Mixture Models: Estimation and Economic Applications

Paul T. Scott

Empirical IO
Fall 2013

Mixture model notation

- x - observed variables
- ζ - unobserved variables assumed to have finite support, Z
- θ parameters of interest

- $p\left(x_i, \zeta_i | \theta\right)$ - complete data likelihood for ith observation
- $p\left(x_i | \theta\right)$ - incomplete data likelihood for ith observation:

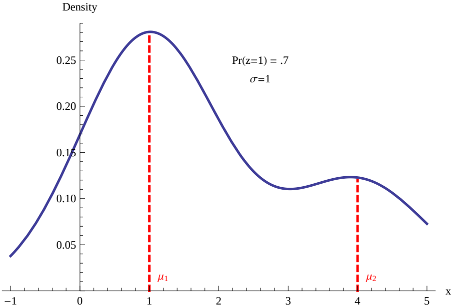
$$p\left(x_i | \theta\right) = \sum_{z \in Z} p\left(x_i, z | \theta\right)$$

- $q_{iz}\left(\theta\right)$ - expectation of incomplete data

$$q_{iz}\left(\theta\right) = Pr\left(\zeta_i = z | x_i, \theta\right)$$

Example 1: mixture of normals

- $\theta = (\mu_1, \mu_2, \sigma, \alpha_1)$
- If $z_i = 1$, then $x_i \sim N(\mu_1, \sigma)$
- If $z_i = 2$, then $x_i \sim N(\mu_2, \sigma)$
- $Pr(z_i = 1) = \alpha_1$



Example 2: discrete choice with heterogeneity

- Panel of bus maintenance decisions indexed by (i, t)
- $x_{it} = (d_{it}, p_t, s_{it})$
  - $d_{it} \in \{0, 1\}$ - agent i's action at time t
  - p_t - price of new bus engine
  - s_{it} - mileage on bus engine. $s_{it} \in \{0, 1, \ldots, 90\}$
- z_{it} - type of route bus takes. $z_{it} \in \{1, 2\}$

Example 3: collusion (Porter, 1983)

- Rob Porter (1983), "A Study of Cartel Stability: The Joint Executive Committee, 1880-1886"
- $$\begin{aligned} \ln Q_t &= \alpha_0 + \alpha_1 \ln P_t + \alpha_2 L_t + U_{1t} \\ \ln P_t &= \beta_0 + \beta_1 \ln Q_t + \beta_2 S_t + \beta_3 I_t + U_{2t} \end{aligned}$$

  where
  - L_t: demand shifters
  - S_t: supply shifters
  - $I_t \in \{0, 1\}$ indicating whether the cartel was in a price war or not
- In previous notation,
  - $x_t = (Q_t, P_t, L_t, S_t)$
  - $z_t = I_t$
  - $\theta = (\alpha, \beta)$
  - to deal with simultaneity, likelihood function $p(x_i, \zeta_i | \theta)$ is FIML

Complete and incomplete data likelihoods

The *incomplete data log-likelihood function* or *unconditional log-likelihood function* for a mixture model involves a sum within an expectation, which makes it very hard to maximize with standard optimization algorithms:

$$\mathcal{L}(x|\theta) = \sum_i \ln \left(\sum_z p(x_i, z|\theta) \right).$$

The EM algorithm is based on the (expected) *complete data log-likelihood function*:

$$Q(x, q|\theta) = \sum_i \sum_z q_{iz} \ln (p(x_i, z|\theta)).$$

Note that Q would simply be the log-likelihood function if ζ were observed.

EM Algorithm overview

- The EM algorithm starts with some initial guess for $\theta^{(0)}$
- In the E-step, we calculate expectations of the q's conditional on the parameter values:

$$q_{iz}^{(m)} = Pr\left(\zeta_i = z|\theta^{(m-1)}\right).$$

- In the M-step, we maximize the value of the complete data likelihood function:

$$\theta^{(m)} = \max_{\theta} Q\left(x, q^{(m)}|\theta\right).$$

- The EM Algorithm iteratively applies E and M steps until $\theta(m)$ converges.

