Introduction to mixtures in discrete choice models

Michel Bierlaire

michel.bierlaire@epfl.ch

Transport and Mobility Laboratory

Outline

- Mixtures
- Capturing correlation
- Alternative specific variance
- Taste heterogeneity
- Latent classes

Mixtures

In statistics, a mixture probability distribution function is a convex combination of other probability distribution functions.
If $f(\varepsilon, \theta)$ is a distribution function, and if $w(\theta)$ is 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. 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 are also possible. If w_i, $i = 1, \ldots, n$ are non negative weights such that

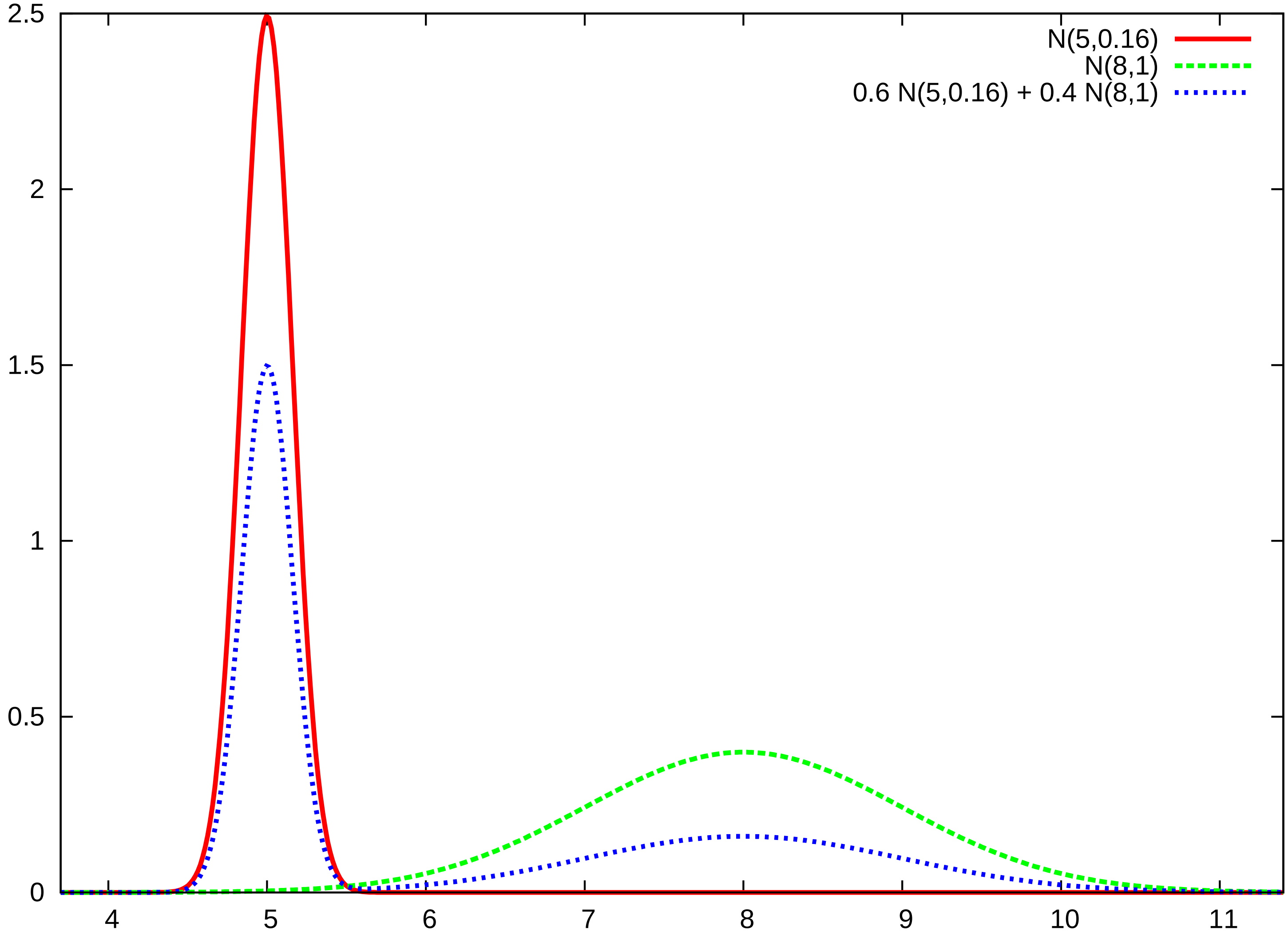
$$\sum_{i=1}^{n} w_i = 1$$

then

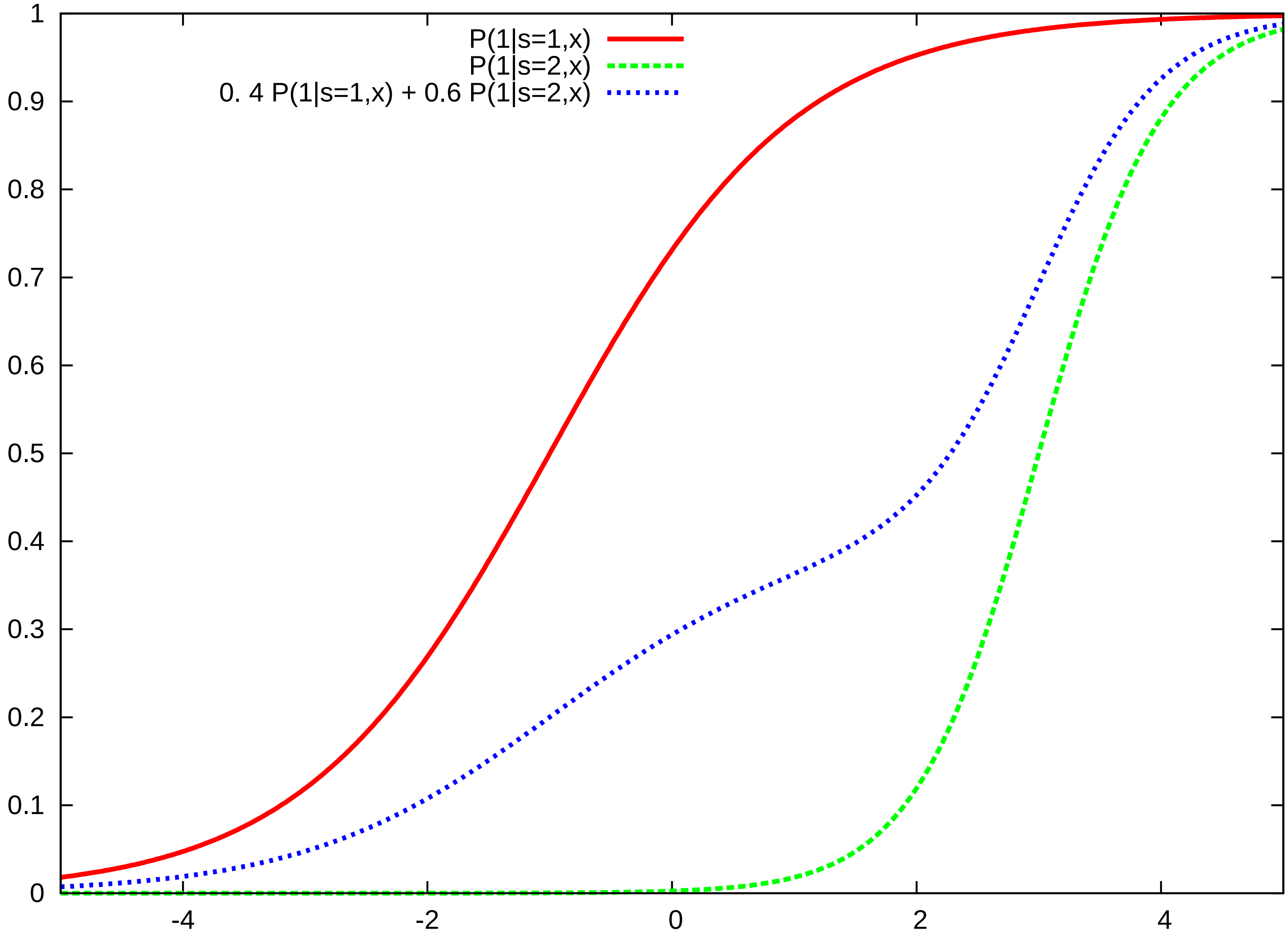
$$g(\varepsilon) = \sum_{i=1}^{n} w_i f(\varepsilon, \theta_i)$$

