Mixed Logit

Stephane Hess
stephane.hess@gmail.com

Outline

1. Deterministic *vs* random heterogeneity
2. Heterogeneity across people and across choices
3. Mixed Logit: overview
4. Mixed Logit: model specification
5. Mixed Logit: estimation
6. Illustrative example
7. MRS from Mixed Logit
8. Error components

Deterministic *vs* random heterogeneity

Deterministic *vs* random heterogeneity

Deterministic heterogeneity: link to observed information

Option 1: discrete segmentations, with separate models

- e.g. male *vs* female, business *vs* leisure
- same as a joint model with segment-specific parameters
- assumes that differences exist in sensitivities to all attributes

Option 2: differences only for some attribute-covariate pairs

- interaction with categorical variables
  - e.g. interaction with gender, implying different sensitivities for men and women
- interactions with continuous covariates
  - e.g. continuous interaction with income, implying different sensitivity for each possible income

Deterministic *vs* random heterogeneity

Need for random taste heterogeneity

- Some taste heterogeneity cannot be explained deterministically
- Data limitations
  - we do not know everything about individuals in our data
  - constraints on behaviour, unobserved socio-demographics, etc
- Idiosyncratic reasons
  - two apparently *identical* individuals may have different sensitivities
- Solution is to allow for *random* heterogeneity
- Preferences are not random, but simply unobserved - we use random heterogeneity to deal with this

Deterministic *vs* random heterogeneity

Mixture models

- Aim:
  - accommodate random taste heterogeneity
- Method:
  - allow choice probabilities to vary as function of (unobserved) distribution of sensitivities

Deterministic *vs* random heterogeneity

Key differences across specifications

- Model specifications without any heterogeneity
  - same probability for all individuals when faced with same choice scenario
- Model specifications with deterministic heterogeneity
  - probabilities vary across individuals
  - we **know** where on that distribution each person is located
- Model specifications with random heterogeneity
  - probabilities also vary across individuals
  - we **do not know** where on that distribution each person is located
- In models combining deterministic with random heterogeneity, we can be more certain about where on the distribution a person is

Deterministic *vs* random heterogeneity

Two broad categories of models

Finite mixtures

- Allow for a limited number of possible values for sensitivities
- Two different implementations:
  - Discrete mixtures: heterogeneity in individual parameters
  - Latent class: finite set of combinations of values for different parameters

Continuous mixtures

- Use continuous statistical distributions to capture heterogeneity
- Most flexible type of random utility model in theory
- Known as Mixed Logit, or Random Parameters Logit

Deterministic *vs* random heterogeneity

Basic idea

- Same underlying idea for finite and continuous mixtures
- Analyst specifies an underlying model, typically (but not necessarily) MNL
  - in technical terms often referred to as the *kernel* of the mixture model
- If sensitivities of an individual were known, would have a probability for the choices as in MNL (or other kernel)
- But sensitivities are not known

Deterministic *vs* random heterogeneity

A VERY simple example

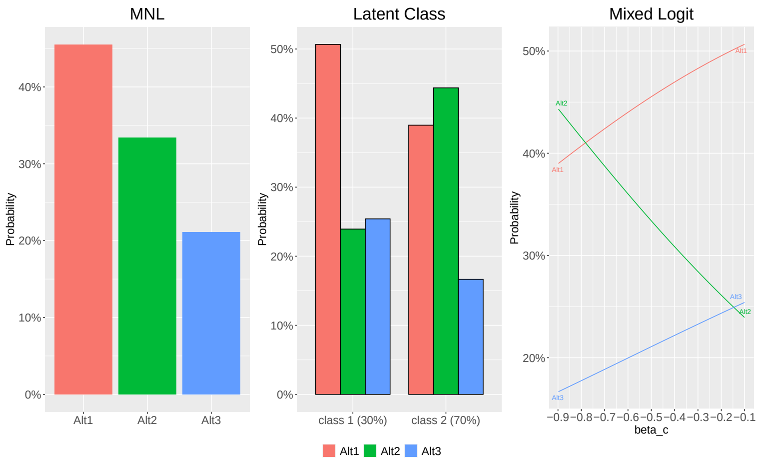
- Random cost (TC) coefficient random, fixed constants and sensitivities to free flow time (FFT), slowed down time (SDT) and tolls (TOLL)

	δ_1	δ_2	FFT	SDT	TC	TOLL
Alt 1	1	0	30	7	2.3	5
Alt 2	0	1	36	8	1.2	6
Alt 3	0	0	15	6	2.5	7
β	0.2	0.1	-0.03	-0.08	β_{TC}	-0.5

- Three cases:
  - MNL: $\beta_{TC} = -0.5$
  - Latent class: two classes, with weights of 30% and 70%, and values of β_{TC} of -0.1 and -0.9
  - Mixed Logit: β_{TC} is distributed uniformly on $[-0.9, -0.1]$

Deterministic *vs* random heterogeneity

Choice probabilities vary as a function of β_{TC}



Deterministic *vs* random heterogeneity

General econometrics: assuming MNL kernel

$$U_{jnt} = \sum_{k=1}^{K} \beta_k \, x_{jnt,k} + \varepsilon_{jnt} + \Delta_{jnt}, \text{ with } \varepsilon \sim EVI$$

$$\Delta_{jnt} = \sum_{k=1}^{K} \left(\beta_{k,n} - \beta_k\right) x_{jnt,k}$$

$$P_{int}\left(\Delta_n\right) = \frac{e^{V_{int}+\Delta_{int}}}{\sum_{j=1}^{J} e^{V_{jnt}+\Delta_{jnt}}}$$

