Mixture Models — Simulation-based Estimation

Michel Bierlaire

Transport and Mobility Laboratory
School of Architecture, Civil and Environmental Engineering
Ecole Polytechnique Fédérale de Lausanne

Outline

1. Mixtures
2. Relaxing the independence assumption
3. Relaxing the identical distribution assumption
4. Taste heterogeneity
5. Latent classes
6. Summary

Mixtures

Mixture probability distribution function

Convex combination of other probability distribution functions.

Property

- Let $f(\varepsilon, \theta)$ be a parametrized family of distribution functions
- Let $w(\theta)$ be a non negative function such that

$$\int_\theta w(\theta) d\theta = 1$$

- Then

$$g(\varepsilon) = \int_\theta w(\theta) f(\varepsilon, \theta) d\theta$$

  is also a distribution function.

Mixtures

We say that g is a w-mixture of f

- If f is a logit model, g is a continuous w-mixture of logit
- If f is a MEV model, g is a continuous w-mixture of MEV

Mixtures

Discrete mixtures

If w_i, $i = 1, \ldots, n$ are non negative weights such that

$$\sum_{i=1}^{n} w_i = 1$$

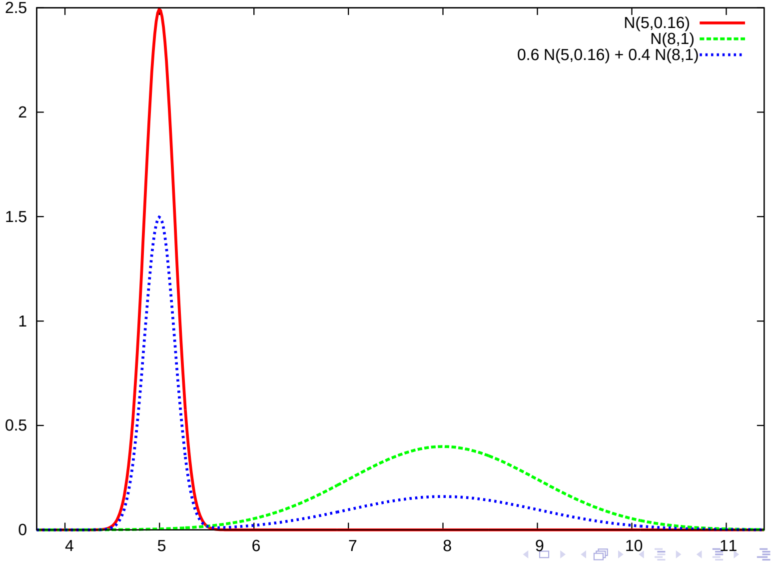
then

$$g(\varepsilon) = \sum_{i=1}^{n} w_i f(\varepsilon, \theta_i)$$

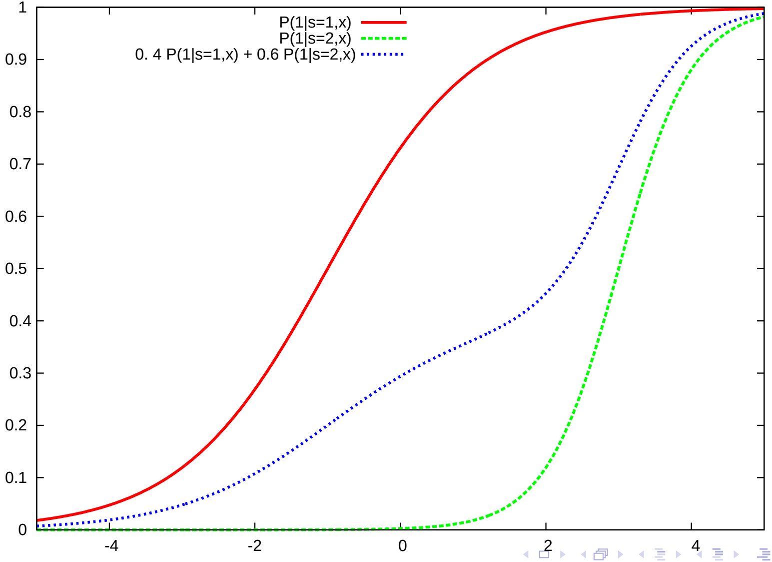
is also a distribution function where θ_i, $i = 1, \ldots, n$ are parameters.

We say that g is a discrete w-mixture of f.

Example: discrete mixture of normal distributions



Example: discrete mixture of binary logit models



Mixtures

General motivation

Generate flexible distributional forms

For discrete choice

- correlation across alternatives
- alternative specific variances
- taste heterogeneity
- ...

Continuous Mixtures of logit

Combining probit and logit

Error components

$$U_{in} = V_{in} + \xi_{in} + \nu_{in}$$

i.i.d EV (logit): tractability

Normal distribution (probit): flexibility

Logit

Specification of the utility functions

$$\begin{array}{rcccl} U_{\text{auto}} & = & \beta X_{\text{auto}} & + & \nu_{\text{auto}} \\ U_{\text{bus}} & = & \beta X_{\text{bus}} & + & \nu_{\text{bus}} \\ U_{\text{subway}} & = & \beta X_{\text{subway}} & + & \nu_{\text{subway}} \end{array}$$

Distributional assumption

ν i.i.d. extreme value

Choice model

$$\Pr(\text{auto}|X,\mathcal{C}) = \frac{e^{\beta X_{\text{auto}}}}{e^{\beta X_{\text{auto}}} + e^{\beta X_{\text{bus}}} + e^{\beta X_{\text{subway}}}}$$

Normal mixture of logit

Specification of the utility functions

$$
\begin{array}{rclclcl}
U_{\text{auto}} & = & \beta X_{\text{auto}} & + & \xi_{\text{auto}} & + & \nu_{\text{auto}} \\
U_{\text{bus}} & = & \beta X_{\text{bus}} & + & \xi_{\text{bus}} & + & \nu_{\text{bus}} \\
U_{\text{subway}} & = & \beta X_{\text{subway}} & + & \xi_{\text{subway}} & + & \nu_{\text{subway}}
\end{array}
$$

Distributional assumptions

- ν i.i.d. extreme value
- $\xi \sim N(0, \Sigma)$

Choice model

$$
\Pr(\text{auto}|X, \xi) = \frac{e^{\beta X_{\text{auto}} + \xi_{\text{auto}}}}{e^{\beta X_{\text{auto}} + \xi_{\text{auto}}} + e^{\beta X_{\text{bus}} + \xi_{\text{bus}}} + e^{\beta X_{\text{subway}} + \xi_{\text{subway}}}}
$$

$$
P(\text{auto}|X) = \int_{\xi} \Pr(\text{auto}|X, \xi) f(\xi) d\xi
$$

Calculation

Choice model

$$P(\text{auto}|X) = \int_{\xi} \Pr(\text{auto}|X, \xi) f(\xi) d\xi$$

Calculation

- Integral has no closed form.
- If one dimension is involved, numerical integration can be used.
- With more dimensions, Monte Carlo simulation must be used.

