Title

stata.com

fmm intro — Introduction to finite mixture models

Description

Finite mixture models (FMMs) are used to classify observations, to adjust for clustering, and to model unobserved heterogeneity. In finite mixture modeling, the observed data are assumed to belong to unobserved subpopulations called classes, and mixtures of probability densities or regression models are used to model the outcome of interest. After fitting the model, class membership probabilities can also be predicted for each observation. This entry discusses some fundamental and theoretical aspects of FMMs and illustrates these aspects with a worked example.

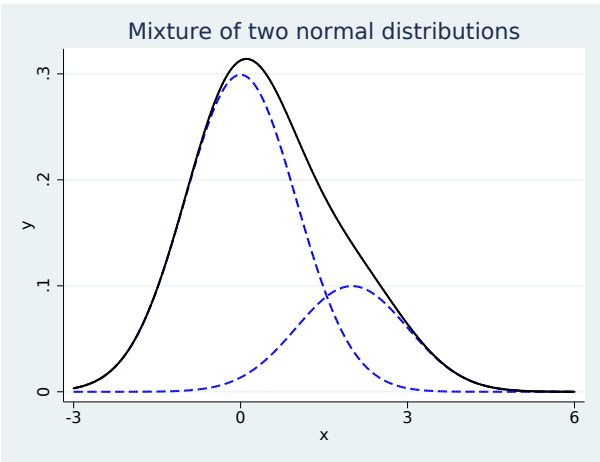
Remarks and examples

stata.com

Remarks are presented under the following headings:

Introduction

The main concept in finite mixture modeling is that the observed data come from distinct, but unobserved, subpopulations. To illustrate, we plot the observed distribution of a whole population (solid line) and the unobserved densities of two underlying subpopulations (dashed lines).



The observed distribution looks approximately normal, with a slight asymmetry because of more values falling above zero than below. This asymmetry occurs because the distribution is a mixture of two normal densities; the right-hand density skews the distribution to the right. We can use FMMs to estimate the means and variances of the two underlying densities along with their proportions in the overall population.

More generally, we can use FMMs to model mixtures containing any number of subpopulations, and the subpopulation-specific models need not be limited to a mixture of normal densities. FMMs allow mixtures of linear and generalized linear regression models, including models for binary, ordinal, nominal, and count responses, and allow the inclusion of covariates with subpopulation-specific effects. We can also make inferences about each subpopulation and classify individual observations into a subpopulation.

Because of their flexibility, FMMs have been used extensively in various fields to classify observations, to adjust for clustering, and to model unobserved heterogeneity. Mixtures of normal densities with equal variances can be used to approximate any arbitrary continuous distribution, which makes FMMs a popular tool to model multimodal, skewed, or asymmetrical data. A mixture of regression models can be used to model phenomena such as clustering of Internet traffic (Jorgensen 2004), demand for medical care (Deb and Trivedi 1997), disease risk (Schlattmann, Dietz, and Böhning 1996), and perceived consumer risk (Wedel and DeSarbo 1993). A mixture of a count model and a degenerate point mass distribution is often used for modeling zero-inflated and truncated count outcomes; see, for example, Jones et al. (2013, chap. 11). McLachlan and Peel (2000) and Frühwirth-Schnatter (2006) provide a comprehensive treatment of finite mixture modeling.

From a broader statistical perspective, FMMs are related to latent class analysis (LCA) models; both are used to identify classes using information from manifest (observed) variables. The difference is that FMMs allow parameters in a regression model for a single dependent variable to differ across classes while traditional LCA fits intercept-only models to multiple dependent variables. FMM is also a subset of structural equation modeling (SEM) where the latent variable is assumed to be categorical; see [SEM] **intro 1**, [SEM] **intro 2**, [SEM] **gsem**, and Skrondal and Rabe-Hesketh (2004, chap. 3) for a theoretical discussion. If your latent variable is continuous and your manifest variables are discrete, you can use item response theory models; see [IRT] **irt**. If both your latent variable and manifest variables are continuous, you can fit a structural equation model; see [SEM] **sem**.

Throughout this manual, we use the terms "class", "group", "type", or "component" to refer to an unobserved subpopulation. We use the terms "class probability" or "component probability" to refer to the probability of belonging to a given component in the mixture. Class probabilities are also referred to in the literature as "mixing weights" or "mixing proportions".