Issue: Δ not known

Deterministic *vs* random heterogeneity

Simple illustration

- $\beta_{C_1}(-0.1)$ for 30% of individuals and $\beta_{C_2}(-0.9)$ for other 70% of individuals
- Conditional choice probabilities:

$$P_{int}\left(\beta_{C_1}\right) = \frac{e^{\delta_i + \beta_{C_1} C_{int} + \sum_{k \neq C} \beta_k x_{int,k}}}{\sum_{j=1}^{J} e^{\delta_j + \beta_{C_1} C_{jnt} + \sum_{k \neq C} \beta_k x_{jnt,k}}} \quad P_{int}\left(\beta_{C_2}\right) = \frac{e^{\delta_i + \beta_{C_2} C_{int} + \sum_{k \neq C} \beta_k x_{int,k}}}{\sum_{j=1}^{J} e^{\delta_j + \beta_{C_2} C_{jnt} + \sum_{k \neq C} \beta_k x_{jnt,k}}}$$

- Unconditional choice probability

$$P_{int}\left(\beta_{C_1}, \beta_{C_2}\right) = 0.3\, P_{int}\left(\beta_{C_1}\right) + 0.7\, P_{int}\left(\beta_{C_2}\right)$$

- Not the same as $P_{int}\left(0.3\beta_{C_1} + 0.7\beta_{C_2}\right)$

Deterministic *vs* random heterogeneity

Calculation for our example

- β_C takes on values of -0.1 and -0.9 with 30% and 70% probability, respectively

	δ_1	δ_2	FFT	SDT	TC	TOLL
Alt 1	1	0	30	7	2.3	5
Alt 2	0	1	36	8	1.2	6
Alt 3	0	0	15	6	2.5	7
β	0.2	0.1	-0.03	-0.08	β_C	-0.5

	$\beta_C = -0.1$		$\beta_C = -0.9$		
	V_j	P_j	V_j	P_j	Unconditional P_j
Alt 1	-3.99	50.66%	-5.83	38.97%	42.48%
Alt 2	-4.74	23.93%	-5.7	44.38%	38.25%
Alt 3	-4.68	25.41%	-6.68	16.66%	19.29%

Deterministic *vs* random heterogeneity

Misattributing random heterogeneity

- As with deterministic heterogeneity, model will use whatever flexibility we give it
- Allowing for heterogeneity in only some attributes risks misattribution

Number of individuals	500		500		500		500	
Number of modelled outcomes	5000		5000		5000		5000	
Estimated parameters	3		3		4		5	
LL(final)	-2160.457		-2150.033		-2150.017		-2149.356	
Adj.Rho-square (0)	0.3758		0.3788		0.3785		0.3784	
AIC	4326.91		4306.07		4308.03		4308.71	
BIC	4346.47		4325.62		4334.1		4341.3	
	estimate	Rob.t-ratio(0)	estimate	Rob.t-ratio(0)	estimate	Rob.t-ratio(0)	estimate	Rob.t-ratio(0)
efficacy (mean for lognormal)	-3.9673	-40.54	-3.8255	-48.46	-3.8257	-48.52	-3.7959	-46.23
efficacy (sd for lognormal)	-0.3579	-6.19	0	NA	-0.0316	-0.83	-0.0928	-0.97
risk (mean for lognormal)	-0.9883	-23.84	-0.9401	-22.92	-0.94	-22.92	-0.9301	-21.87
risk (sd for lognormal)	0	NA	0.262	8.54	0.2614	8.55	0.3277	5.27
cholesky term	0	NA	0	NA	0	NA	0.158	1.29

Heterogeneity across people and across choices

Heterogeneity across people and across choices

Different possibilities

- Main scope for heterogeneity is across decision makers
- But also scope across choices for the same individual
  - choices at different points in time, or for different types of choices (e.g. long and short trips)
- With deterministic heterogeneity, easy to accommodate both at the same time
- With random heterogeneity, main interest is across people, but can also accommodate intra-individual heterogeneity

Key reference: *Hess, S. & Train, K.E. (2011), Recovery of inter- and intra-personal heterogeneity using mixed logit models, Transportation Research Part B, 45(7), pp. 973-990.*

Example with latent class: *Song, F., Hess, S. & Dekker, T. (2023), Uncovering the link between intra-individual heterogeneity and variety seeking: the case of new shared mobility, Transportation, forthcoming.*

Mixed Logit: overview

Mixed Logit: overview

Introduction

- Mixed Logit is the key example of a model allowing for continuous random heterogeneity
- Very powerful model, widely used in academia and practice
- The material in this session is more **advanced** and **theoretical**, due to the very nature of the model
- You do not necessarily need to understand all the mathematical detail
- But it is important to understand that this is a **complex** model and that analyst decisions have major impacts on results

Mixed Logit: overview

Setting the scene

- Two treatments, with a simple time/money trade-off
- Lower income people have a higher cost sensitivity and lower time sensitivity
- With deterministic heterogeneity, can calculate probability for treatment choices for both groups of patients
  - these are probabilities *conditional* on observing income, and hence the time & cost sensitivity according to the model
- The probabilities in the population are distributed according to the size of the two groups of patients
- But we also know the location of each person!