Simulation

In order to approximate

$$P(i|X) = \int_{\xi} \Pr(i|X, \xi) f(\xi) d\xi$$

- Draw from $f(\xi)$ to obtain $r_1, \ldots, r_R$
- Compute

$$\begin{aligned} P(i|X) \approx \tilde{P}(i|X) &= \frac{1}{R} \sum_{k=1}^{R} P(i|X, r_k) \\ &= \frac{1}{R} \sum_{k=1}^{R} \frac{e^{V_{1n}+r_k}}{e^{V_{1n}+r_k} + e^{V_{2n}+r_k} + e^{V_{3n}}} \end{aligned}$$

Simulation

Can approximate as close as needed

$$P(i|X) = \lim_{R \to \infty} \frac{1}{R} \sum_{k=1}^{R} P(i|X, r_k).$$

In practice

- Efficient methods to draw from the distribution.
- R must be large enough.

Outline

Capturing correlations: nesting

Specification of the utility functions

$$\begin{array}{rclcccl} U_{\text{auto}} & = & \beta X_{\text{auto}} & & & + & \nu_{\text{auto}} \\ U_{\text{bus}} & = & \beta X_{\text{bus}} & + & \sigma_{\text{transit}}\eta_{\text{transit}} & + & \nu_{\text{bus}} \\ U_{\text{subway}} & = & \beta X_{\text{subway}} & + & \sigma_{\text{transit}}\eta_{\text{transit}} & + & \nu_{\text{subway}} \end{array}$$

Distributional assumptions

- ν i.i.d. extreme value,
- $\eta_{\text{transit}} \sim N(0,1)$, $\sigma^2_{\text{transit}}$ =cov(bus,subway)

Choice model

$$\Pr(\text{auto}|X,\eta_{\text{transit}}) = \frac{e^{\beta X_{\text{auto}}}}{e^{\beta X_{\text{auto}}} + e^{\beta X_{\text{bus}}+\sigma_{\text{transit}}\eta_{\text{transit}}} + e^{\beta X_{\text{subway}}+\sigma_{\text{transit}}\eta_{\text{transit}}}}$$

$$P(\text{auto}|X) = \int \Pr(\text{auto}|X,\eta) f(\eta) d\eta$$

Nesting structure

Example: residential telephone

	Ct. BM	Ct. SM	Ct. LF	Ct. EF	β_C	σ_M	σ_F
BM	1	0	0	0	ln(cost(BM))	η_M	0
SM	0	1	0	0	ln(cost(SM))	η_M	0
LF	0	0	1	0	ln(cost(LF))	0	η_F
EF	0	0	0	1	ln(cost(EF))	0	η_F
MF	0	0	0	0	ln(cost(MF))	0	η_F

Nesting structure

Identification issues

- If there are two nests, only one σ is identified
- If there are more than two nests, all σ's are identified

Walker (2001)

Results with 5000 draws

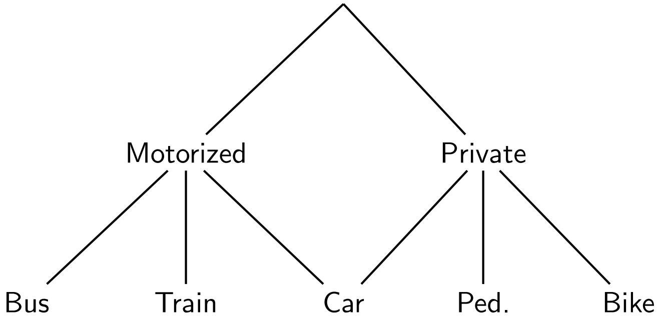
	NL		NML		NML $\sigma_F = 0$		NML $\sigma_M = 0$		NML $\sigma_F = \sigma_M$	
$\mathcal{L}$	-473.219		-472.768		-473.146		-472.779		-472.846	
	Estim.	Scaled	Estim.	Scaled	Estim.	Scaled	Estim.	Scaled	Estim.	Scaled
Ct .BM	-1.78	1.00	-3.81	1.00	-3.79	1.00	-3.81	1.00	-3.81	1.00
Ct. EF	-0.558	0.313	-1.20	0.314	-1.19	0.313	-1.20	0.314	-1.20	0.314
Ct. LF	-0.512	0.287	-1.10	0.287	-1.09	0.287	-1.09	0.287	-1.09	0.287
Ct. SM	-1.41	0.788	-3.02	0.791	-3.00	0.790	-3.01	0.791	-3.02	0.791
β_C	-1.49	0.835	-3.26	0.855	-3.24	0.855	-3.26	0.855	-3.26	0.854
μ_{FLAT}	2.29									
μ_{MEAS}	2.06									
σ_F			3.02		0.00		3.06		2.17	
σ_M			0.530		3.02		0.00		2.17	
$\sigma_F^2 + \sigma_M^2$			9.40		9.15		9.37		9.43	

Comments

- The scale of the parameters is different between NL and the mixture model
- Normalization can be performed in several ways
  - $\sigma_F = 0$
  - $\sigma_M = 0$
  - $\sigma_F = \sigma_M$
- Final log likelihood should be the same
- But... estimation relies on simulation
- Only an approximation of the log likelihood is available
- Final log likelihood with 50000 draws:

Unnormalized:	-472.872	$\sigma_M = \sigma_F$:	-472.875
$\sigma_F = 0$:	-472.884	$\sigma_M = 0$:	-472.901

