0.059075  0.052590   1.1233  0.2613096
parcomp  -0.155268  0.046725  -3.3230  0.0008904 ***
(Intercept) -3.042436  0.215808 -14.0979 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$Comp.2
      Estimate Std. Error z value Pr(>|z|)
gdppc10k -0.940451  0.038622 -24.3499 < 2.2e-16 ***
logpop   0.313097  0.014995  20.8798 < 2.2e-16 ***
cwar     0.223818  0.095751   2.3375  0.019413 *
iwar     1.275715  0.087299  14.6132 < 2.2e-16 ***
log_checks -0.079940  0.063392  -1.2611  0.207288
polconiii -0.562600  0.211472  -2.6604  0.007805 **
xconst   -0.122089  0.020075  -6.0818  1.189e-09 ***
laworder -0.055720  0.021798  -2.5562  0.010583 *
subfed   -0.068404  0.104666  -0.6535  0.513407
(Intercept) -1.185614  0.152271  -7.7862  6.904e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

20 / 24

Concomitant Variable Output

```
> out.a.refit@concomitant
$Comp.2
      Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.37866   0.35108  1.0785  0.2808
parlresp_cComplete -0.80307   0.51434 -1.5613  0.1184
parlresp_cIncomplete -0.47378   0.63924 -0.7412  0.4586
parlresp_cIrrelevant  0.49948   0.50378  0.9915  0.3215
```

21 / 24

Distribution Free Test of Two Models

We could think about what would happen if we did the distribution-free test of the two different theories:

```
> mod1 <- lm(rep1 ~lgates + aclp + bnr + xrcomp + parcomp +
+ gdppc10k + logpop + cwar + iwar, data=dat, y=T)
> mod2 <- lm(rep1 ~log_checks + polconiii + xconst + laworder +
+ subfed + gdppc10k + logpop + cwar + iwar, data=dat, y=T)
> smod1 <- summary(mod1)
> smod2 <- summary(mod2)
> il1 <- log(dnorm(mod1$y, mod1$fitted, smod1$sigma))
> il2 <- log(dnorm(mod2$y, mod2$fitted, smod2$sigma))
> cuts <- qbinom(c(.05,.95), size=length(mod1$y), prob=.5)
> d <- sum(il1 > il2)
> 1-pbinom(d, length(il1), .5)

[1] 0.00766228
```

22 / 24

Comparison

- Note that the distribution free test suggests strong evidence in favor of the elections model (model 1).
- The mixture model identifies that both theories are useful, though the posterior probability of being in the second model is better for a larger proportion of the observations.
- The mixture model forces the observations to be grouped on countries.

23 / 24

Conclusion

- Mixture models are a useful way to test competing theories when we don't assume one theory can explain all observations equally well.
- These are very flexible tools in practice that permit users to estimate many different mixture regression models (including GLMs, GAMs, and mixed-effect models).
- Mixture Models provide information on what theories are useful for which observations and permit the incorporation of information into the model to help predict to which theory observations belong.

24 / 24