Mixtures of models

Michel Bierlaire

michel.bierlaire@epfl.ch

Transport and Mobility Laboratory

Mixtures

In statistics, a mixture density is a pdf which is a convex linear combination of other pdf's.
If $f(\varepsilon, \theta)$ is a pdf, and if $w(\theta)$ is a nonnegative 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 pdf. We say that g is a mixture of f.
If f is the pdf of a logit model, it is a mixture of logit
If f is the pdf of a MEV model, it is a mixture of MEV

Mixtures

Discrete mixtures are also possible. If $f(\varepsilon, \theta)$ is a pdf, and if w_i, $i = 1, \ldots, n$ are nonnegative weights such that

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

associated with parameter values θ_i, $i = 1, \ldots, n$ then

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

is also a pdf. We say that g is a discrete mixture of f.

Mixtures

Two important motivations:

- Define more complex error terms
  - heteroscedasticity
  - correlation across alternatives
- Capture taste heterogeneity

Capturing correlations

Logit

$$U_{in} = V_{in} + \varepsilon_{in}$$

where ε_{in} iid EV
Idea for the derivation of the nested logit model:

$$U_{in} = V_{in} + \varepsilon_m + \varepsilon_{in}$$

where ε_m is the error term specific to nest m.
Assumptions for the nested logit model:

- ε_m are independent across m
- $\varepsilon_m + \varepsilon'_m \sim$ EV$(0, \mu)$, and
- $\varepsilon'_m = \max_{i \in C_m}(V_i + \varepsilon_{im}) - \frac{1}{\mu_m} \ln \sum_{i \in C_m} e^{\mu_m V_i}$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Capturing correlations

- Assumptions are convenient for the derivation of the model
- They are not natural or intuitive

Consider a trinomial model, where alternatives 1 and 2 are correlated

$$\begin{aligned} U_{1n} &= V_{1n} + \varepsilon_m & +\varepsilon_{1n} \\ U_{2n} &= V_{2n} + \varepsilon_m & +\varepsilon_{2n} \\ U_{3n} &= V_{3n} & +\varepsilon_{3n} \end{aligned}$$

If ε_{in} are iid EV and ε_m is given, we have

$$P_n(1|\varepsilon_m, \mathcal{C}_n) = \frac{e^{V_{1n}+\varepsilon_m}}{e^{V_{1n}+\varepsilon_m} + e^{V_{2n}+\varepsilon_m} + e^{V_{3n}}}$$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Capturing correlations

But... ε_m is not given!
If we know its density function, we have

$$P_n(1|\mathcal{C}_n) = \int_{\varepsilon_m} P_n(1|\varepsilon_m, \mathcal{C}_n) f(\varepsilon_m) d\varepsilon_m$$

This is a mixture of logit models

In general, it is hopeless to obtain an analytical form for $P_n(1|\mathcal{C}_n)$
Simulation must be used.

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Simulation: reminders

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

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 $U(a,b)$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

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 $N(0,1)$

Simulation: transformations of standard normal

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

$$s = br + a$$

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

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

$$e^r$$

  is a draw from a lognormal $LN(a, b^2)$ with mean

$$e^{a+(b^2/2)}$$

  and variance

$$e^{2a+b^2}(e^{b^2} - 1)$$

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

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

Simulation: multivariate normal

- If r_1,...,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

Simulation: multivariate normal

Example:

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

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

Simulation for mixtures of logit

- In order to approximate

$$P_n(1|\mathcal{C}_n) = \int_{\varepsilon_m} P_n(1|\varepsilon_m, \mathcal{C}_n) f(\varepsilon_m) d\varepsilon_m$$

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

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

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Maximum simulated likelihood

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

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

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Maximum simulated likelihood

Warning:

- $\tilde{P}_n(j|\theta, \mathcal{C}_n)$ is an unbiased estimator of $P_n(j|\theta, \mathcal{C}_n)$

$$E[\tilde{P}_n(j|\theta, \mathcal{C}_n)] = P_n(j|\theta, \mathcal{C}_n)$$

- $\ln \tilde{P}_n(j|\theta, \mathcal{C}_n)$ is not an unbiased estimator of $\ln P_n(j|\theta, \mathcal{C}_n)$

$$\ln E[\tilde{P}_n(j|\theta, \mathcal{C}_n)] \neq E[\ln \tilde{P}_n(j|\theta, \mathcal{C}_n)]$$

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.