Treatment	**1**	**2**
Wait (days)	28	14
Cost (£)	100	250

Low income (60% of sample)

β_t	-0.04	
β_c	-0.01	
V	-2.12	-3.06
P	0.72	0.28

High income (40% of sample)

β_t	-0.06	
β_c	-0.005	
V	-2.18	-2.09
P	0.48	0.52

Mixed Logit: overview

Distribution of probabilities

- Let β_n give the (vector of) sensitivities for person n
- If we "know" β_n, we can calculate probabilities
  - e.g. with linear-in-attributes MNL, we have $P_{in}(\beta_n, x) = \frac{e^{\beta_n x_{in}}}{\sum_{j=1}^{J} e^{\beta_n x_{jn}}}$
- In models with deterministic heterogeneity, we observe the source of heterogeneity
  - we can then calculate person-specific probabilities and can also show the distribution of probabilities across the sample
- Problem: with random heterogeneity, we only have the distribution, not each person's location on that distribution

Mixed Logit: overview

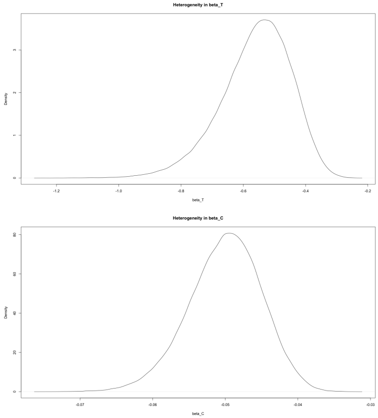
Mixed Logit in a nutshell

- Let $P_{in}(\beta_n, x)$ again be the probability of person n choosing alternative i
- The value of β_n is now not "observed", but only known up to a probability
- In particular, we have that β_n follows a continuous (multivariate) distribution over individuals, i.e. $\beta_n \sim f(\beta_n \mid \Omega)$
- We know from MNL that if β_n varies across people, then so do the probabilities
- In Mixed Logit, this simply implies that the probabilities follow a continuous distribution across individuals

Mixed Logit: overview

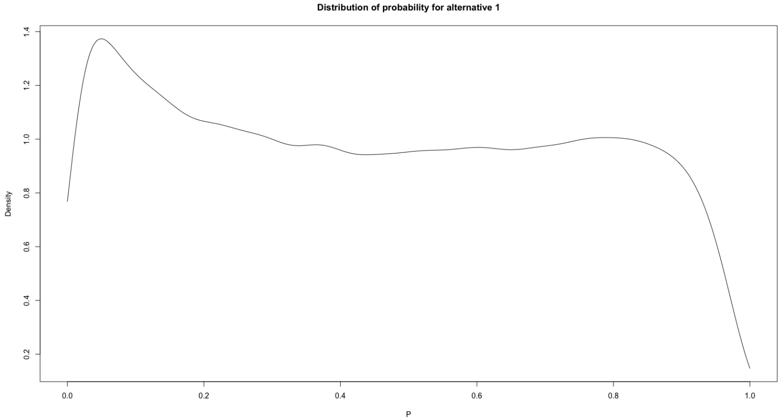
Illustration

- Negative lognormal distribution for waiting time and cost coefficient
- Ensures purely negative response to time and cost
- What does this mean for the choice probabilities?



Mixed Logit: overview

Resulting probability for alternative 1



Mixed Logit: model specification

Mixed Logit: model specification

Key decisions

- An analyst needs to decide:
  - which model parameters follow random distributions
  - what distributions are used
  - whether univariate or multivariate distributions are used
- These decisions have major impacts on model results and interpretation

Mixed Logit: model specification

Random parameters

- In theory, we should allow for random heterogeneity in all parameters
  - this would let the data speak
  - and avoid misattribution
- In practice, we need to consider empirical identification (data limitations) and computational costs
- Should think carefully which parameters are most likely to have variation in a population

Mixed Logit: model specification

Distributional assumptions

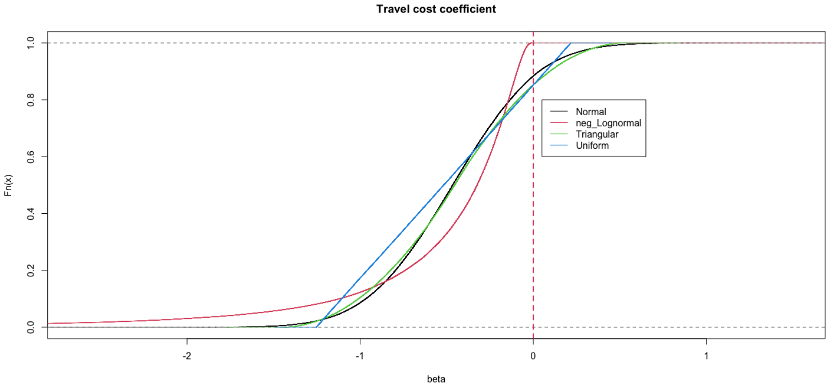
- Too many applications by default rely on Normal distributions
  - unbounded, and behaviourally not meaningful in many cases
  - problems in computing MRS/WTP
- Many other options exist
  - lognormal distribution (exponential of a Normal)
  - triangular distribution (sum of two independent uniforms with same support)
  - ...
- True shape can only be revealed by moving away from parametric distributions

Reference on inappropriate distributions: *Hess, S., Bierlaire, M. & Polak, J.W. (2005), Estimation of value of travel-time savings using Mixed Logit models, Transportation Research Part A, 39(2-3), pp. 221-236.*

Reference on flexible distributions: *Fosgerau, S. & Mabit, S. (2013), Easy and flexible mixture distributions, Economics Letters 120 (2), 206-210.*