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 decomposed into two parts

$$U_{in} = V_{in} + \xi + \nu$$

i.i.d EV (logit): tractability

Normal distribution (probit): flexibility

Logit

- Utility:

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

- ν i.i.d. extreme value
- Probability:

$$
\Lambda(\text{auto}|X) = \frac{e^{\beta X_{\text{auto}}}}{e^{\beta X_{\text{auto}}} + e^{\beta X_{\text{bus}}} + e^{\beta X_{\text{subway}}}}
$$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Normal mixture of logit

- Utility:

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

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

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

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

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Capturing correlations: nesting

- Utility:

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

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

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

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

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Nesting structure

Example: residential telephone

	ASC_BM	ASC_SM	ASC_LF	ASC_EF	BETA_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..

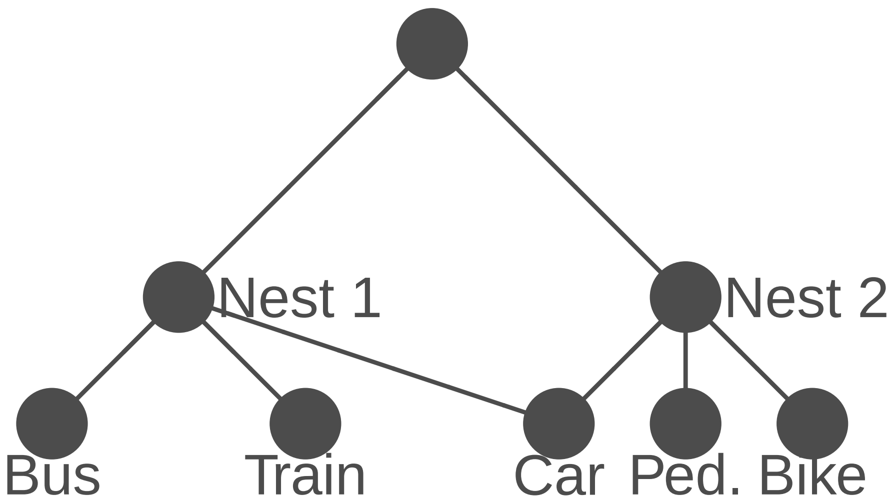
	NL		NML		NML $\sigma_F = 0$		NML $\sigma_M = 0$		NML $\sigma_F = \sigma_M$	
$\mathcal{L}$	-473.219		-472.768		-473.146		-472.779		-472.846	
	Value	Scaled	Value	Scaled	Value	Scaled	Value	Scaled	Value	Scaled
ASC_BM	-1.784	1.000	-3.81247	1.000	-3.79131	1.000	-3.80999	1.000	-3.81327	1.000
ASC_EF	-0.558	0.313	-1.19899	0.314	-1.18549	0.313	-1.19711	0.314	-1.19672	0.314
ASC_LF	-0.512	0.287	-1.09535	0.287	-1.08704	0.287	-1.0942	0.287	-1.0948	0.287
ASC_SM	-1.405	0.788	-3.01659	0.791	-2.9963	0.790	-3.01426	0.791	-3.0171	0.791
B_LOGCOST	-1.490	0.835	-3.25782	0.855	-3.24268	0.855	-3.2558	0.855	-3.25805	0.854
FLAT	2.292									
MEAS	2.063									
σ_F			3.02027		0		3.06144		2.17138	
σ_M			0.52875		3.024833		0		2.17138	
$\sigma_F^2 + \sigma_M^2$			9.402		9.150		9.372		9.430	

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



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

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

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

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

Alternative specific variance

- Error terms in logit are i.i.d. and, in particular, have the same variance

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

- ε_{in} i.i.d. extreme value $\Rightarrow$ $\text{Var}(\varepsilon_{in}) = \pi^2/6\mu^2$
- In order allow for different variances, we use mixtures

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

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

- Variance:

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

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

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

	Logit		ASV		ASV norm.	
$\mathcal{L}$	-5315.39		-5241.01		-5242.10	
	Value	Scaled	Value	Scaled	Value	Scaled
ASC_CAR	0.189	1.000	0.248	1.000	0.241	1.000
ASC_SM	0.451	2.384	0.903	3.637	0.882	3.657
B_COST	-0.011	-0.057	-0.018	-0.072	-0.018	-0.073
B_FR	-0.005	-0.028	-0.008	-0.031	-0.008	-0.032
B_TIME	-0.013	-0.067	-0.017	-0.069	-0.017	-0.071
SIGMA_CAR			0.020			
SIGMA_TRAIN			0.039		0.061	
SIGMA_SM			3.224		3.180	

Taste heterogeneity

- 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

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

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

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

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

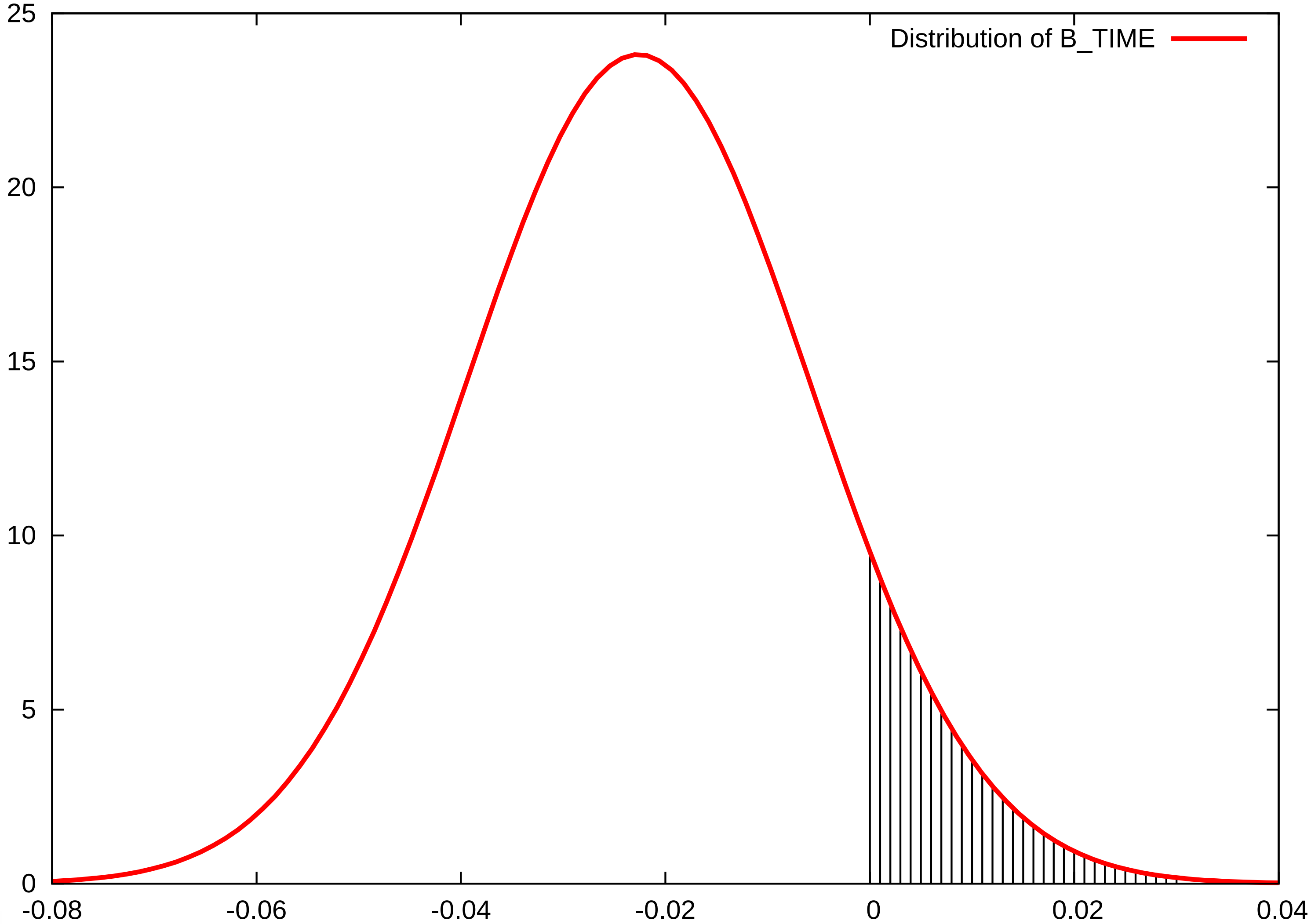
TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Random parameters