Cross nesting



Cross nesting

Specification

$$\begin{array}{rcllll} U_{\text{bus}} & = & V_{\text{bus}} & +\xi_1 & & +\varepsilon_{\text{bus}} \\ U_{\text{train}} & = & V_{\text{train}} & +\xi_1 & & +\varepsilon_{\text{train}} \\ U_{\text{car}} & = & V_{\text{car}} & +\xi_1 & +\xi_2 & +\varepsilon_{\text{car}} \\ U_{\text{ped}} & = & V_{\text{ped}} & & +\xi_2 & +\varepsilon_{\text{ped}} \\ U_{\text{bike}} & = & V_{\text{bike}} & & +\xi_2 & +\varepsilon_{\text{bike}} \end{array}$$

Choice model

$$P(\text{car}) = \int_{\xi_1} \int_{\xi_2} P(\text{car}|\xi_1, \xi_2) f(\xi_1) f(\xi_2) d\xi_2 d\xi_1$$

Identification issue

- Not all parameters can be identified
- For logit, one ASC has to be constrained to zero
- Identification of NML is important and tricky
- See Walker, Ben-Akiva & Bolduc (2007) for a detailed analysis

Outline

Alternative specific variance

Logit: i.i.d. error terms

- In particular, they have the same variance

$$U_{in} = \beta^T x_{in} + \mathrm{ASC}_i + \varepsilon_{in}$$

- ε_{in} i.i.d. $EV(0,\mu) \Rightarrow \mathrm{Var}(\varepsilon_{in}) = \pi^2/6\mu^2$

Relax the identical distribution assumption

$$U_{in} = \beta^T x_{in} + \mathrm{ASC}_i + \sigma_i \xi_i + \varepsilon_{in}$$

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

Variance

$$\mathrm{Var}(\sigma_i \xi_i + \varepsilon_{in}) = \sigma_i^2 + \frac{\pi^2}{6\mu^2}$$

Alternative specific variance

Identification issue

- Not all σs are identified
- One of them must be constrained to zero
- Not necessarily the one associated with the ASC constrained to zero
- In theory, the smallest σ must be constrained to zero
- In practice, we don't know a priori which one it is
- Solution:
  1. Estimate a model with a full set of σs
  2. Identify the smallest one and constrain it to zero.

Alternative specific variance

Example with Swissmetro

	ASC_CAR	ASC_SBB	ASC_SM	B_COST	B_FR	B_TIME
Car	1	0	0	cost	0	time
Train	0	0	0	cost	freq.	time
Swissmetro	0	0	1	cost	freq.	time

+ alternative specific variance

Comparison (using 500 draws)

	Logit		ASV		ASV norm.	
$\mathcal{L}$	-5315.39		-5240.414		-5240.414	
	Estim.	Scaled	Estim.	Scaled	Estim.	Scaled
ASC_CAR	0.189	-0.175	0.248	-0.140	0.248	-0.140
ASC_SM	0.451	-0.418	0.900	-0.508	0.901	-0.509
B_COST	-1.08	1.00	-1.77	1.00	-1.77	1.00
B_FR	-5.35	4.95	-7.78	4.40	-7.78	4.40
B_TIME	-1.28	1.19	-1.71	0.966	-1.71	0.966
SIGMA_CAR			0.0107			
SIGMA_TRAIN			0.0284		0.0282	
SIGMA_SM			-3.21		-3.22	

Identification issue: process

Examine the variance-covariance matrix

1. Specify the model of interest
2. Take the differences in utilities
3. Apply the order condition: necessary condition
4. Apply the rank condition: sufficient condition
5. Apply the equality condition: verify equivalence

Heteroscedastic: specification

Model

$$
\begin{array}{lllllll}
U_1 & = & \beta x_1 & +\sigma_1\xi_1 & & & +\varepsilon_1 \\
U_2 & = & \beta x_2 & & +\sigma_2\xi_2 & & & +\varepsilon_2 \\
U_3 & = & \beta x_3 & & & +\sigma_3\xi_3 & & +\varepsilon_3 \\
U_4 & = & \beta x_4 & & & & +\sigma_4\xi_4 & +\varepsilon_4
\end{array}
$$

where $\xi_i \sim N(0,1)$, $\varepsilon_i \sim EV(0,\mu)$

Covariance matrix

$$
\mathrm{Cov}(U) = \begin{pmatrix}
\sigma_1^2 + \gamma/\mu^2 & 0 & 0 & 0 \\
0 & \sigma_2^2 + \gamma/\mu^2 & 0 & 0 \\
0 & 0 & \sigma_3^2 + \gamma/\mu^2 & 0 \\
0 & 0 & 0 & \sigma_4^2 + \gamma/\mu^2
\end{pmatrix}
$$

Heteroscedastic: differences

Utility differences

$$
\begin{aligned}
U_1 - U_4 &= \beta(x_1 - x_4) + (\sigma_1\xi_1 - \sigma_4\xi_4) + (\varepsilon_1 - \varepsilon_4) \\
U_2 - U_4 &= \beta(x_2 - x_4) + (\sigma_2\xi_2 - \sigma_4\xi_4) + (\varepsilon_2 - \varepsilon_4) \\
U_3 - U_4 &= \beta(x_3 - x_4) + (\sigma_3\xi_3 - \sigma_4\xi_4) + (\varepsilon_3 - \varepsilon_4)
\end{aligned}
$$

Covariance of utility differences

$\mathrm{Cov}(\Delta U) =$

$$
\begin{pmatrix}
\sigma_1^2 + \sigma_4^2 + 2\gamma/\mu^2 & \sigma_4^2 + \gamma/\mu^2 & \sigma_4^2 + \gamma/\mu^2 \\
\sigma_4^2 + \gamma/\mu^2 & \sigma_2^2 + \sigma_4^2 + 2\gamma/\mu^2 & \sigma_4^2 + \gamma/\mu^2 \\
\sigma_4^2 + \gamma/\mu^2 & \sigma_4^2 + \gamma/\mu^2 & \sigma_3^2 + \sigma_4^2 + 2\gamma/\mu^2
\end{pmatrix}
$$