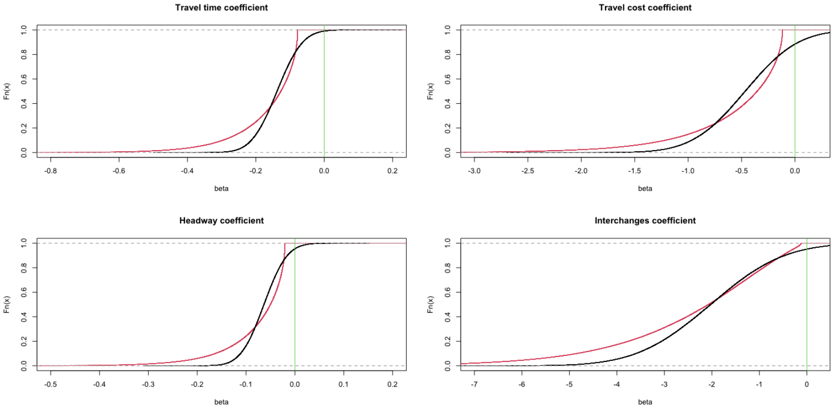
Mixed Logit: model specification

Example with parametric distributions



Mixed Logit: model specification

Non-parametric distribution confirms issues



Mixed Logit: model specification

Multi-variate distributions

- The majority of applications rely on univariate distributions
- In practice, this may not be reasonable
  - people who care more about time may care less about cost, and vice versa
  - some people may be more sensitive overall than others
- Multi-variate distributions improve fit, reduce bias and allow model to allow for scale heterogeneity (but of course cannot disentangle it!)

Key reference: *Hess, S. & Train, K.E. (2017), Correlation and scale in mixed logit models, Journal of Choice Modelling, 23, pp. 1-8.*

Mixed Logit: estimation

Mixed Logit: estimation

Maximum likelihood estimation

- Log-likelihood: $LL(\Omega \mid x, Y) = \sum_{n=1}^{N} log\left(P_{jn_n^*}(\Omega, x)\right)$
- MLE: $\widehat{\Omega} = \arg\max_{\Omega} LL(\Omega \mid x, Y)$
- Optimisation requires $P_{jn_n^*}(\Omega, x)\,,\ \forall n$
  - i.e. estimation requires us to calculate the probabilities of the choices in the data
- The issue now is how to calculate the probabilities for choices in a Mixed Logit model

Mixed Logit: estimation

Econometrics

- Let $P_{in}(\beta_n, x)$ be probability of person n choosing alternative i
- We have a continuous distribution of β over individuals, $\beta_n \sim f(\beta_n \mid \Omega)$
- We do not know where on the distribution person n is
- Unconditional (on β_n) choice probability:

$$P_{in}(\Omega, x) = \int_{\beta_n} \left[P_{in}(\beta_n, x)\ f(\beta_n \mid \Omega)\right] \mathrm{d}\beta_n$$

- Probabilities given by an integral without a closed form solution
- Need to use approximation via numerical integration over distributions of β
  - often done using Monte Carlo simulation, giving us a *simulated log-likelihood*
  - can also approximate using Gaussian quadrature or other numerical integration techniques

Mixed Logit: estimation

Parameters

- With MNL (and other fixed coefficients models), we estimate values of β
  - this includes constants, parameters multiplying attributes, interactions, etc
- The situation changes when we include random components in our model, such as random coefficients
- Example: cost coefficient (β_c) follows a random distribution
  - we do not obtain an estimate for β_c
  - we obtain estimates for the parameters of the distribution of β_c, e.g. mean and std dev
- We have $\beta_n \sim f\left(\beta_n \mid \Omega\right)$
  - Ω is a vector of parameters for the multivariate distribution of β in our data
  - we obtain estimates for Ω
  - for any elements of β that are not random, we obtain a point estimate

Mixed Logit: estimation

Likelihood and log-likelihood: cross-sectional

- *"Cross-sectional"* specification with heterogeneity across people and choices

$$L(\Omega \mid x, Y) = \prod_{n=1}^{N} \prod_{t=1}^{T_n} \int_{\beta_{nt}} \left[P_{ntj^*_{nt}} (\beta_n, x) \;\; f(\beta_{nt} \mid \Omega) \right] \mathrm{d}\beta_{nt}$$

$$LL(\Omega \mid x, Y) = \sum_{n=1}^{N} \sum_{t=1}^{T_n} log \left(\int_{\beta_n} \left[P_{ntj^*_{nt}} (\beta_{nt}, x) \;\; f(\beta_n \mid \Omega) \right] \mathrm{d}\beta_{nt} \right)$$

where j^*_{nt} is chosen by person n in situation t

- Issue of an integral without a closed form solution

Mixed Logit: estimation

Numerical integration using simulation

$$LL(\Omega \mid x, Y) = \sum_{n=1}^{N} \sum_{t=1}^{T_n} log \left(\int_{\beta_n} \left[P_{ntj_{nt}^*} (\beta_{nt}, x) \ f(\beta_n \mid \Omega) \right] \mathrm{d}\beta_{nt} \right)$$

- Simulation: approximate integral by averaging over large number of draws
- Let $\beta_{nt}^{(r)}$ with $r = 1, \ldots, R$ be a random (multivariate) draw from $f(\beta_{nt} \mid \Omega)$

$$SLL(\Omega \mid x, Y) = \sum_{n=1}^{N} \sum_{t=1}^{T_n} log \left(\frac{1}{R} \sum_{r=1}^{R} \left[P_{ntj_{nt}^*} \left(\beta_{nt}^{(r)}, x \right) \right] \right)$$

- No weights in average as shape of distribution taken into account when taking draws