Finite mixture models

FMMs are probabilistic models that combine two or more density functions. In an FMM, the observed responses $\boldsymbol{y}$ are assumed to come from g distinct classes $f_1, f_2, \ldots, f_g$ in proportions $\pi_1, \pi_2, \ldots, \pi_g$. In its simplest form, we can write the density of a g-component mixture model as

$$f(\boldsymbol{y}) = \sum_{i=1}^{g} \pi_i f_i(\boldsymbol{y}|\boldsymbol{x}'\boldsymbol{\beta}_i)$$

where π_i is the probability for the ith class, $0 \leq \pi_i \leq 1$ and $\sum \pi_i = 1$, and $f_i(\cdot)$ is the conditional probability density function for the observed response in the ith class model.

fmm uses the multinomial logistic distribution to model the probabilities for the latent classes. The probability for the ith latent class is given by

$$\pi_i = \frac{\exp(\gamma_i)}{\sum_{j=1}^{g} \exp(\gamma_j)}$$

where γ_i is the linear prediction for the ith latent class. By default, the first latent class is the base level so that $\gamma_1 = 0$ and $\exp(\gamma_1) = 1$.

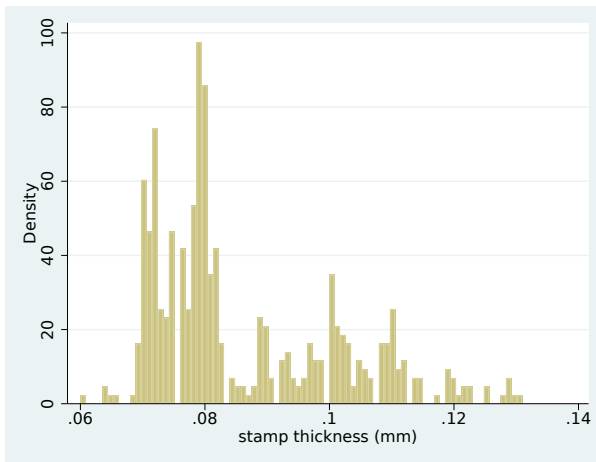
The likelihood is computed as the sum of the probability-weighted conditional likelihood from each latent class; see *Methods and formulas* in [FMM] **fmm** for details.

Mixture of normal distributions—FMM by example

The 1872 Hidalgo stamp of Mexico was printed on different paper types, which was typical of stamps of that era. For collectors, a stamp from a printing that used thicker paper is more valuable. We can use an FMM to predict the probability that a stamp is from a printing that used thick paper.

stamp.dta contains data on 485 measurements of stamp thickness, recorded to a thousandth of a millimeter. Here we plot the histogram of the measurements.

```
. use http://www.stata-press.com/data/r15/stamp
(1872 Hidalgo stamp of Mexico)
. histogram thickness, bin(80)
(bin=80, start=.06, width=.0008875)
```



At a minimum, the histogram suggests bimodality in the data, but we follow Izenman and Sommer (1988) and fit a mixture of three normal distributions to the data, each with its own mean and variance. We also estimate the proportion that each distribution contributes to the overall density. You can think of the three distributions as representing three different types of paper (thick, medium, thin) that the stamps were printed on. More specifically, our model is

$$f(\boldsymbol{y}) = \pi_1 N(\mu_1, \sigma_1^2) + \pi_2 N(\mu_2, \sigma_2^2) + \pi_3 N(\mu_3, \sigma_3^2)$$

The probability of being in each class is estimated using multinomial logistic regression

$$\pi_1 = \frac{1}{1 + \exp(\gamma_2) + \exp(\gamma_3)}$$

$$\pi_2 = \frac{\exp(\gamma_2)}{1 + \exp(\gamma_2) + \exp(\gamma_3)}$$

$$\pi_3 = \frac{\exp(\gamma_3)}{1 + \exp(\gamma_2) + \exp(\gamma_3)}$$

where the γ_i are intercepts in the multinomial logit model. By default, the first class is treated as the base, so $\gamma_1 = 0$.

To fit this model, we type

```
. fmm 3: regress thickness
```

We type `fmm 3:` because we have a mixture of three components. We type `regress thickness` to tell `fmm` to fit a linear regression model for each component. With no covariates, `regress` reduces to estimating the mean and variance of a Gaussian (normal) density for each component.