EM Algorithm overview

- As I will illustrate, the E and M steps are often easy computationally (in contrast to maximization of incomplete data likelihood function).
- Each EM iteration increases $\mathcal{L}(x|\theta)$.
- Thus, iterating on the E and M steps will monotonically increase $\mathcal{L}\left(x|\theta^{(m)}\right)$, and $\theta^{(m)}$ will typically converge to a local maximum of $\mathcal{L}(x|\theta)$.
- $\Rightarrow$ EM Algorithm transforms a hard optimization problem into a series of easy optimization problems

Monotonicity

Monotonicity

$$\mathcal{L}\left(x|\theta^{(m)}\right) \geq \mathcal{L}\left(x|\theta^{(m-1)}\right)$$

Monotonicity

Monotonicity

$\mathcal{L}\left(x|\theta^{(m)}\right) \geq \mathcal{L}\left(x|\theta^{(m-1)}\right)$

$$
\begin{aligned}
\mathcal{L}\left(x|\theta^{(m)}\right) &= \sum_i \ln\left(\sum_z p\left(x_i|\zeta_i,\theta^{(m)}\right) p\left(\zeta_i|\theta^{(m)}\right)\right) \\
&= \sum_i \ln\left(\sum_z p\left(\zeta_i = z|x,\theta^{(m-1)}\right) \frac{p\left(x_i|\zeta_i,\theta^{(m)}\right)p\left(\zeta_i|\theta^{(m)}\right)}{p\left(\zeta_i=z|x,\theta^{(n-1)}\right)}\right) \\
&\geq \sum_i \sum_z p\left(\zeta_i = z|x,\theta^{(m-1)}\right) \ln\left(\frac{p\left(x_i|\zeta_i,\theta^{(m)}\right)p\left(\zeta_i|\theta^{(m)}\right)}{p\left(\zeta_i=z|x,\theta^{(m-1)}\right)}\right)
\end{aligned}
$$

where the inequality follows from Jensen's inequality

Monotonicity

$$
\begin{aligned}
\mathcal{L}\left(x|\theta^{(m)}\right) &= \sum_i \ln\left(\sum_z p\left(x_i|\zeta_i,\theta^{(m)}\right) p\left(\zeta_i\theta^{(m)}\right)\right) \\
&= \sum_i \ln\left(\sum_z p\left(\zeta_i = z|x,\theta^{(m-1)}\right) \frac{p\left(x_i|\zeta_i,\theta^{(m)}\right)p\left(\zeta_i|\theta^{(m)}\right)}{p\left(\zeta_i = z|x,\theta^{(n-1)}\right)}\right) \\
&\geq \sum_i \sum_z p\left(\zeta_i = z|x,\theta^{(m-1)}\right) \ln\left(\frac{p\left(x_i|\zeta_i,\theta^{(m)}\right)p\left(\zeta_i|\theta^{(m)}\right)}{p\left(\zeta_i = z|x,\theta^{(m-1)}\right)}\right) \\
&\geq \sum_i \sum_z p\left(\zeta_i = z|x,\theta^{(m-1)}\right) \ln\left(\frac{p\left(x_i|\zeta_i,\theta^{(m-1)}\right)p\left(\zeta_i\theta^{(m-1)}\right)}{p\left(\zeta_i = z|x,\theta^{(m-1)}\right)}\right)
\end{aligned}
$$

where the second inequality follows because $\theta^{(m)}$ is selected to maximize

$$\sum_i \sum_z p\left(\zeta_i = z|x,\theta^{(m-1)}\right) \ln\left(p\left(x_i|\zeta_i,\theta\right) p\left(\zeta_i|\theta\right)\right)$$

Monotonicity

$$
\begin{aligned}
\mathcal{L}\left(x|\theta^{(m)}\right) &= \sum_i \ln\left(\sum_z p\left(x_i|\zeta_i,\theta^{(m)}\right) p\left(\zeta_i|\theta^{(m)}\right)\right) \\
&= \sum_i \ln\left(\sum_z p\left(\zeta_i = z|x,\theta^{(m-1)}\right) \frac{p\left(x_i|\zeta_i,\theta^{(m)}\right)p\left(\zeta_i|\theta^{(m)}\right)}{p\left(\zeta_i = z|x,\theta^{(n-1)}\right)}\right) \\
&\geq \sum_i \sum_z p\left(\zeta_i = z|x,\theta^{(m-1)}\right) \ln\left(\frac{p\left(x_i|\zeta_i,\theta^{(m)}\right)p\left(\zeta_i|\theta^{(m)}\right)}{p\left(\zeta_i = z|x,\theta^{(m-1)}\right)}\right) \\
&\geq \sum_i \sum_z p\left(\zeta_i = z|x,\theta^{(m-1)}\right) \ln\left(\frac{p\left(x_i|\zeta_i,\theta^{(m-1)}\right)p\left(\zeta_i|\theta^{(m-1)}\right)}{p\left(\zeta_i = z|x,\theta^{(m-1)}\right)}\right) \\
&= \mathcal{L}\left(x|\theta^{(m-1)}\right)
\end{aligned}
$$