	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

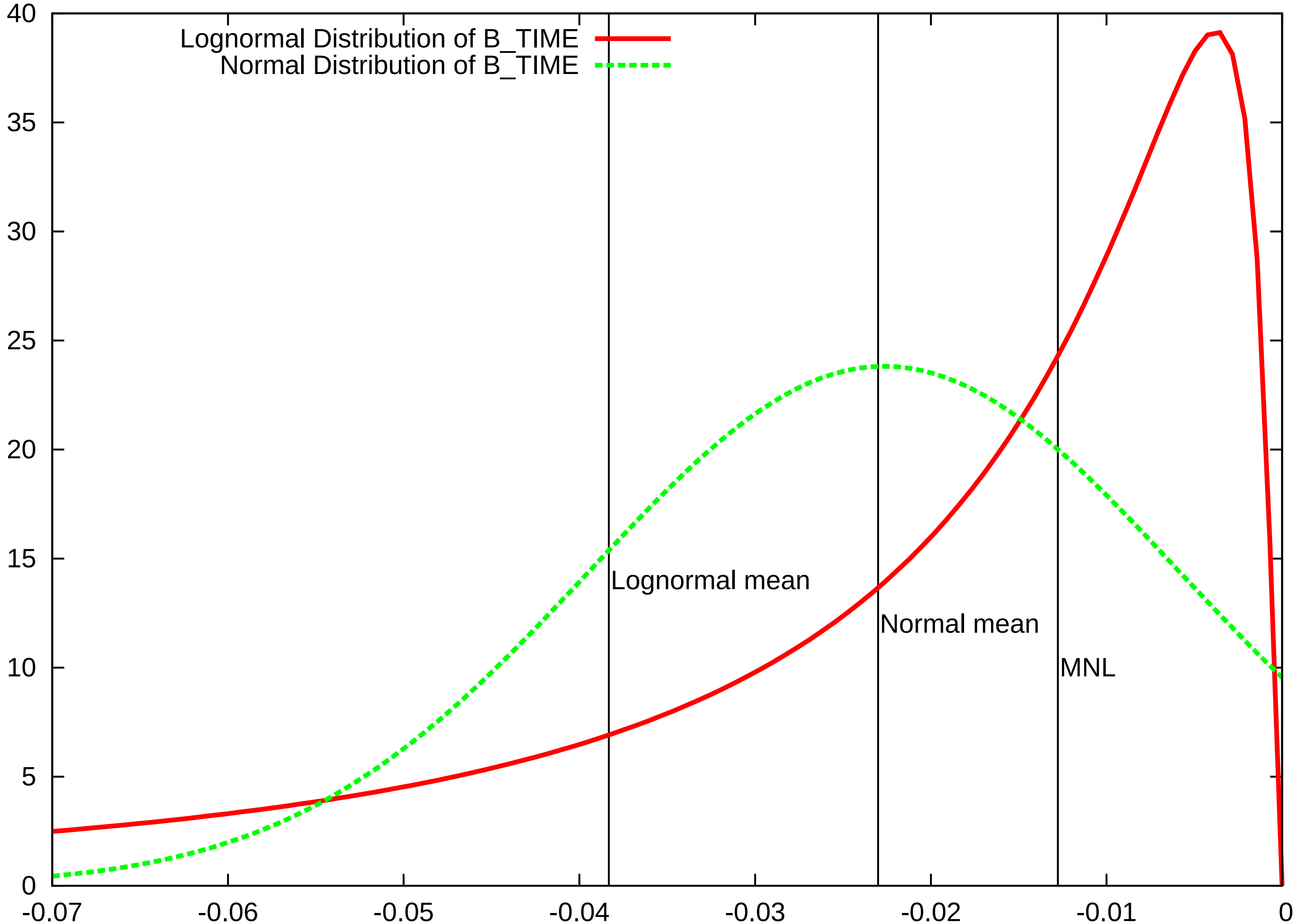
```
[Utilities]
11 SBB_SP TRAIN_AV_SP ASC_SBB_SP * one        +
                      B_COST     * TRAIN_COST +
                      B_FR       * TRAIN_FR
21 SM_SP SM_AV        ASC_SM_SP  * one        +
                      B_COST     * SM_COST    +
                      B_FR * SM_FR
31 Car_SP CAR_AV_SP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

	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

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

[LinearConstraints]
W1 + W2 = 1.0
```

Random parameters

	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

Latent classes

- Latent classes capture unobserved heterogeneity
- They can represent different:
  - Choice sets
  - Decision protocols
  - Tastes
  - Model structures
  - etc.

