Multinomial Discrete Choice: Mixed Logit

Chris Conlon

September 1, 2020

Grad IO

Mixed/ Random Coefficients Logit

An alternative is to allow for individuals to have random coefficients β_i:

$$U_{ij} = \beta_i x_j + \varepsilon_{ij}, \quad \beta_i \sim f(\beta_i|\theta), \quad \varepsilon_{ij} \sim \text{Type I EV}$$

As an alternative, we could have specified an error components structure on ε_i.

$$U_{ij} = \beta x_{ij} + \underbrace{\nu_i z_{ij} + \varepsilon_{ij}}_{\tilde{\varepsilon}_{ij}}$$

- x_{ij} are observed per usual and ε_{ij} is IID Type I EV.
- The key is that ν_i is unobserved and mean zero (often normally distributed).
- This allows for a heteroskedastic structure on ε_i, but only one which we can project down onto the space of x.

Mixed Logit

We relax the IIA property by mixing over various logits:

$$
\begin{aligned}
u_{ijt} &= x_j \beta_i + \varepsilon_{ij} & &= x_j \beta + \mu_{ij} + \varepsilon_{ij} \\
s_{ij}(\theta) &= \int \frac{\exp[x_j \beta_i]}{1 + \sum_k \exp[x_k \beta_i]} f(\boldsymbol{\beta_i}|\theta) & &= \int \frac{\exp[x_j \beta + \mu_{ij}]}{1 + \sum_k \exp[x_k \beta + \mu_{ik}]} f(\boldsymbol{\mu_i}|\theta)
\end{aligned}
$$

- Each individual draws a vector β_i / μ_i (separately from ε_i).
- Conditional on β_i / μ_i each person follows an IIA logit model.
- However we integrate (or mix) over many such individuals giving us a mixed logit or heirarchical model (if you are a statistician)
- In practice these are not that different from linear random effects models you have learned about in econometrics.
- It helps to think about fixing β_i / μ_i first and then integrating out over ε_i

Kinds of heterogeneity

$$s_{ij}(\theta) = \int \frac{\exp[x_j \beta_i]}{1 + \sum_k \exp[x_k \beta_i]} f(\boldsymbol{\beta_i} | \theta) \approx \sum_{s=1}^{S} w_i^s \frac{\exp[x_j \beta_i^s]}{1 + \sum_k \exp[x_k \beta_i^s]}$$

- We can only approximate the integral with weights w_i^s and nodes β_i^s
    - We can allow for there to be two types of β_i in the population (high-type, low-type). latent class model.
    - We can allow β_i to follow an independent normal distribution for each component of x_{ij} such as $\beta_i = \overline{\beta} + \nu_i \sigma$.
    - We can allow for correlated normal draws $\beta_i \sim N(\mu_\beta, \Sigma_\beta)$.
    - Can allow for non-normal distributions too (lognormal, exponential).
    - Why is normal so easy?

Adding Heterogeneity

- The structure is extremely flexible but at a cost.
- We generally must perform the integration numerically.
- High-dimensional numerical integration is difficult. In fact, integration in dimension 8 or higher makes me very nervous.
- We need to be parsimonious in how many variables have unobservable heterogeneity.
- Again observed heterogeneity does not make life difficult so the more of that the better!
    - If we see individual income, education, distance to hospital, etc. we can always interact that with observed characteristics without doing any additional integration.

Mixed Logit

How does it work?

- Well we are mixing over individuals who conditional on β_i or μ_i follow logit substitution patterns, however they may differ wildly in their s_{ij} and hence their substitution patterns.
- For example if we are buying cameras: I may care a lot about price, you may care a lot about megapixels, and someone else may care mostly about zoom.
- The basic idea is that we need to explain the heteroskedasticity of $Cov(\varepsilon_i, \varepsilon_j)$ what random coefficients do is let us use a basis from our X's.
- If our X's are able to span the space effectively, then an RC logit model can approximate any arbitrary RUM (such as probit) (McFadden and Train 2002).
- Of course if you have 1000 products and two random coefficients, you are asking for a lot.