The result of typing our estimation command is

```
. fmm 3: regress thickness
Fitting class model:
Iteration 0:   (class) log likelihood =  -532.8249
Iteration 1:   (class) log likelihood =  -532.8249
Fitting outcome model:
Iteration 0:   (outcome) log likelihood =  1949.1228
Iteration 1:   (outcome) log likelihood =  1949.1228
Refining starting values:
Iteration 0:   (EM) log likelihood =  1396.8814
Iteration 1:   (EM) log likelihood =  1404.8995
Iteration 2:   (EM) log likelihood =  1412.4626
Iteration 3:   (EM) log likelihood =  1416.9678
Iteration 4:   (EM) log likelihood =  1419.0044
Iteration 5:   (EM) log likelihood =  1419.0582
Iteration 6:   (EM) log likelihood =  1417.9719
Iteration 7:   (EM) log likelihood =  1416.4213
Iteration 8:   (EM) log likelihood =  1414.8176
Iteration 9:   (EM) log likelihood =  1413.3462
Iteration 10:  (EM) log likelihood =  1412.0695
Iteration 11:  (EM) log likelihood =   1410.992
Iteration 12:  (EM) log likelihood =  1410.0961
Iteration 13:  (EM) log likelihood =  1409.3574
Iteration 14:  (EM) log likelihood =  1408.7518
Iteration 15:  (EM) log likelihood =  1408.2578
Iteration 16:  (EM) log likelihood =  1407.8564
Iteration 17:  (EM) log likelihood =  1407.5315
Iteration 18:  (EM) log likelihood =  1407.2694
Iteration 19:  (EM) log likelihood =  1407.0695
Iteration 20:  (EM) log likelihood =  1406.9013
Note: EM algorithm reached maximum iterations.
Fitting full model:
Iteration 0:   log likelihood =  1516.5252
Iteration 1:   log likelihood =  1517.1348  (not concave)
```

```
Iteration 2:   log likelihood =  1517.8203  (not concave)
Iteration 3:   log likelihood =   1518.153
Iteration 4:   log likelihood =  1518.6491
Iteration 5:   log likelihood =  1518.8474
Iteration 6:   log likelihood =  1518.8484
Iteration 7:   log likelihood =  1518.8484
Finite mixture model                            Number of obs     =        485
Log likelihood =  1518.8484
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
1.Class	(base outcome)					
2.Class						
_cons	.6410696	.1625089	3.94	0.000	.3225581	.9595812
3.Class						
_cons	.8101538	.1493673	5.42	0.000	.5173992	1.102908

```
Class          : 1
Response       : thickness
Model          : regress
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
thickness						
_cons	.0712183	.0002011	354.20	0.000	.0708242	.0716124
var(e.thic~s)	1.71e-06	4.49e-07			1.02e-06	2.86e-06

```
Class          : 2
Response       : thickness
Model          : regress
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
thickness						
_cons	.0786016	.0002496	314.86	0.000	.0781123	.0790909
var(e.thic~s)	5.74e-06	9.98e-07			4.08e-06	8.07e-06

```
Class          : 3
Response       : thickness
Model          : regress
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
thickness						
_cons	.0988789	.0012583	78.58	0.000	.0964127	.1013451
var(e.thic~s)	.0001967	.0000223			.0001575	.0002456

The output shows four iteration logs. The first three are for models that are fit to obtain starting values. Finding good starting values is often challenging for mixture models. `fmm` provides a variety of options for specifying and computing starting values; see *Options* in [FMM] **fmm** for more information.