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Latent classes

$$P(i) = \sum_{s=1}^{S} \Lambda(i|s)Q(s)$$

- $\Lambda(i|s)$ is the class-specific choice model
  - *probability of choosing i given that the individual belongs to class s*
- $Q(s)$ is the class membership model
  - *probability of belonging to class s*

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Example: residential location

- Hypothesis
  - Lifestyle preferences exist (e.g., suburb vs. urban)
  - Lifestyle differences lead to differences in considerations, criterion, and preferences for residential location choices
- Infer "lifestyle" preferences from choice behavior using latent class choice model
  - Latent classes = lifestyle
  - Choice model = location decisions

Example: residential location

	(Alternative 1)	(Alternative 2)	(Alternative 3)	(Alternative 4)	(Alt. 5)
	Buy Single Family	Buy Multi-Family	Rent Single Family	Rent Multi-Family	Move out of the Metro Area
Type of Dwelling :	*single house*	*apartment*	*duplex / row house*	*condominium*	
Residence Size :	*< 1,000 sq. f t.*	*500-1,000 sq. f t.*	*1,500 - 2,000 sq. f t.*	*< 500 sq. f t.*	
Lot Size :	*< 5,000 sq. f t.*	*n/a*	*5,000 - 7,500 sq. f t.*	*n/a*	
Parking :	*street parking only*	*street parking only*	*driveway, no garage*	*reserved, uncovered*	
Price or Monthly Rents :	*< $75K*	*$50K - $100K*	*> $1,200*	*$300 - $600*	
Community Type :	*mixed use*	*mixed use*	*rural*	*urban*	
Housing Mix :	*mostly single f amily*	*mostly multi-f amily*	*mostly multi-f amily*	*mostly multi-f amily*	
Age of Development :	*10-15 years*	*0-5 years*	*10-15 years*	*0 - 5 years*	
Mix of Residential Ownership :	*mostly own*	*mostly own*	*mostly rent*	*mostly own*	
Shops/Services/Entertainment :	*community square*	*basic shops*	*community square*	*basic, specialty shops*	
Local Parks :	*none*	*yes*	*none*	*none*	
Bicycle Paths :	*none*	*yes*	*yes*	*yes*	
School Quality :	*very good*	*very good*	*f air*	*f air*	
Neighborhood Safety :	*average*	*average*	*average*	*average*	
Shopping Prices Relative to Avg :	*20% more*	*20% more*	*same*	*10% more*	
Walking Time to Shops :	*20-30 minutes*	*20-30 minutes*	*< 10 minutes*	*10 - 20 minutes*	
Bus Fare, Travel Time to Shops :	*$1.00, 15-20 minutes*	*$1.00, > 20 minutes*	*$0.50, 5 - 10 minutes*	*$0.50, < 5 minutes*	
Travel Time to Work by Auto :	*> 20 minutes*	*15-20 minutes*	*15 - 20 minutes*	*< 10 minutes*	
Travel Time to Work by Transit :	*> 45 minutes*	*30-45 minutes*	*30 - 45 minutes*	*15 - 30 minutes*	

Latent lifestyle segmentation

<u>Class 1</u>

Suburban, school, auto affluent, more established families

<u>Class 2</u>

Transit, school, less affluent, younger families

<u>Class 3</u>

High density, urban activity, older, non-family, professionals

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Summary

- Logit mixtures models
  - Computationally more complex than MEV
  - Allow for more flexibility than MEV
- Continuous mixtures: alternative specific variance, nesting structures, random parameters

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

- Discrete mixtures: well-defined latent classes of decision makers

$$P(i) = \sum_{s=1}^{S} \Lambda(i|s) Q(s).$$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Tips for applications

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