Mixed Logit: estimation

Likelihood and log-likelihood: panel specification

- More realistic *"panel"* specification that assume tastes vary across people, but stay constant for same individual

$$L\left(\Omega \mid x, Y\right) = \prod_{n=1}^{N} \int_{\beta_n} \left[\prod_{t=1}^{T_n} P_{ntj_{nt}^*}\left(\beta_n, x\right)\ f\left(\beta_n \mid \Omega\right)\right] \mathrm{d}\beta_n$$

$$LL\left(\Omega \mid x, Y\right) = \sum_{n=1}^{N} log\left(\int_{\beta_n} \left[\prod_{t=1}^{T_n} P_{ntj_{nt}^*}\left(\beta_n, x\right)\ f\left(\beta_n \mid \Omega\right)\right] \mathrm{d}\beta_n\right)$$

where j_{nt}^* is chosen by person n in situation t

- Still have the issue of an integral without a closed form solution

Mixed Logit: estimation

Numerical integration using simulation

$$LL\left(\Omega \mid x, Y\right) = \sum_{n=1}^{N} log\left(\int_{\beta_n}\left[\prod_{t=1}^{T_n} P_{ntj^*_{nt}}\left(\beta_n, x\right)\ f\left(\beta_n \mid \Omega\right)\right]\mathrm{d}\beta_n\right)$$

- Now use draws at person level and approximates probability of sequence of choices
- Let $\beta_n^{(r)}$ with $r = 1, \ldots, R$ be a random (multivariate) draw from $f\left(\beta_n \mid \Omega\right)$

$$SLL\left(\Omega \mid x, Y\right) = \sum_{n=1}^{N} log\left(\frac{1}{R}\sum_{r=1}^{R}\left[\prod_{t=1}^{T_n} P_{ntj^*_{nt}}\left(\beta_n^{(r)}, x\right)\right]\right)$$

Mixed Logit: estimation

Precision *vs* computational cost

- We use an approximation:

$$\boxed{SLL\left(\Omega \mid x, Y\right) = \sum_{n=1}^{N} log\left(\frac{1}{R}\sum_{r=1}^{R}\left[\prod_{t=1}^{T_n} P_{ntj^*_{nt}}\left(\beta_n^{(r)}, x\right)\right]\right)}$$

- Need R times as many calculations as without random β
- High R leads to costly estimation (and application) process
- As $SLL\left(\Omega \mid x, Y\right)$ is an approximation of true $LL\left(\Omega \mid x, Y\right)$, have simulation noise
- Quality of approximation increases as R rises
  - simulated LL closer to *true* LL
  - greater accuracy in estimates of Ω obtained through maximisation of $SLL\left(\Omega \mid x, Y\right)$

Mixed Logit: estimation

Excel example

- See Excel file MMNL_probs.xlsx
- This looks at simulating a single choice probability
- Spreadsheet shows this with different distributional settings
- Check impact of increasing draws up to $5,000$
- Also convince yourself that working with average draw is not the same as averaging probabilities across draws

Mixed Logit: estimation

Impact of simulation noise

- Significant simulation error with low number of draws
  - Even more significant in more complex models
- Would translate into error in log-likelihood
  - means the model we're estimating is not the one we think we're estimating
  - has strong impact on parameter estimates!
  - bad idea to use high number of draws only for final model

Mixed Logit: estimation

Improving precision of simulation

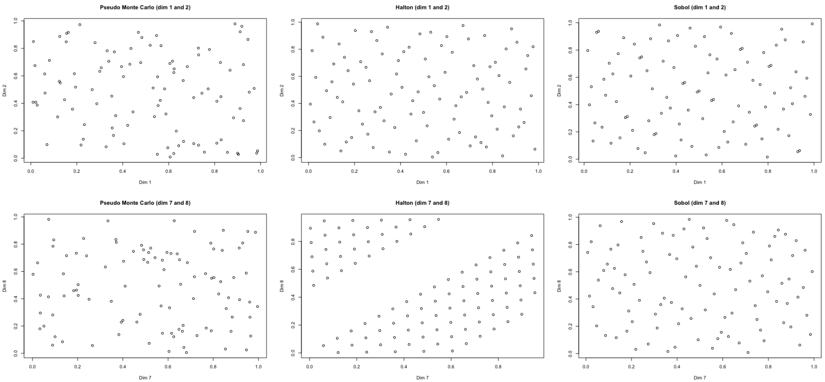
- Crucial to guarantee low level of simulation error
- Use of very high number of random draws impractical
- One solution: use quasi-random draws
  - constructed with aim of greater uniformity
- But don't fall into the trap of thinking a low number of QMC draws is acceptable
- And only use Halton draws with up to 5 random parameters

Key reference: *Bhat, C. (2001), 'Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model', Transportation Research B 35, 677-693.*

Key reference: *Hess, S., Train, K.E. & Polak, J.W. (2006), On the use of a Modified Latin Hypercube Sampling (MLHS) approach in the estimation of a Mixed Logit model for vehicle choice, Transportation Research Part B, 40(2), pp. 147-163.*

Mixed Logit: estimation

Pseudo Monte Carlo and Quasi Monte Carlo



Mixed Logit: estimation

Common questions

- *How many draws should I use to estimate my models with random components?*
  - There is no correct answer to this question. More draws is always better
- *But the log-likelihood of my model is better with fewer draws, so isn't that good?*
  - This in fact shows that fewer draws offers a poor approximation to the real model
  - Once the number of draws is large enough, the model fit will stay much more stable
- *Why does my model converge with a low number of draws, but fails with a high number of draws?*
  - The fact that the model does not converge with a high number of draws shows that there is a problem with the model
  - It is known that using a low number of draws can mean a model that is overspecified still converges and can give every impression of being identified

Illustrative example

Illustrative example

Application to Swiss VTT data

- Binary unlabelled public transport route choice, with alternatives described by travel time (TT), travel cost (TC), headway (HW), interchanges (CH)

$$\boxed{V_{jnt} = \delta_j + \beta_{tt} TT_{jnt} + \beta_{tc} TC_{jnt} + \beta_{hw} HW_{jnt} + \beta_{ch} CH_{jnt}}$$

- For Mixed Logit, use negative lognormal distributions for the four β parameters, e.g.:

$$\boxed{\beta_{tt} = -e^{\mu_{log(\beta_{tt})} + \sigma_{log(\beta_{tt})} \cdot \xi_{tt}}}$$

- Ensures sign of β is purely negative
- $\xi_{tt} \sim N(0,1)$, so sign of σ estimate irrelevant (it's not saying that the sd is negative!)