Estimation of example 1: mixture of normals

- $\theta = (\mu_1, \mu_2, \sigma, \alpha_1)$
- If $z_i = 1$, then $x_i \sim N(\mu_1, \sigma)$
- If $z_i = 2$, then $x_i \sim N(\mu_2, \sigma)$
- $Pr(z_i = 1) = \alpha_1$

In the E step, we just apply Bayes's Theorem to find q's

$$q_{i1}^{(m)} = Pr\left(z_i = 1 | x_i, \theta^{(m)}\right) = \frac{\alpha_1^{(m)} f\left(x_i | \mu_1^{(m)}, \sigma^{(m)}\right)}{\alpha_1^{(m)} f\left(x_i | \mu_1^{(m)}, \sigma^{(m)}\right) + \left(1 - \alpha_1^{(m)}\right) f\left(x_i | \mu_2^{(m)}, \sigma^{(m)}\right)}$$

where $f(x|\mu, \sigma)$ is the density at x of the normal distribution with mean μ and standard deviation σ^2.

Estimation of example 1: mixture of normals

- In the M step, maximizing the complete data likelihood function amounts to taking weighted means:

$$\mu_z^{(m)} = \sum_i q_{iz}^{(m)} x_i$$

$$\sigma^{(m)} = \sqrt{\frac{\sum_z \sum_i q_{iz}^{(m)} (x_i - \mu_z)^2}{\sum_z \sum_i q_{iz}^{(m)}}}$$

$$\alpha_z^{(m)} = N^{-1} \sum_i q_{iz}^{(m)}$$

Estimation of example 1: mixture of normals

- Note: in a mixture model with covariates that enter linearly, the M step involves weighted OLS instead of a weighted mean
- Bottom line: E and M step are both easy computationally, so iterating on them goes quickly.
- In general, the EM algorithm can stop at local maxima, so some care is needed to ensure a global optimum is attained (e.g., multiple starting points).

"Finite Mixture Distributions, Sequential Likelihood and the EM Algorithm"
Arcidiacono and Jones (2003)

Setup

- x_i: ith observation
- z: mixture component
- $f_z(x_i; \theta_1, \theta_2) = f_{1z}(x_i; \theta_1) f_{2z}(x_i; \theta_1, \theta_2)$ distribution function of x for component z
- α_z: unconditional probability of component z
- n.b., different notation from the paper

Sequential estimation background

- Forget about the mixture model for this slide:

$$f\left(x_i;\theta_1,\theta_2\right)=f_1\left(x_i;\theta_1\right)f_2\left(x_i,\theta_1,\theta_2\right)$$

- We could estimate θ_1 and θ_2 by choosing them to jointly maximize $\sum_i \ln f$, or we could estimate:

$$\begin{aligned}\tilde{\theta}_1 &= \max_{\theta_1}\sum \ln f_1\left(x_i;\theta_1\right)\\ \tilde{\theta}_2 &= \max_{\theta_2}\sum \ln f_2\left(x_i;\tilde{\theta}_1,\theta_2\right)\end{aligned}$$

- e.g., Hotz and Miller: conditional choice probabilities are estimated before profit function is estimated.

Sequential M step

- Main idea: combine sequential estimation and EM algorithm.
- Normal EM algorithm would estimate $\left(\theta_1^{(m)}, \theta_2^{(m)}\right)$ to jointly maximize

$$\left(\theta_1^{(m)}, \theta_2^{(m)}\right) = \arg\max_{(\theta_1,\theta_2)} \sum_i \sum_z q_{iz}^{(m)} \ln\left(f_{1z}\left(x_i;\theta_1\right) f_{2z}\left(x_i;\theta_1,\theta_2\right)\right)$$

- Arcidiacono and Jones's ESM algorithm estimates $\left(\theta_1^{(m)}, \theta_2^{(m)}\right)$ to satisfy:

$$\begin{aligned} \theta_1^{(m)} &= \arg\max_{\theta_1} \textstyle\sum_i \sum_z q_{iz}^{(m)} \ln\left(f_{1z}\left(x_i;\theta_1\right)\right) \\ \theta_2^{(m)} &= \arg\max_{\theta_2} \textstyle\sum_i \sum_z q_{iz}^{(m)} \ln\left(f_{2z}\left(x_i;\theta_1^{(m)},\theta_2\right)\right) \end{aligned}$$

- Does this strategy work? What are its asymptotic properties?