Heteroscedastic: order condition

Upper bound

- S is the number of estimable parameters
- J is the number of alternatives

$$S \leq \frac{J(J-1)}{2} - 1$$

- It represents the number of entries in the lower part of the (symmetric) var-cov matrix
- minus 1 for the scale
- $J = 4$ implies $S \leq 5$

Heteroscedastic: rank condition

Idea

- Number of estimable parameters =
- number of linearly independent equations
- -1 for the scale

$\mathrm{Cov}(\Delta U) =$

$$\begin{pmatrix} \sigma_1^2+\sigma_4^2+2\gamma/\mu^2 & & \\ \sigma_4^2+\gamma/\mu^2 & \sigma_2^2+\sigma_4^2+2\gamma/\mu^2 & \\ \sigma_4^2+\gamma/\mu^2 & \sigma_4^2+\gamma/\mu^2 & \sigma_3^2+\sigma_4^2+2\gamma/\mu^2 \end{pmatrix}$$

dependent

scale

Heteroscedastic: rank condition

Three parameters out of five can be estimated

Formally...

1. Identify unique elements of $\mathrm{Cov}(\Delta U)$
2. Compute the Jacobian wrt σ_1^2, σ_2^2, σ_3^2, σ_4^2, γ/μ^2
3. Compute the rank

$$\begin{pmatrix} \sigma_1^2 + \sigma_4^2 + 2\gamma/\mu^2 \\ \sigma_2^2 + \sigma_4^2 + 2\gamma/\mu^2 \\ \sigma_3^2 + \sigma_4^2 + 2\gamma/\mu^2 \\ \sigma_4^2 + \gamma/\mu^2 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 & 0 & 1 & 2 \\ 0 & 1 & 0 & 1 & 2 \\ 0 & 0 & 1 & 1 & 2 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

$S = \text{Rank - } 1 = 3$

Heteroscedastic: equality condition

Normalization

- We know how many parameters can be identified
- There are infinitely many normalizations
- The normalized model is equivalent to the original one
- Obvious normalizations, like constraining extra-parameters to 0 or another constant, may not be valid

Heteroscedastic: equality condition

Error components

$$\begin{array}{rcllll} U_n & = & \beta^T x_n & + & L_n\xi_n & + & \varepsilon_n \\ \mathrm{Cov}(U_n) & = & & & L_n L_n^T & + & (\gamma/\mu^2)I \\ \mathrm{Cov}(\Delta_j U_n) & = & & & \Delta_j L_n L_n^T \Delta_j^T & + & (\gamma/\mu^2)\Delta_j\Delta_j^T \end{array}$$

Notations

$$\Delta_2 = \begin{pmatrix} 1 & -1 & 0 \\ 0 & -1 & 1 \end{pmatrix}$$

$$\begin{array}{rcllll} \mathrm{Cov}(\Delta_j U_n) = & \Omega_n & = & \Sigma_n & + & \Gamma_n \\ & \Omega_n^{\text{norm}} & = & \Sigma_n^{\text{norm}} & + & \Gamma_n^{\text{norm}} \end{array}$$

Heteroscedastic: equality condition

The following conditions must hold

- Covariance matrices must be equal

$$\Omega_n = \Omega_n^{\text{norm}}$$

- Σ_n^{norm} must be positive semi-definite

Heteroscedastic: equality condition

Example with 3 alternatives

$$\begin{array}{lclllll} U_1 & = & \beta x_1 & +\sigma_1\xi_1 & & & +\varepsilon_1 \\ U_2 & = & \beta x_2 & & +\sigma_2\xi_2 & & +\varepsilon_2 \\ U_3 & = & \beta x_3 & & & +\sigma_3\xi_3 & +\varepsilon_3 \end{array}$$

$$\mathrm{Cov}(\Delta_3 U) = \Omega = \begin{pmatrix} \sigma_1^2+\sigma_3^2+2\gamma/\mu^2 & \\ \sigma_3^2+\gamma/\mu^2 & \sigma_2^2+\sigma_3^2+2\gamma/\mu^2 \end{pmatrix}$$

- Parameters: $\{\sigma_1, \sigma_2, \sigma_3, \mu\}$
- Rank condition: $S = 2$
- μ is used for the scale

Heteroscedastic: equality condition

Change of variables

- Denote $\nu_i = \sigma_i^2 \mu^2$ (scaled parameters)
- Normalization condition: $\nu_3 = K$

$$\Omega = \begin{pmatrix} (\nu_1 + \nu_3 + 2\gamma)/\mu^2 & \\ (\nu_3 + \gamma)/\mu^2 & (\nu_2 + \nu_3 + 2\gamma)/\mu^2 \end{pmatrix}$$

$$\Omega^{\text{norm}} = \begin{pmatrix} (\nu_1^N + K + 2\gamma)/\mu_N^2 & \\ (K + \gamma)/\mu_N^2 & (\nu_2^N + K + 2\gamma)/\mu_N^2 \end{pmatrix}$$

where index N stands for "normalized"

Heteroscedastic: equality condition

First equality condition: $\Omega = \Omega^{\text{norm}}$

$$
\begin{array}{rcl}
(\nu_3+\gamma)/\mu^2 &=& (K+\gamma)/\mu_N^2 \\
(\nu_1+\nu_3+2\gamma)/\mu^2 &=& (\nu_1^N+K+2\gamma)/\mu_N^2 \\
(\nu_2+\nu_3+2\gamma)/\mu^2 &=& (\nu_2^N+K+2\gamma)/\mu_N^2
\end{array}
$$

that is, writing the normalized parameters as functions of others,

$$
\begin{array}{rcl}
\mu_N^2 &=& \mu^2(K+\gamma)/(\nu_3+\gamma) \\
\nu_1^N &=& (K+\gamma)(\nu_1+\nu_3+2\gamma)/(\nu_3+\gamma)-K-2\gamma \\
\nu_2^N &=& (K+\gamma)(\nu_2+\nu_3+2\gamma)/(\nu_3+\gamma)-K-2\gamma
\end{array}
$$

Heteroscedastic: equality condition

Second equality condition

$$\Sigma^{\text{norm}} = \frac{1}{\mu_N^2} \begin{pmatrix} \nu_1^N & 0 & 0 \\ 0 & \nu_2^N & 0 \\ 0 & 0 & K \end{pmatrix}$$

must be positive semi-definite, that is

$$\mu_N > 0,\ \nu_1^N \geq 0,\ \nu_2^N \geq 0,\ K \geq 0.$$

Putting everything together, we obtain

$$K \geq \frac{(\nu_3 - \nu_i)\gamma}{\nu_i + \gamma},\ i = 1, 2$$

Heteroscedastic: equality condition

Condition to be verified for the normalization to be valid

$$K \geq \frac{(\nu_3 - \nu_i)\gamma}{\nu_i + \gamma}, \; i = 1, 2$$

- If $\nu_3 \leq \nu_i$, $i = 1, 2$, then the rhs is negative, and any $K \geq 0$ would do. Typically, $K = 0$.
- If not, K must be chosen large enough
- In practice, always select the alternative with minimum variance.