Illustrative example

Results

- Big improvement in model fit, and all standard deviations different from zero

```
Model name                   : MNL_swiss
Model description            : MNL model on Swiss route choice data
Estimation method            : bgw
Modelled outcomes            : 3492

LL(final)                    : -1665.62
Estimated parameters         : 5


Estimates (robust covariance matrix, 1-sided p-values):

     estimate std. error t-ratio p (1-sided)
asc1  -0.0159     0.0457    -0.3         0.4
b_tt  -0.0598     0.0067    -8.9      <2e-16 ***
b_tc  -0.1317     0.0236    -5.6       1e-08 ***
b_hw  -0.0374     0.0023   -16.2      <2e-16 ***
b_ch  -1.1521     0.0614   -18.8      <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Model name                   : MMNL_swiss_panel_all_negLN
Model description            : MMNL model with negative Lognormal distributions
Estimation method            : bgw
Modelled outcomes            : 3492

LL(final)                    : -1442.84
Estimated parameters         : 9


Estimates (robust covariance matrix, 1-sided p-values):

             estimate std. error t-ratio p (1-sided)
asc1           -0.039      0.071    -0.6       0.289
b_log_tt_mu    -1.985      0.110   -18.1      <2e-16 ***
b_log_tt_sig   -0.527      0.061    -8.7      <2e-16 ***
b_log_tc_mu    -0.961      0.179    -5.4       4e-08 ***
b_log_tc_sig   -0.940      0.069   -13.7      <2e-16 ***
b_log_hw_mu    -2.923      0.090   -32.4      <2e-16 ***
b_log_hw_sig    0.774      0.324     2.4       0.008 **
b_log_ch_mu     0.618      0.082     7.5       3e-14 ***
b_log_ch_sig   -0.887      0.128    -6.9       2e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
                                 LL par
MNL_swiss                  -1665.62   5
MMNL_swiss_panel_all_negLN -1442.84   9
Difference                   222.78   4

Likelihood ratio test-value:    445.56
Degrees of freedom:             4
Likelihood ratio test p-value:  3.96e-95
```

MRS from Mixed Logit

MRS from Mixed Logit

Introduction

- In models that incorporate random heterogeneity, MRS become random too
  - if at least one of the two coefficients varies across people
- If denominator is fixed (e.g. no random heterogeneity in cost sensitivity), then there are no problems
  - MRS given by dividing a random distribution by a constant
  - But bad idea to fix the cost coefficient just to achieve this
- In other cases, MRS given by ratio of random distributions
- Question is how we then calculate moments?

MRS from Mixed Logit

Illustration for Mixed Logit (Lognormal)

- Negative lognormal distributions for all parameters (except ASC)
- $\beta = -e^{\mu+\sigma\xi}$, with $\xi \sim N(0,1)$
- Sign of σ is thus irrelevant
- Simulate the VTT as ratio of 1 million draws each from time and cost coefficient
- Wrong to just take the ratio of means
- With lognormals, MRS also becomes lognormal, and could calculate moments analytically

```
Estimates:
               Estimate        s.e.   t.rat.(0)    Rob.s.e. Rob.t.rat.(0)
asc1           -0.03921     0.06319     -0.6206     0.07062       -0.5552
b_log_tt_mu    -1.98548     0.08787    -22.5964     0.10983      -18.0770
b_log_tt_sig   -0.52693     0.06286     -8.3824     0.06075       -8.6740
b_log_tc_mu    -0.96129     0.13163     -7.3032     0.17945       -5.3569
b_log_tc_sig   -0.93991     0.06874    -13.6726     0.06853      -13.7158
b_log_hw_mu    -2.92323     0.08205    -35.6292     0.09026      -32.3875
b_log_hw_sig    0.77384     0.17082      4.5302     0.32420        2.3870
b_log_ch_mu     0.61831     0.07343      8.4205     0.08210        7.5315
b_log_ch_sig   -0.88725     0.10142     -8.7486     0.12772       -6.9469

> beta_tt = -exp(rnorm(10^6,model$estimate["b_log_tt_mu"],abs(model$estimate["b_log_tt_sig"])))
> beta_tc = -exp(rnorm(10^6,model$estimate["b_log_tc_mu"],abs(model$estimate["b_log_tc_sig"])))
> vtt     = beta_tt / beta_tc * 60
> mean(vtt)
[1] 38.54417
> sd(vtt)
[1] 56.77404
```

MRS from Mixed Logit

Illustration for Mixed Logit (Normal)