Moments, part 1

- The true parameters (θ^*, α^*) satisfy:

$$(\theta^*, \alpha^*) = \arg\max_{(\theta,\alpha)} E_{x,z}\left[\ln\left(\alpha_z f_{1z}\left(x_i; \theta_1\right) f_{2z}\left(x_i; \theta_1, \theta_2\right)\right)\right].$$

- By the law of total probability,

$$(\theta^*, \alpha^*) = \arg\max_{(\theta,\alpha)} E_x\left[\sum_z Pr\left(z|x; \theta^*, \alpha^*\right) \ln\left(\alpha_z f_{1z}\left(x_i; \theta_1\right) f_{2z}\left(x_i; \theta_1, \theta_2\right)\right)\right].$$

- The first-order conditions for θ_2 is:

$$\sum_z Pr\left(z|x; \theta^*, \alpha^*\right) \frac{\partial \ln\left(f_{2z}\left(x_i; \theta_1^*, \theta_2\right)\right)}{\partial \theta_2} = 0.$$

- And θ_1 can be estimated just from the f_1 likelihood functions:

$$\sum_z Pr\left(z|x; \theta^*, \alpha^*\right) \frac{\partial \ln\left(f_{1z}\left(x_i; \theta_1\right)\right)}{\partial \theta_1} = 0.$$

Moments, part 2

- A set of moments satisfied by the true parameters:

$$E\begin{pmatrix} \sum_z Pr\left(z|x;\theta^*,\alpha^*\right)\frac{\partial \ln\left(f_{2z}\left(x_i;\theta_1^*,\theta_2\right)\right)}{\partial\theta_2} \\ \sum_z Pr\left(z|x;\theta^*,\alpha^*\right)\frac{\partial \ln(f_{1z}(x_i;\theta_1))}{\partial\theta_1} \\ Pr\left(1|x;\theta^*,\alpha^*\right)-\alpha_1 \\ \vdots \\ Pr\left(Z|x;\theta^*,\alpha^*\right)-\alpha_Z \end{pmatrix}=0$$

- If the ESM algorithm converges, it converges to parameters satisfying the empirical analog of these moments.
- $\Rightarrow$ so now we're talking about a GMM estimator, and the ESM algorithm might be a useful tool to find the point estimate

Arcidiacono and Jones (2003)

TABLE I

Simulation Results

	Estimation Method			
	Complete	Incomplete	FIML	ESM
Mean $\hat{\ }_C$	0 2078	0 2932	0 2255	0 2226
Standard Deviation $\hat{\ }_C$	0 0330	0 0323	0 0496	0 0565
Mean Squared Error × 100	0 1141	0 9731	0 3082	0 3667
(FIML FLOPs)/(ESM FLOPs)				22.48

Note: Each simulation was conducted 100 times with 3000 observations. The distributions of unknown state variables were approximated with 10-point discrete distributions. Mean squared error refers to the squared differences between estimates of $_C$ and its true value of 0.2.

Further comments

- Some econometric models feature difficult likelihood functions but easy sequential estimation approaches. ESM offers a way to extend these estimation approaches to mixture models.
- ESM algorithm yields a GMM estimator which is less efficient asymptotically than the maximum likelihood estimator
- ESM algorithm doesn't have monotonicity property of EM algorithm (don't confuse ESM with ECM or GEM, which retain monotonicity)
- However, Arcidiacono and Jones find ESM still converges, and I have experienced the same

Identification of example 2: discrete choice with heterogeneity

- Perhaps it's intuitive how a mixture of normals is identified, but it's harder to see how a discrete choice model with unobservable heterogeneity is identified

- Note: there is clearly no identification in a cross section. When the aggregate probability of action j is .5, we could have homogeneous agents who all have choice probabilities of .5, or the population could be split between agents who always choose action j and agents who never do.

- Thus, identification of discrete choice models with unobservable heterogeneity comes from the panel data structure.

Identification of example 2: discrete choice with heterogeneity

- For a thorough treatment of identification of DDC models with unobservable heterogeneity, see Kasahara and Shimotsu (2009), "Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Processes"
- For more basic intuition, see Hall and Zhou (2003), "Nonparametric Estimation of Component Distributions in a Multivariate Mixture."