Elasticities

At the level of an individual substitution patterns follow a plain logit:

$$\epsilon_{s_j,p_j} = \frac{\partial s_j}{\partial p_j} \cdot \frac{p_j}{s_j} = \frac{p_j}{s_j} \cdot \int \frac{\partial s_{ij}}{\partial p_j} \, dF_i = \frac{p_j}{s_j} \int \beta_i \cdot s_{ij} \cdot (1 - s_{ij}) \, dF_i$$

$$\epsilon_{s_j,p_k} = \frac{\partial s_j}{\partial p_k} \cdot \frac{p_k}{s_j} = \frac{p_k}{s_j} \cdot \int \frac{\partial s_{ij}}{\partial p_k} \, dF_i = -\frac{p_k}{s_j} \int \beta_i \cdot s_{ij} \cdot s_{ik} \, dF_i$$

Notation: here $s_j \equiv s_j(x_i)$ denotes an individual choice probability (we re-use the i subscript) and $s_{ij}(\boldsymbol{\mu_i}) = s_{ij}$.

Substitution (Conlon Mortimer 2020)

At the level of an individual substitution patterns follow a plain logit:

$$D_{jk,i}(x) = \frac{\frac{\partial s_{ik}}{\partial p_j}}{\left|\frac{\partial s_{ij}}{\partial p_j}\right|} = \frac{s_{ik}(x)}{1 - s_{ij}(x)}$$

But at the aggregate level, we mix over heterogeneous individuals:

$$D_{jk}(x) = \int D_{jk,i}(x) w_i(z_j, z_j') dF_i$$

Different interventions (price changes, product removals, quality changes, etc.) give different weighting schemes.

Estimation Details: MSL/MSLE

How do we estimate these models? (Derived from multinomial log-likelihood):

$$\theta_{MLE} = \arg\min_{\theta} - \sum_{i=1}^{N} y_{ij} \ln \widehat{s}_{ij}(\theta)$$

- We need to perform numerical integration to get $\widehat{s}_{ij}(\theta)$ and its derivative: $\widehat{\frac{\partial s_{ij}}{\partial \theta}}$.
- Consistency
    - Technically for fixed number of MC draws the MSLE estimator is inconsistent. Why?
    - Very accurate integration is even more important!

Estimation Alternatives: MSM

Can also do a simulated GMM estimator with moment conditions:

$$E\left[\left(y_{ij} - \widehat{s}_{ij}(\theta)\right)|z_{ij}\right] = 0$$

- We need to perform numerical integration to get $\widehat{s}_{ij}(\theta)$ and its derivative: $\widehat{\frac{\partial s_{ij}}{\partial \theta}}$.
- The optimal instruments in the sense of Chamberlain (1987) or Amemiya (1977) are the derivative of moments with respect to parameters (scores): $z_{ij} = \frac{\partial \log s_{ij}}{\partial \theta}(\theta)$.
- The true scores are infeasible (they depend on θ_0).
    - At true scores: same FOC as MLE.
    - Simulated scores have same consistency issue as MSLE.

Estimation Alternatives: MSS

As an alternative we could work with the score function directly:

$$\sum_{i=1}^{n} \frac{\widehat{\partial \log s_{ij}}}{\partial \theta}(\theta) = \sum_{i=1}^{N} \frac{1}{\widehat{s_{ij}}(\theta)} \cdot \frac{\widehat{\partial s_{ij}}}{\partial \theta}(\theta) = 0$$

- We can get unbiased estimated of derivative (linearity)
- We can't get unbiased estimate of $\frac{1}{\widehat{s_{ij}}(\theta)}$
    - Train's book suggests Accept-Reject
    - Maybe importance sampling would work?

Estimation Details: Hints

How bad is the simulation error?

- Depends how small your shares are.
- Since you care about $\log s_{jt}$ when shares are small, tiny errors can be enormous.
- Often it is pretty bad.

Estimation Details: Recommendations

You should read these separate notes, but my recommendations are:

1. Numerical integration:
   - Use Gauss-Hermite quadrature rules to integrate over normal densities.
   - Follow Heiss and Winschel (JoE) and do sparse grids in medium dimensions
   - Try `http://sparse-grids.de`.
2. Nonlinear optimization: with mixtures the problem is non-convex (ie: very hard)
   - You must provide analytic gradients.
   - You should use a quasi-Newton solver and may need to try several.
   - Try multiple starting values.
3. Generalized Method of Moments
   - You can read more about optimal IV in these settings.