- Too many applications still by default use Normal distributions
- Unbounded and symmetrical
- May not be behaviourally realistic
- In simulating the ratio, we already see an odd standard deviation

```
Estimates:
             Estimate        s.e.   t.rat.(0)    Rob.s.e. Rob.t.rat.(0)
asc1         -0.04528    0.060399     -0.7497    0.067188       -0.6740
b_tt_mu      -0.13444    0.010750    -12.5054    0.013614       -9.8753
b_tt_sig     -0.05560    0.007889     -7.0478    0.008118       -6.8486
b_tc_mu      -0.41762    0.038631    -10.8107    0.051996       -8.0318
b_tc_sig     -0.37026    0.039354     -9.4086    0.052696       -7.0264
b_hw_mu      -0.06145    0.004400    -13.9668    0.005288      -11.6214
b_hw_sig     -0.03813    0.005373     -7.0963    0.006797       -5.6094
b_ch_mu      -2.01883    0.121406    -16.6287    0.143891      -14.0303
b_ch_sig      1.18199    0.130220      9.0768    0.143263        8.2504

> beta_tt = rnorm(10^6,model$estimate["b_tt_mu"],abs(model$estimate["b_tt_sig"]))
> beta_tc = rnorm(10^6,model$estimate["b_tc_mu"],abs(model$estimate["b_tc_sig"]))
> vtt     = beta_tt / beta_tc * 60
> mean(vtt)
[1] 11.96089
> sd(vtt)
[1] 9356.123
```

MRS from Mixed Logit

The problem with Normals

- Every time we do this simulation, we get different results
- Including some with the wrong signs
- And this is with one million draws

```
> beta_tt = rnorm(10^6,model$estimate["b_tt_mu"],abs(model$estimate["b_tt_sig"]))
> beta_tc = rnorm(10^6,model$estimate["b_tc_mu"],abs(model$estimate["b_tc_sig"]))
> vtt     = beta_tt / beta_tc * 60
> mean(vtt)
[1] 21.34947
> sd(vtt)
[1] 18455.51
>
> beta_tt = rnorm(10^6,model$estimate["b_tt_mu"],abs(model$estimate["b_tt_sig"]))
> beta_tc = rnorm(10^6,model$estimate["b_tc_mu"],abs(model$estimate["b_tc_sig"]))
> vtt     = beta_tt / beta_tc * 60
> mean(vtt)
[1] 35.44493
> sd(vtt)
[1] 18170.47
>
> beta_tt = rnorm(10^6,model$estimate["b_tt_mu"],abs(model$estimate["b_tt_sig"]))
> beta_tc = rnorm(10^6,model$estimate["b_tc_mu"],abs(model$estimate["b_tc_sig"]))
> vtt     = beta_tt / beta_tc * 60
> mean(vtt)
[1] 58.58022
> sd(vtt)
[1] 25320.84
>
> beta_tt = rnorm(10^6,model$estimate["b_tt_mu"],abs(model$estimate["b_tt_sig"]))
> beta_tc = rnorm(10^6,model$estimate["b_tc_mu"],abs(model$estimate["b_tc_sig"]))
> vtt     = beta_tt / beta_tc * 60
> mean(vtt)
[1] -16.85933
> sd(vtt)
[1] 30459.87
```

MRS from Mixed Logit

(In)existence of moments

- No difficulties if denominator is kept fixed
- Issues arise as soon as denominator is random
- Mathematical proofs show that moments of WTP distribution do not exist for many common choices of distribution (Daly, Hess & Train, 2012)
- Many analysts still ignore this, and just simulate ratio
  - this may mask the whole issue!
- Issues with extreme values arise if distribution of β_C includes zero or approaches zero at a given rate

Key reference: *Daly, A.J., Hess, S. & Train, K.E. (2012), Assuring finite moments for willingness to pay in random coefficients models, Transportation 39(1), pp. 19-31.*

MRS from Mixed Logit

Observations

- Values close to zero will lead to extreme values
- Presence of positive values also causes problems
- Redoing simulation will give different results each time
- Medians of course exist, but they are not what you want for economic appraisal
- Censoring by removing wrongly signed values is very arbitrary
- From a behavioural/micro-economic perspective, do not want Normal for cost coefficient anyway
- But using a fixed coefficient also gives an inferior model
- Two solutions:
  - appropriate distributional assumptions
  - WTP space

Error components

Error components

ECL introduction

- Model structure described thus far for Mixed Logit:
  - Random Coefficients Logit (RCL) formulation of MMNL
  - Exploits MMNL structure to allow for random taste heterogeneity
- MMNL does not exhibit IIA assumption by default, but that does not imply that correlations are captured automatically
- But: MMNL model can be exploited in other ways
- Error components Logit (ECL) formulation
  - Exploits MMNL structure to allow for:
    - Heteroscedasticity
    - Inter-alternative correlation
  - Accommodated through additional $N(0, \sigma)$ random terms

Error components

Heteroskedasticity

- Three alternatives, with different variances for error terms

$$U_A = V_A + \sigma_1 \xi_1 + \varepsilon_A,\ U_B = V_B + \sigma_2 \xi_2 + \varepsilon_B,\ U_C = V_C + \varepsilon_C$$

  where $\xi_j \sim N(0,1)$
- Normalisation required for identification with cross-sectional data (and empirically potentially also with panel data), hence no additional term for one of the alternatives

$$Var(U_A) = \sigma_1^2 + \frac{\pi^2}{6},\ Var(U_B) = \sigma_2^2 + \frac{\pi^2}{6},\ Var(U_C) = \frac{\pi^2}{6}$$

- Easy to implement in *Apollo* by creating random terms that are not multiplying an attribute in the utilities

Error components

Inter-alternative correlation

- Approximation to NL model with A and B nested together

$$U_A = V_A + \sigma_1 \xi_1 + \varepsilon_A, U_B = V_B + \sigma_1 \xi_1 + \varepsilon_B$$

$$U_C = V_C + \varepsilon_C$$

  where $\xi_1 \sim N(0,1)$

- $Cov(U_A, U_B) = \sigma_1^2;\; Corr(U_A, U_B) = \frac{\sigma_1^2}{\sigma_1^2 + \frac{\pi^2}{6}}$
- But this specification introduces heteroskedasticity

$$Var(U_A) = \sigma_1^2 + \frac{\pi^2}{6},\; Var(U_B) = \sigma_1^2 + \frac{\pi^2}{6},\; Var(U_C) = \frac{\pi^2}{6}$$