"Conditional Choice Probability Estimation
of Dynamic Discrete Choice Models
with Unobserved Heterogeneity"
Arcidiacono and Miller (2011)

Overview

- They show how CCP-based estimation techniques for DDC models can be adapted to deal with unobservable heterogeneity or unobserved state variables with discrete distributions
- Their main approach is based on ESM algorithm, but they propose an alternative two-stage approach in which the EM algorithm is only used to estimate CCP's in a first stage.
- They formalize the notion of *finite dependence*, which allows for computationally simple applications of the Hotz-Miller inversion

Notation

- θ - parameters to be estimated
  - θ_1 - parameters affecting state transitions
  - θ_2 - parameters of profit function
- z - mixing components
- α_z - probability of component z
- d_{jit} - dummy for decision j by agent i in period t
- $p(x, z)$ - choice probabilities conditional on observed state x and unobserved state z
- l - log likelihood function

(Notation here slightly different than the paper.)

E step

- Let's take the likelihood function $l\left(d_{it}|x_{it}, z', \hat{p}^{(m-1)}, \theta^{(m)}\right)$ for granted for now.
- E step is pretty standard:

$$q_{iz}^{(m)} = \frac{\alpha_z^{(m-1)} \prod_{t=1}^{T} l\left(d_{it}|x_{it}, z', \hat{p}^{(m-1)}, \theta^{(m)}\right)}{\sum_{z'} \alpha_{z'}^{(m-1)} \prod_{t=1}^{T} l\left(d_{it}|x_{it}, z', \hat{p}^{(m-1)}, \theta^{(m)}\right)}$$

M step, first approach

$$\alpha_z^{(m)} = \frac{1}{N} \sum_{i=1}^{N} q_{iz}^{(m)}$$

$$\theta^{(m)} = \arg\max_{\theta} \sum_i \sum_z \sum_t q_{iz}^{(m)} l \left(d_{it} | x_{it}, z', \hat{p}^{(m-1)}, \theta^{(m)} \right)$$

There are two options for updating p:

$$p_j^{(m)}(x, z) = \frac{\sum_i \sum_t d_{jit} q_{iz}^{(m)} I(x_{it} = x)}{\sum_i \sum_t q_{iz}^{(m)} I(x_{it} = x)}$$

$$p_j^{(m)}(x, z) = l \left(d_{it} | x_{it}, z', \hat{p}^{(m-1)}, \theta^{(m)} \right)$$

M step, second approach

- In the alternative approach, in the first stage EM estimation, we only worry about estimating p, α, and θ_1
- The utility function, θ_2, is then estimated in a second stage, after the EM algorithm has completed.

Applying Hotz-Miller

- The likelihood function is based on the Hotz-Miller inversion, as we have seen before.
- For example, let's suppose logit errors and that action 0 is a terminal action, always leading conditional payoffs of zero.

$$
\begin{aligned}
l\left(d_{it}=j|x_{it},z',p,\theta\right) &= \frac{\exp\left(u_j(x,z;\theta)+\beta E\left[\bar{V}\left(x',z;\theta\right)+\gamma|j,x\right]\right)}{\sum_{j'}\exp\left(u_{j'}(x,z;\theta)+\beta E\left[\bar{V}(x',z;\theta)+\gamma|j',x\right]\right)} \\
&= \frac{\exp\left(u_j(x,z;\theta)+\beta E\left[\ln\sum_{j''}p_{j''}\left(x',z\right)/p_0\left(x',z\right)+\gamma|j,x\right]\right)}{\sum_{j'}\exp\left(u_{j'}(x,z;\theta)+\beta E\left[\ln\sum_{j''}p_{j''}(x',z)/p_0(x',z)+\gamma|j',x\right]\right)}
\end{aligned}
$$

where I have used

$$
\begin{aligned}
\bar{V}\left(x',z;\theta\right) &= \ln\left(\sum_{j''}p_{j''}\left(x',z\right)/p_0\left(x',z\right)\exp\left(v_0\left(x',z,p,\theta\right)\right)\right)+\gamma \\
&= \ln\sum_{j''}p\left(x',z\right)/p_0\left(x',z\right)+\gamma
\end{aligned}
$$

Finite Dependence

- We can always derive a relatively simple expression for the likelihood function in terms of choice probabilities and the utility function, as long as we have *finite dependence*.
- Finite dependence requires that there is always a sequence of actions that, starting from two different initial actions, will lead to the same state(s) in expectation within a finite number of periods.
- Renewal actions and terminal actions are particularly convenient forms of finite dependence.