Outline

Taste heterogeneity

Motivation

- Population is heterogeneous
- Taste heterogeneity is captured by segmentation
- Deterministic segmentation is desirable but not always possible
- Distribution of a parameter in the population

Random parameters

$$\begin{aligned} U_i &= \beta_t T_i + \beta_c C_i + \varepsilon_i \\ U_j &= \beta_t T_j + \beta_c C_j + \varepsilon_j \end{aligned}$$

Let $\beta_t \sim N(\bar{\beta}_t, \sigma_t^2)$, or, equivalently,

$$\beta_t = \bar{\beta}_t + \sigma_t \xi, \text{ with } \xi \sim N(0,1).$$

$$\begin{aligned} U_i &= \bar{\beta}_t T_i + \sigma_t \xi T_i + \beta_c C_i + \varepsilon_i \\ U_j &= \bar{\beta}_t T_j + \sigma_t \xi T_j + \beta_c C_j + \varepsilon_j \end{aligned}$$

If ε_i and ε_j are i.i.d. EV and ξ is given, we have

$$P(i|\xi) = \frac{e^{\bar{\beta}_t T_i + \sigma_t \xi T_i + \beta_c C_i}}{e^{\bar{\beta}_t T_i + \sigma_t \xi T_i + \beta_c C_i} + e^{\bar{\beta}_t T_j + \sigma_t \xi T_j + \beta_c C_j}}, \text{ and}$$

$$P(i) = \int_\xi P(i|\xi) f(\xi) d\xi.$$

Random parameters

Example with Swissmetro

	ASC_CAR	ASC_SBB	ASC_SM	B_COST	B_FR	B_TIME
Car	1	0	0	cost	0	time
Train	0	0	0	cost	freq.	time
Swissmetro	0	0	1	cost	freq.	time

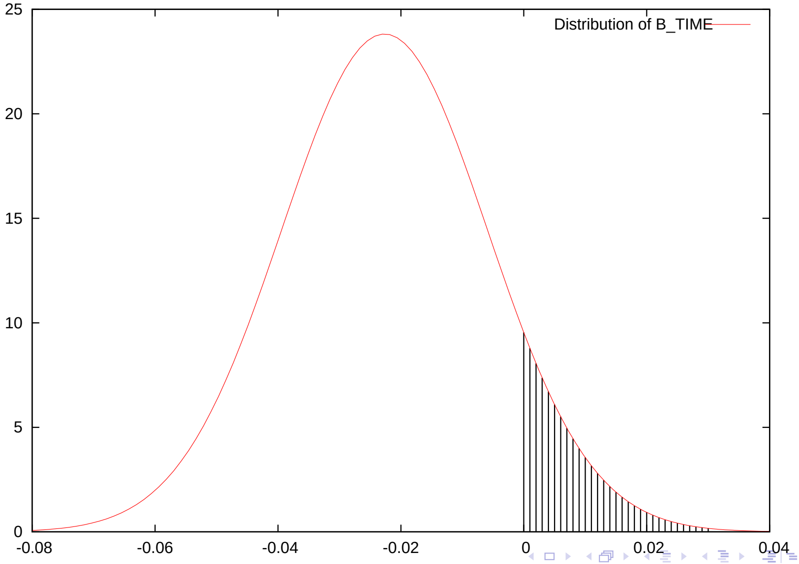
B_TIME randomly distributed across the population, normal distribution

Random parameters

Estimation results

	Logit	RC
$\mathcal{L}$	-5315.4	-5198.0
ASC_CAR_SP	0.189	0.118
ASC_SM_SP	0.451	0.107
B_COST	-0.011	-0.013
B_FR	-0.005	-0.006
B_TIME	-0.013	-0.023
S_TIME		0.017
Prob(B_TIME $\geq$ 0)		8.8%
χ^2		234.84

Random parameters



Random parameters

Example with Swissmetro

	ASC_CAR	ASC_SBB	ASC_SM	B_COST	B_FR	B_TIME
Car	1	0	0	cost	0	time
Train	0	0	0	cost	freq.	time
Swissmetro	0	0	1	cost	freq.	time

B_TIME randomly distributed across the population, log normal distribution

Random parameters

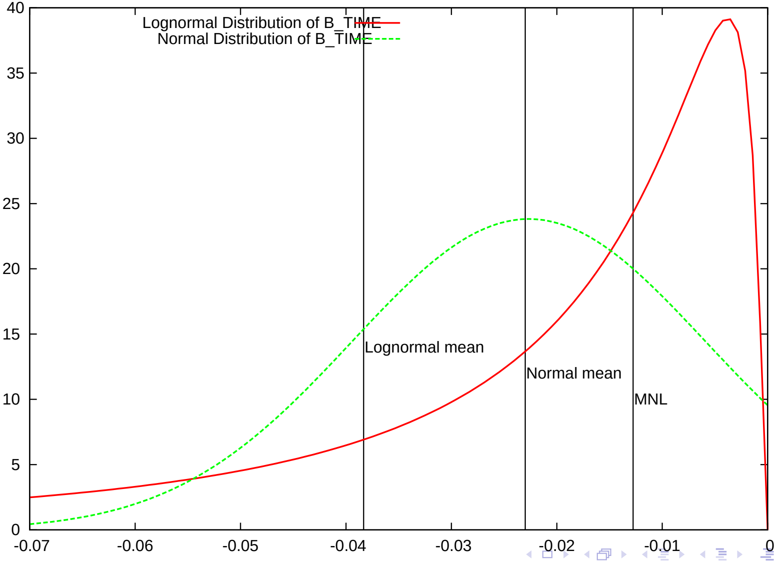
```
[Utilities]
11 SBB_SP TRAIN_AV_SP ASC_SBB_SP * one        +
                      B_COST     * TRAIN_COST +
                      B_FR       * TRAIN_FR
21 SM_SP SM_AV        ASC_SM_SP  * one        +
                      B_COST     * SM_COST    +
                      B_FR * SM_FR
31 Car_SP CAR_AV_SP   ASC_CAR_SP * one        +
                      B_COST     * CAR_CO

[GeneralizedUtilities]
11 - exp( B_TIME [ S_TIME ] ) * TRAIN_TT
21 - exp( B_TIME [ S_TIME ] ) * SM_TT
31 - exp( B_TIME [ S_TIME ] ) * CAR_TT
```