Error components

Homoskedastic spec with inter-alternative correlation

- Approximation to NL model with A and B nested together

$$U_A = V_A + \sigma_1 \xi_1 + \varepsilon_A, U_B = V_B + \sigma_1 \xi_1 + \varepsilon_B$$

$$U_C = V_C + \sigma_1 \xi_2 + \varepsilon_C$$

  where $\xi_j \sim N(0,1)$
- $Cov(U_A, U_B) = \sigma_1^2$; $Corr(U_A, U_B) = \frac{\sigma_1^2}{\sigma_1^2 + \frac{\pi^2}{6}}$
- $Var(U_j) = \sigma_1^2 + \frac{\pi^2}{6}$
- Similar approach for more complex structures, but with additional identification rules

ECL discussion

- ECL model can account for
  - Inter-alternative correlation
  - Heteroscedasticity
  - Random taste heterogeneity (in joint RCL/ECL structure)
- Joint RCL/ECL model can theoretically approximate any RUM
- But:
  - additional error components lead to rise in computational cost
  - additional identification issues arise (see Walker et al., 2007)
- Should explain as much of correlation as possible using GEV structure (not always possible)

Error components

GEV mixture models

- ECL models allow for joint representation of:
  - random taste heterogeneity
  - inter-alternative correlation
  - heteroscedasticity
- Great flexibility, and reduced risk of confounding
- But:
  - estimation cost can be prohibitive
  - major issues in specification
- Solution:
  - use general GEV mixture models

Error components

Same idea as for mixed logit

- RCL model

$$P_i(\Omega) = \int_\beta \left[P_i(\beta)\, f(\beta \mid \Omega)\right] \mathrm{d}\beta$$

- Conditional on distribution of β, obtain MNL model
- Let $P_i(\beta, \lambda)$ give choice probabilities for NL model
- Then:

$$P_i(\Omega, \lambda) = \int_\beta \left[P_i(\beta, \lambda)\, f(\beta \mid \Omega)\right] \mathrm{d}\beta$$

  gives choice probabilities for NL mixture model
- Correlation explained by underlying NL model
- Random taste heterogeneity explained by mixture

Error components

Other structures

- Can have mixtures of NL, CNL, etc...
- Advantages:
  - No additional error components for correlation
    - lower computational cost
  - Fewer identification issues
- But:
  - Still need error components for heteroscedasticity
  - GEV structures cannot capture all correlation
  - More complicated integrand

Summary

Summary

Key points from this class

- Mixed Logit is incredibly powerful tool
- Here, focus has been on random coefficients
- Can also use for error components, to capture correlation and heteroskedasticity
- Analyst decisions have major impacts on model results and interpretation

Questions?

Apollo

www.ApolloChoiceModelling.com

The most flexible choice modelling software (up to a probability)

Posterior analysis

Posterior analysis

Additional material

Core idea

- Latent class and mixed logit give findings that show how coefficients vary across individuals in our sample
- Further insights are possible by moving from the unconditional (i.e. sample population level) distribution to a conditional distribution
- Equates to inferring the **most likely** position of each sampled individual on the distribution of sensitivities
- Some people call this individual level parameters
  - this is not factually correct!
  - individual-level parameters would require an individual level model
  - the results here still come from a sample level model
  - and they have a lot of uncertainty attached to them

Posterior analysis

Posteriors from Mixed Logit

- Let β give a vector of coefficients that are jointly distributed according to $f(\beta \mid \Omega)$
- Let Y_n give the sequence of observed choices for individual n, with $L(Y_n \mid \beta)$ giving the probability of observing this sequence of choices with a specific value for the vector β
- Probability of observing the specific value of β given the choices of individual n is then given by

$$L(\beta \mid Y_n) = \frac{L(Y_n \mid \beta)\, f(\beta \mid \Omega)}{\int_\beta L(Y_n \mid \beta)\, f(\beta \mid \Omega)\, \mathrm{d}\beta}$$

- Posterior distribution given by an integral without a closed form solution

Posterior analysis

Additional material

Conditional means

- Can approximate moments using numerical simulation
- Mean for the conditional distribution for individual n:

$$\widehat{\beta_n} = \frac{\sum_{r=1}^{R} \left[L\left(Y_n \mid \beta_r\right) \beta_r \right]}{\sum_{r=1}^{R} L\left(Y_n \mid \beta_r\right)},$$

  where β_r with $r = 1, \ldots, R$ are independent multi-dimensional draws from $f\left(\beta \mid \Omega\right)$
- $\widehat{\beta_n}$ gives the posterior mean for various marginal utility coefficients, conditional on choices for individual n

Conditional standard deviations

- Posteriors for each individual follow a random distribution, and conditional means simply give the expected values of this distribution
  - not the *actual* sensitivities for that individual
- Distribution of conditional means across individuals is not a distribution of sensitivities across individuals!
- Aggregating full conditional distributions across individuals yields unconditional distribution
- Conditional standard deviation given by:

$$\widetilde{\beta}_n = \sqrt{\frac{\sum_{r=1}^{R}\left[L\left(Y_n \mid \beta_r\right)\left(\beta_r - \widehat{\beta}_n\right)^2\right]}{\sum_{r=1}^{R} L\left(Y_n \mid \beta_r\right)}}$$

Posterior analysis

Common points for Latent Class and Mixed Logit

- Conditionals can be used for further analysis (e.g. clustering analysis)
- But shape of the unconditional distribution can limit insights
  - conditionals still depend on sample level model
  - classes and distributions
- Need many choices per individual to gain a lot of information for conditionals