The first output table presents the estimated class probabilities on a multinomial logistic scale. We can transform these estimates into probabilities as follows:

$$\pi_1 = \frac{1}{1 + \exp(0.64) + \exp(0.81)} \approx 0.19$$

$$\pi_2 = \frac{\exp(0.64)}{1 + \exp(0.64) + \exp(0.81)} \approx 0.37$$

$$\pi_3 = \frac{\exp(0.81)}{1 + \exp(0.64) + \exp(0.81)} \approx 0.44$$

More conveniently, we can use the `estat lcprob` command, which calculates these probabilities and the associated standard errors and confidence intervals; see [FMM] **estat lcprob**.

```
. estat lcprob
Latent class marginal probabilities             Number of obs     =        485
```

	Margin	Delta-method Std. Err.	[95% Conf.	Interval]
Class				
1	.1942968	.0221242	.1545535	.2413428
2	.3688746	.0286318	.3147305	.4265356
3	.4368286	.027885	.383149	.49203

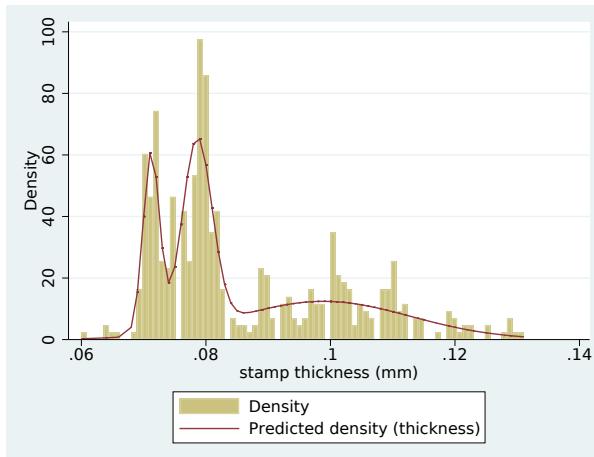
The three remaining tables of the `fmm` output show the estimated means and variances of each normal distribution.

The resulting mixture density, with maximum likelihood estimates of means, variances, and class probabilities, is given by

$$0.19 \times N(0.071, 0.0000017) + 0.37 \times N(0.079, 0.0000057) + 0.44 \times N(0.099, 0.0001967)$$

This equation gives the predicted density of stamp thickness, and we can plot it against the empirical distribution of stamp thickness as follows:

```
. predict den, density marginal
. histogram thickness, bin(80) addplot(line den thickness)
(bin=80, start=.06, width=.0008875)
```



We see that the first two components with small variances model the left-hand side of the empirical distribution, whereas the third component with much larger variance covers the long tail on the right-hand side of the empirical distribution.

We can use the predictions of the posterior probability of class membership to evaluate the probability of being in each class for each stamp. For the first stamp in our dataset, the probability of being in class 3, the thick paper type, is 1.

```
. predict pr*, classposteriorpr
. format %4.3f pr*
. list thickness pr* in 1, abbreviate(10)
```

	thickness	pr1	pr2	pr3
1.	.06	0.000	0.000	1.000

Because there are no covariates in the model, the posterior probabilities are the same for any stamp with a given thickness and are as follows.

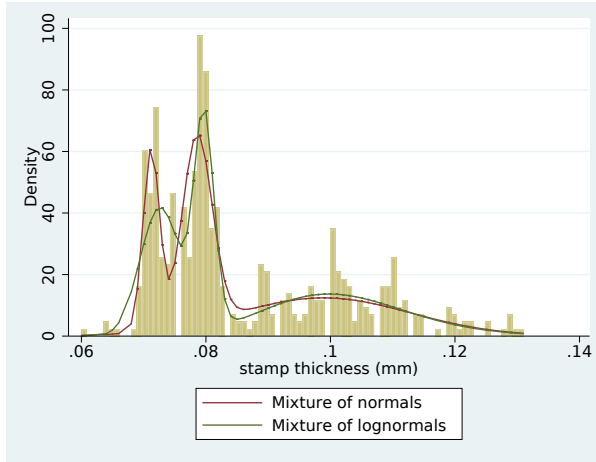
thickness	pr1	pr2	pr3
.06	0	0	1
.064	0	0	1
.065	.001	0	.999
.066	.026	0	.974
.068	.723	.001	.276
.069	.915	.001	.083
.07	.96	.002	.037
.071	.965	.007	.028
.072	.937	.026	.037
.073	.789	.134	.076
.074	.335	.525	.14
.075	.038	.838	.123
.076	.002	.91	.088
.077	0	.93	.07
.078	0	.936	.064
.079	0	.93	.07
.08	0	.912	.088
.081	0	.871	.129
.082	0	.788	.212
.083	0	.635	.365
.084	0	.406	.594
.085	0	.185	.815
.086	0	.06	.94
.087	0	.015	.985
.088	0	.003	.997
.089	0	.001	.999
.09-.131	0	0	1

The third mixture component has a relatively large variance, so the four thinnest measures end up being incorrectly classified into the thick paper type. Because stamp thickness cannot be negative, we can improve the model fit if we use a density with support only on the positive real line, such as the lognormal distribution.

```
. fmm 3: glm thickness, family(lognormal)
```

(*output omitted*)

We plot the predicted density from the mixture of normals with the density from the mixture of lognormals.