Random parameters

Estimation results

	Logit	RC-norm.	RC-logn.	
	-5315.4	-5198.0	-5215.81	
ASC_CAR_SP	0.189	0.118	0.122	
ASC_SM_SP	0.451	0.107	0.069	
B_COST	-0.011	-0.013	-0.014	
B_FR	-0.005	-0.006	-0.006	
B_TIME	-0.013	-0.023	-4.033	-0.038
S_TIME		0.017	1.242	0.073
Prob($\beta > 0$)		8.8%	0.0%	
χ^2		234.84	199.16	

Random parameters



Random parameters

Example with Swissmetro

	ASC_CAR	ASC_SBB	ASC_SM	B_COST	B_FR	B_TIME
Car	1	0	0	cost	0	time
Train	0	0	0	cost	freq.	time
Swissmetro	0	0	1	cost	freq.	time

B_TIME randomly distributed across the population, discrete distribution

$$P(\beta_{\text{time}} = \hat{\beta}) = \omega_1 \quad P(\beta_{\text{time}} = 0) = \omega_2 = 1 - \omega_1$$

Random parameters

Syntax for Biogeme

```
[DiscreteDistributions]
B_TIME < B_TIME_1 ( W1 ) B_TIME_2 ( W2 ) >

[LinearConstraints]
W1 + W2 = 1.0
```

Random parameters

Estimation results

	Logit	RC-norm.	RC-logn.		RC-disc.
	-5315.4	-5198.0	-5215.8		-5191.1
ASC_CAR_SP	0.189	0.118	0.122		0.111
ASC_SM_SP	0.451	0.107	0.069		0.108
B_COST	-0.011	-0.013	-0.014		-0.013
B_FR	-0.005	-0.006	-0.006		-0.006
B_TIME	-0.013	-0.023	-4.033	-0.038	-0.028
					0.000
S_TIME		0.017	1.242	0.073	
W1					0.749
W2					0.251
Prob($\beta > 0$)		8.8%	0.0%		0.0%
χ^2		234.84	199.16		248.6

Outline

Latent classes

Capture unobserved heterogeneity

They can represent different:

- Choice sets
- Decision protocols
- Tastes
- Model structures
- etc.

Latent classes

Model structure

$$P_n(i|\mathcal{C}_n) = \sum_{s=1}^{S} P_n(i|\mathcal{C}_n, s) Q_n(s)$$

- $P_n(i|\mathcal{C}_n, s)$ is the class-specific choice model
  - probability of choosing i given that the individual n belongs to class s
- $Q_n(s)$ is the class membership model
  - probability of belonging to class s

Outline

Summary

Logit mixtures models

- Computationally more complex than MEV
- Allow for more flexibility than MEV

Continuous mixtures

Alternative specific variance, nesting structures, random parameters

$$P_n(i) = \int_\xi P_n(i|\xi) f(\xi) d\xi$$

Discrete mixtures

Latent classes of decision makers

$$P_n(i|\mathcal{C}_n) = \sum_{s=1}^{S} P_n(i|\mathcal{C}_n, s) Q_n(s)$$

Tips for applications

- Be careful: simulation can mask specification and identification issues
- Do not forget about the systematic portion

Appendix: Simulation

How to calculate?

$$P(i) = \int_{\xi} \Pr(i|\xi) f(\xi) d\xi$$

No closed form formula

Monte Carlo simulation

- Randomly draw numbers such that their frequency matches the density $f(\xi)$
- Let $\xi^1, \ldots, \xi^R$ be these numbers
- The choice model can be approximated by

$$P(i) \approx \frac{1}{R} \sum_{r=1}^{R} \Pr(i|r), \text{ as } \lim_{R \to \infty} \frac{1}{R} \sum_{r=1}^{R} \Pr(i|r) = \int_{\xi} \Pr(i|\xi) f(\xi) d\xi$$

Appendix: Simulation

Approximation

$$P(i) \approx \frac{1}{R} \sum_{r=1}^{R} \Pr(i|r).$$

The kernel is a logit model, easy to compute

$$\Pr(i|r) = \frac{e^{V_{1n}+r}}{e^{V_{1n}+r} + e^{V_{2n}+r} + e^{V_{3n}}}$$

Therefore, it amounts to generating the appropriate draws.

Appendix: Simulation

Pseudo-random numbers generators

Although deterministically generated, numbers exhibit the properties of random draws

- Uniform distribution
- Standard normal distribution
- Transformation of standard normal
- Inverse CDF
- Multivariate normal

