Reviewing from the previous video:

When s is unobserved, the log likelihood function is not additively separable:

$$\ell = \sum_i \log \left(\sum_s \pi_s \prod_t \mathcal{L}(Y_{1t}|X_{1t}, \alpha_1, s)\mathcal{L}(Y_{2t}|X_{2t}, \alpha_2, s) \right)$$

where $\mathcal{L}$ is a likelihood function

Econometrica, Vol. 71, No. 3 (May, 2003), 933–946

FINITE MIXTURE DISTRIBUTIONS, SEQUENTIAL LIKELIHOOD AND THE EM ALGORITHM

BY PETER ARCIDIACONO AND JOHN BAILEY JONES[1]

A popular way to account for unobserved heterogeneity is to assume that the data are drawn from a finite mixture distribution. A barrier to using finite mixture models is that parameters that could previously be estimated in stages must now be estimated jointly: using mixture distributions destroys any additive separability of the log-likelihood function. We show, however, that an extension of the EM algorithm reintroduces additive separability, thus allowing one to estimate parameters sequentially during each maximization step. In establishing this result, we develop a broad class of estimators for mixture models. Returning to the likelihood problem, we show that, relative to full information maximum likelihood, our sequential estimator can generate large computational savings with little loss of efficiency.

KEYWORDS: Unobserved heterogeneity, mixture distributions, EM algorithm, dynamic discrete choice.

1. INTRODUCTION

ONE WAY TO ACCOUNT for unobserved heterogeneity in data, and the related problem of self-selection, is to assume that the data are drawn from a finite mixture distribution. Under this approach, each observation is assumed to belong to one of several different "types," each of which has its own distribution. While the econometrician does not observe each observation's type, if her model is sufficiently structured she can infer it by applying Bayes' Theorem.

Models with finite mixtures have appeared in numerous applications.[2] In labor economics, Keane and Wolpin (1997) and Eckstein and Wolpin (1999) use mixtures to control for person-specific differences in models of dynamic discrete choice. Finite mixture models form the basis of Hamilton's (1989, 1990) influential regime-switching model of economic time series. A particularly important application has been to use finite mixture models as nonparametric approximations to more general mixture models. Important papers in this vein include Laird (1978), Lindsay (1983), and Heckman and Singer (1984). More recently, Cameron and Heckman (1998, 2001) use this sort of nonparametric maximum likelihood estimation to study the effect of family background on educational achievement. Mroz (1999) uses mixtures to control for endogeneity in a binary explanatory variable. He shows that "discrete factor approximations" to a continuous latent variable often

[1] We thank Donald Andrews, Arie Beresteanu, Mark Coppejans, Michael McCracken, Tom Mroz, Barbara Rossi, Wilbert van der Klaauw, and two anonymous referees for valuable comments.

[2] Although we focus on economic applications, finite mixture models have been used widely in other fields as well. Titterington, Smith, and Makov (1985) and McLachlan and Peel (2000) provide lists.

We can get additive separability of the finite mixture model with the EM algorithm

EM stands for “Expectation-Maximization”

We can get additive separability of the finite mixture model with the EM algorithm

EM stands for “Expectation-Maximization”

The algorithm iterates on two steps:

E-step: estimate parameters of the mixing distribution (the π’s)

M-step: pretend you observe the unobserved variable and estimate the α’s

We can get additive separability of the finite mixture model with the EM algorithm

EM stands for "Expectation-Maximization"

The algorithm iterates on two steps:

E-step: estimate parameters of the mixing distribution (the π's)

M-step: pretend you observe the unobserved variable and estimate the α's

The EM algorithm is used in other applications to fill in missing data

In this case, the missing data is the permanent unobserved heterogeneity

With the EM algorithm, the non-separable likelihood function

$$\ell = \sum_i \log \left(\sum_s \pi_s \prod_t \mathcal{L}(Y_{1t}|X_{1t}, \alpha_1, s)\mathcal{L}(Y_{2t}|X_{2t}, \alpha_2, s) \right)$$