The mixture of lognormals correctly classifies the thinnest stamps into the thin paper type, which is confirmed by the predicted posterior probabilities.

thickness	pr1	pr2	pr3
.06	.889	0	.111
.064	.992	0	.008
.065	.994	0	.006
.066	.996	0	.004
.068	.997	0	.003
.069	.997	0	.003
.07	.996	0	.004
.071	.996	0	.004
.072	.995	0	.005
.073	.992	0	.008
.074	.987	.001	.011
.075	.965	.017	.018
.076	.849	.124	.027
.077	.532	.437	.031
.078	.233	.741	.026
.079	.102	.874	.024
.08	.056	.915	.028
.081	.041	.911	.048
.082	.039	.85	.111
.083	.042	.654	.305
.084	.034	.288	.678
.085	.017	.056	.928
.086	.006	.006	.988
.087	.002	0	.998
.088	.001	0	.999
.89-.131	0	0	1

Beyond mixtures of distributions

We have just scratched the surface of what can be done with `fmm`. We can fit mixtures of linear and generalized linear regression models where the effect of the covariates and the covariates themselves differ by class; see [FMM] **fmm estimation** for a list of supported outcome models. We can also model class probabilities with common or class-specific covariates.

More complicated FMMs can be fit using `gsem` within the LCA framework. `gsem` allows more than one response variable per component and more than one categorical latent variable; see, for instance, [SEM] **example 54g**, where we fit a mixture of Poisson regression models to multiple responses. See *Latent class analysis (LCA)* in [SEM] **intro 2** and *Latent class models* in [SEM] **intro 5** for an overview of latent class modeling with `gsem`.

Acknowledgment

We gratefully acknowledge the previous work by Partha Deb at Hunter College and the Graduate Center, City University of New York; see Deb (2007).

References

Deb, P. 2007. fmm: Stata module to estimate finite mixture models. Boston College Department of Economics, Statistical Software Components s456895. https://ideas.repec.org/c/boc/bocode/s456895.html.

Deb, P., and P. K. Trivedi. 1997. Demand for medical care by the elderly: A finite mixture approach. *Journal of Applied Econometrics* 12: 313–336.

Frühwirth-Schnatter, S. 2006. *Finite Mixture and Markov Switching Models*. New York: Springer.

Izenman, A. J., and C. J. Sommer. 1988. Philatelic mixtures and multimodal densities. *Journal of the American Statistical Association* 83: 941–953.

Jones, A. M., N. Rice, T. Bago D'Uva, and S. Balia. 2013. *Applied Health Economics*. 2nd ed. New York: Routledge.

Jorgensen, M. 2004. Using multinomial mixture models to cluster Internet traffic. *Australian & New Zealand Journal of Statistics* 46: 205–218.

McLachlan, G. J., and D. Peel. 2000. *Finite Mixture Models*. New York: Wiley.

Schlattmann, P., E. Dietz, and D. Böhning. 1996. Covariate adjusted mixture models and disease mapping with the program dismapwin. *Statistics in Medicine* 15: 919–929.

Skrondal, A., and S. Rabe-Hesketh. 2004. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton, FL: Chapman & Hall/CRC.

Wedel, M., and W. S. DeSarbo. 1993. A latent class binomial logit methodology for the analysis of paired comparison choice data: An application reinvestigating the determinants of perceived risk. *Decision Sciences* 24: 1157–1170.

Also see

[FMM] **fmm** — Finite mixture models using the fmm prefix

[FMM] **example 1a** — Mixture of linear regression models

[FMM] **example 2** — Mixture of Poisson regression models

[FMM] **example 3** — Zero-inflated models

[FMM] **example 4** — Mixture cure models for survival data

[FMM] **Glossary**

[SEM] **gsem** — Generalized structural equation model estimation command