Appendix: Simulation

Uniform distribution

- Almost all programming languages provide generators for a uniform $U(0,1)$
- If r is a draw from a $U(0,1)$, then

$$s = (b - a)r + a$$

is a draw from a $U(a,b)$

Appendix: Simulation

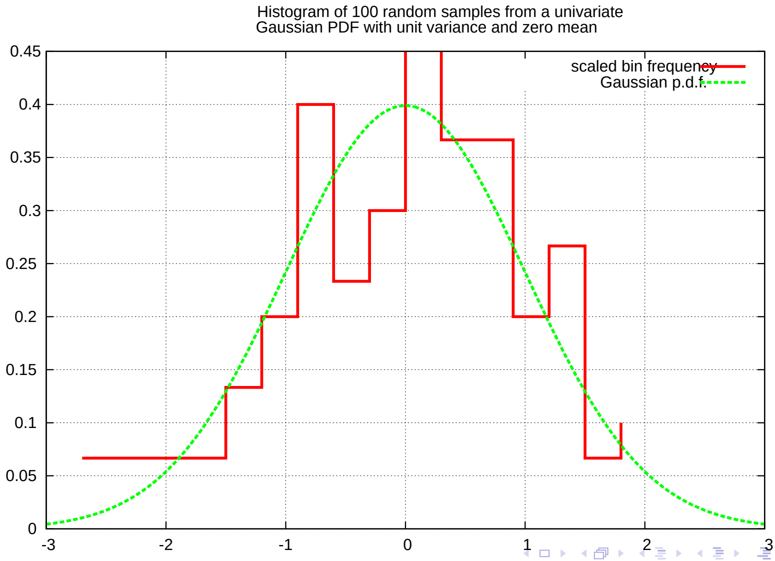
Standard normal

- If r_1 and r_2 are independent draws from $U(0,1)$, then

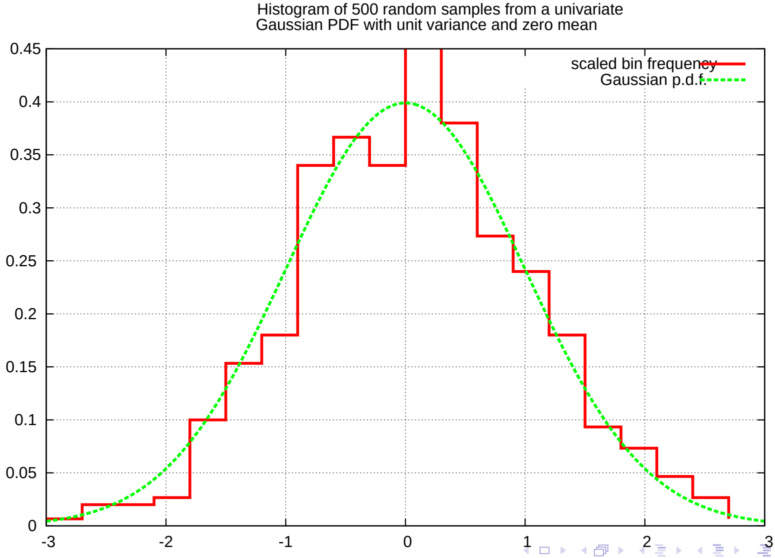
$$\begin{aligned} s_1 &= \sqrt{-2 \ln r_1} \sin(2\pi r_2) \\ s_2 &= \sqrt{-2 \ln r_1} \cos(2\pi r_2) \end{aligned}$$