With the EM algorithm, the non-separable likelihood function

$$\ell = \sum_i \log\left(\sum_s \pi_s \prod_t \mathcal{L}(Y_{1t}|X_{1t}, \alpha_1, s)\mathcal{L}(Y_{2t}|X_{2t}, \alpha_2, s)\right)$$

can be written in a form that is separable:

$$\ell = \sum_i \sum_s q_{is} \sum_t \{\ell_1(Y_{1t}|X_{1t}, \alpha_1, s) + \ell_2(Y_{2t}|X_{2t}, \alpha_2, s)\}$$

where q_{is} is the (posterior) probability that i belongs to group s

With the EM algorithm, the non-separable likelihood function

$$\ell = \sum_i \log\left(\sum_s \pi_s \prod_t \mathcal{L}(Y_{1t}|X_{1t}, \alpha_1, s)\mathcal{L}(Y_{2t}|X_{2t}, \alpha_2, s)\right)$$

can be written in a form that is separable:

$$\ell = \sum_i \sum_s q_{is} \sum_t \{\ell_1(Y_{1t}|X_{1t}, \alpha_1, s) + \ell_2(Y_{2t}|X_{2t}, \alpha_2, s)\}$$

where q_{is} is the (posterior) probability that i belongs to group s

q_{is} satisfies $\pi_s = \frac{1}{N}\sum_i q_{is}$

We can now estimate the model in stages because of the restoration of separability

The only twist is that we need to weight by the q's in each estimation stage

We can now estimate the model in stages because of the restoration of separability

The only twist is that we need to weight by the q's in each estimation stage

Stage 1 of M-step: estimate $\ell(Y_{1t}|X_{1t}, \alpha_1, s)$ weighting by the q's

Stage 2 of M-step: estimate $\ell(Y_{2t}|X_{2t}, \alpha_2, s)$ weighting by the q's

We can now estimate the model in stages because of the restoration of separability

The only twist is that we need to weight by the q's in each estimation stage

Stage 1 of M-step: estimate $\ell(Y_{1t}|X_{1t}, \alpha_1, s)$ weighting by the q's

Stage 2 of M-step: estimate $\ell(Y_{2t}|X_{2t}, \alpha_2, s)$ weighting by the q's

E-step: update the q's by calculating

$$q_{is} = \frac{\pi_s \prod_t \mathcal{L}(Y_{1t}|X_{1t}, \alpha_1, s)\mathcal{L}(Y_{2t}|X_{2t}, \alpha_2, s)}{\sum_m \pi_m \prod_t \mathcal{L}(Y_{1t}|X_{1t}, \alpha_1, m)\mathcal{L}(Y_{2t}|X_{2t}, \alpha_2, m)}$$

We can now estimate the model in stages because of the restoration of separability

The only twist is that we need to weight by the q's in each estimation stage

Stage 1 of M-step: estimate $\ell(Y_{1t}|X_{1t}, \alpha_1, s)$ weighting by the q's

Stage 2 of M-step: estimate $\ell(Y_{2t}|X_{2t}, \alpha_2, s)$ weighting by the q's

E-step: update the q's by calculating

$$q_{is} = \frac{\pi_s \prod_t \mathcal{L}(Y_{1t}|X_{1t}, \alpha_1, s)\mathcal{L}(Y_{2t}|X_{2t}, \alpha_2, s)}{\sum_m \pi_m \prod_t \mathcal{L}(Y_{1t}|X_{1t}, \alpha_1, m)\mathcal{L}(Y_{2t}|X_{2t}, \alpha_2, m)}$$

Iterate on E and M steps until the q's converge

With permanent unobserved heterogeneity, we no longer have global concavity

- This means that different starting values will give different estimates

With permanent unobserved heterogeneity, we no longer have global concavity

- This means that different starting values will give different estimates

Another issue is standard errors

With permanent unobserved heterogeneity, we no longer have global concavity

- This means that different starting values will give different estimates

Another issue is standard errors

- With stages, each stage introduces estimation error into the following stages

With permanent unobserved heterogeneity, we no longer have global concavity

- This means that different starting values will give different estimates

Another issue is standard errors

- With stages, each stage introduces estimation error into the following stages

- We take the estimate as given, but it's actually subject to sampling error

With permanent unobserved heterogeneity, we no longer have global concavity

- This means that different starting values will give different estimates

Another issue is standard errors

- With stages, each stage introduces estimation error into the following stages

- We take the estimate as given, but it's actually subject to sampling error

Both issues (local optima and estimation error) are problem-specific

You need to understand your specific case