are independent draws from $N(0,1)$

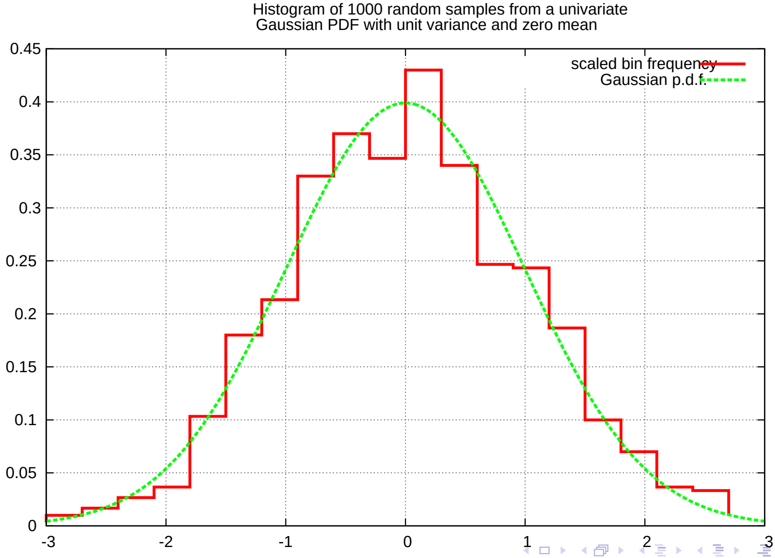
Appendix: Simulation: standard normal



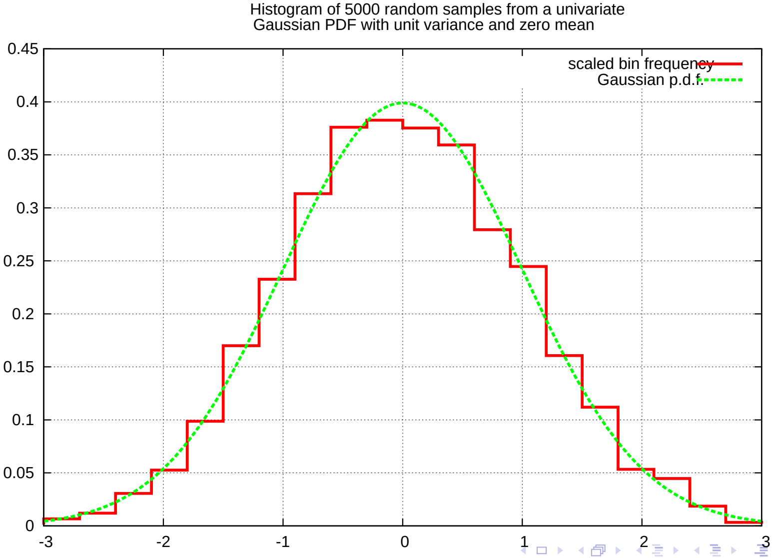
Appendix: Simulation: standard normal



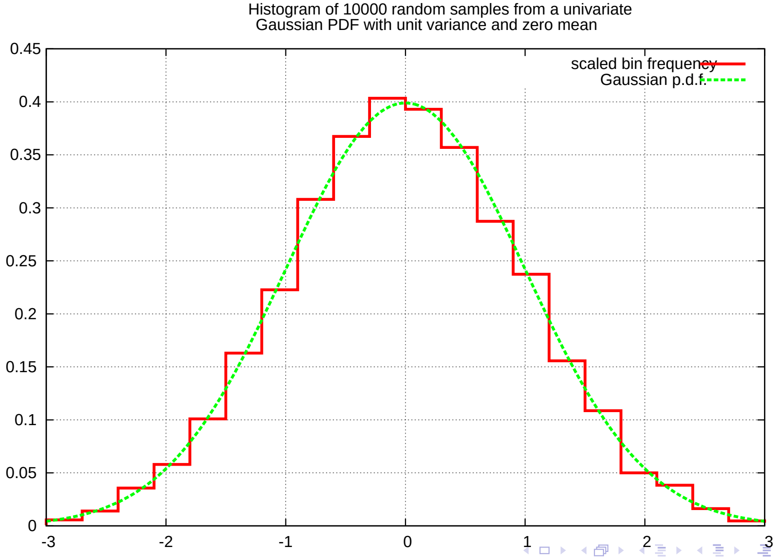
Appendix: Simulation: standard normal



Appendix: Simulation: standard normal



Appendix: Simulation: standard normal



Appendix: Simulation

Normal distribution

If r is a draw from $N(0,1)$, then

$$s = br + a$$

is a draw from $N(a, b^2)$

Log normal distribution

If r is a draw from $N(a, b^2)$, then

$$e^r$$

is a draw from a log normal $LN(a, b^2)$ with mean $e^{a+(b^2/2)}$ and variance $e^{2a+b^2}(e^{b^2}-1)$

Appendix: Simulation

Inverse CDF

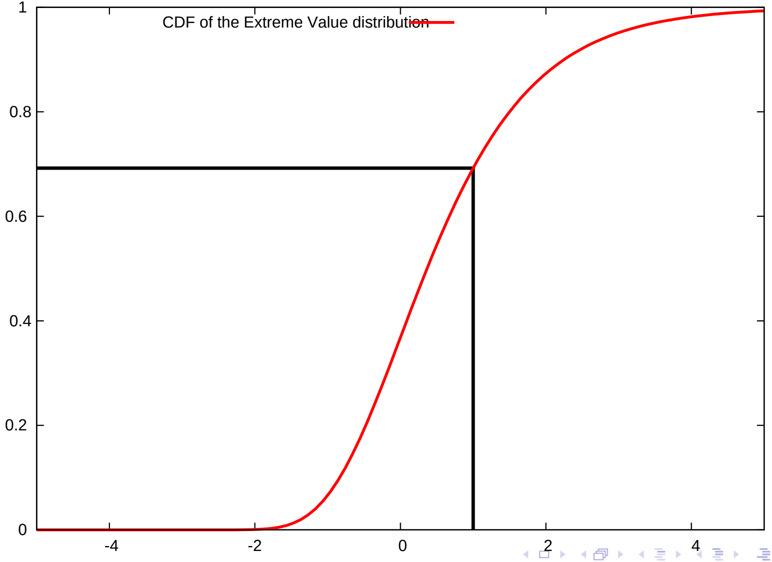
- Consider a univariate r.v. with CDF $F(\varepsilon)$
- If F is invertible and if r is a draw from $U(0,1)$, then

$$s = F^{-1}(r)$$

  is a draw from the given r.v.
- Example: EV with

$$F(\varepsilon) = e^{-e^{-\varepsilon}} \quad F^{-1}(r) = -\ln(-\ln r)$$

Appendix: Simulation: inverse CDF



Appendix: Simulation

Multivariate normal

If $r_1, \ldots, r_n$ are independent draws from $N(0,1)$, and

$$r = \begin{pmatrix} r_1 \\ \vdots \\ r_n \end{pmatrix}$$

then

$$s = a + Lr$$

is a vector of draws from the n-variate normal $N(a, LL^T)$, where

- L is lower triangular, and
- LL^T is the Cholesky factorization of the variance-covariance matrix

Appendix: Simulation

Example

$$L = \begin{pmatrix} \ell_{11} & 0 & 0 \\ \ell_{21} & \ell_{22} & 0 \\ \ell_{31} & \ell_{32} & \ell_{33} \end{pmatrix}$$

$$\begin{array}{lllllll} s_1 & = & \ell_{11} r_1 & & & & \\ s_2 & = & \ell_{21} r_1 & + & \ell_{22} r_2 & & \\ s_3 & = & \ell_{31} r_1 & + & \ell_{32} r_2 & + & \ell_{33} r_3 \end{array}$$

Appendix: Simulation

Mixtures of logit

$$P(i|X) = \int_{\xi} \Pr(i|X,\xi) f(\xi) d\xi$$

- Draw from $f(\xi)$ to obtain $r_1, \ldots, r_R$
- Compute

$$\begin{aligned} P(i|X) \approx \tilde{P}(i|X) &= \frac{1}{R} \sum_{k=1}^{R} P(i|X, r_k) \\ &= \frac{1}{R} \sum_{k=1}^{R} \frac{e^{V_{1n}+r_k}}{e^{V_{1n}+r_k} + e^{V_{2n}+r_k} + e^{V_{3n}}} \end{aligned}$$

Appendix: Maximum simulated likelihood

Solve

$$\max_{\theta} \mathcal{L}(\theta) = \sum_{n=1}^{N} \left(\sum_{j=1}^{J} y_{jn} \ln \tilde{P}(j;\theta) \right)$$

where $y_{jn} = 1$ if ind. n has chosen alt. j, 0 otherwise.

Vector of parameters θ contains

- usual (fixed) parameters of the choice model
- parameters of the density of the random parameters
- For instance, if $\beta_j \sim N(\mu_j, \sigma_j^2)$, μ_j and σ_j are parameters to be estimated

Appendix: Maximum simulated likelihood

Warning

- $\tilde{P}(j;\theta)$ is an unbiased estimator of $P(j;\theta)$

$$E[\tilde{P}_n(j;\theta)] = P(j;\theta)$$

- $\ln \tilde{P}(j;\theta)$ is not an unbiased estimator of $\ln P(j;\theta)$

$$\ln E[\tilde{P}(j;\theta] \neq E[\ln \tilde{P}(j;\theta)]$$

- Under some conditions, it is a consistent (asymptotically unbiased) estimator, so that many draws are necessary.

Appendix: Maximum simulated likelihood

Properties of MSL

- If R is fixed, MSL is inconsistent
- If R rises at any rate with N, MSL is consistent
- If R rises faster than $\sqrt{N}$, MSL is asymptotically equivalent to ML.