Modeling

$$P_n(1|\mathcal{C}_n) = \int_{\varepsilon_m} P_n(1|\varepsilon_m, \mathcal{C}_n) f(\varepsilon_m) d\varepsilon_m$$

Mixtures of logit can be used to model, depending on the role of ε_m in the kernel model.

- Heteroscedasticity
- Nesting structures
- Taste variations
- and many more...

Heteroscedasticity

- Error terms in logit are i.i.d. and, in particular, homoscedastic

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

- In order to introduce heteroscedasticity in the model, we use random ASCs

$$\text{ASC}_i \sim N(\overline{\text{ASC}}_i, \sigma_i^2)$$

so that

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

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

Heteroscedasticity

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.

Heteroscedastic model

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

Heteroscedastic model: ASCs random

	Logit		Hetero		Hetero norm.	
$\mathcal{L}$	-5315.39		-5241.01		-5242.10	
	Value	Scaled	Value	Scaled	Value	Scaled
ASC_CAR_SP	0.189	1.000	0.248	1.000	0.241	1.000
ASC_SM_SP	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_SP			0.020			
SIGMA_SBB_SP			-0.039		-0.061	
SIGMA_SM_SP			-3.224		-3.180	

Nesting structure

- Structure of nested logit can be mimicked with error components
- For each nest m, define a random term

$$\sigma_m \xi_m$$

where $\sigma_m \in \mathbb{R}$ and $\xi_m \sim N(0,1)$.

- σ_m represents the standard error of the r.v. $\xi_m \sim N(0,1)$
- If alternative i belongs to nest m, its utility writes

$$U_{in} = V_{in} + \sigma_m \xi_m + \varepsilon_{in}$$

where ε_{in} is, as usual, i.i.d EV.

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))	ξ_M	0
SM	0	1	0	0	$\ln$(cost(SM))	ξ_M	0
LF	0	0	1	0	$\ln$(cost(LF))	0	ξ_F
EF	0	0	0	1	$\ln$(cost(EF))	0	ξ_F
MF	0	0	0	0	$\ln$(cost(MF))	0	ξ_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		MLogit		MLogit $\sigma_F = 0$		MLogit $\sigma_M = 0$		MLogit $\sigma_F = \sigma_M$	
$\mathcal{L}$	-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	

Identification issue

- Not all parameters can be identified
- For instance, one ASC has to be constrained to zero
- Identification of MLogit is important and tricky
- Unidentified model: infinite number of estimates, sharing the same likelihood
- Need to impose restrictions to obtain a unique solution

Heteroscedastic model

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

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

The smallest σ must be set to 0

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Two nests

$$
\begin{array}{lllll}
U_1 = \beta x_1 & +\sigma_1\xi_1 & & +\varepsilon_1 \\
U_2 = \beta x_2 & +\sigma_1\xi_1 & & +\varepsilon_2 \\
U_3 = \beta x_3 & +\sigma_1\xi_1 & & +\varepsilon_3 \\
U_4 = \beta x_4 & & +\sigma_2\xi_2 & +\varepsilon_4 \\
U_5 = \beta x_5 & & +\sigma_2\xi_2 & +\varepsilon_5
\end{array}
$$

One σ must be constrained to 0

Three nests

$$
\begin{array}{lllllll}
U_1 & = & \beta x_1 & +\sigma_1\xi_1 & & & +\varepsilon_1 \\
U_2 & = & \beta x_2 & +\sigma_1\xi_1 & & & +\varepsilon_2 \\
U_3 & = & \beta x_3 & & +\sigma_2\xi_2 & & +\varepsilon_3 \\
U_4 & = & \beta x_4 & & & +\sigma_3\xi_3 & +\varepsilon_4 \\
U_5 & = & \beta x_5 & & & +\sigma_3\xi_3 & +\varepsilon_5
\end{array}
$$

All σs are estimable

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

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

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

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

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

Heteroscedastic: differences

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

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

- 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

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 Cov(ΔU)
2. Compute the Jacobian wrt σ_1^2, σ_2^2, σ_3^2, σ_4^2, γ/μ^2
3. Compute the rank

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

$S =$ Rank - 1 = 3

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Heteroscedastic: equality condition

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

Heteroscedastic: equality condition

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

Notations:

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

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

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:

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

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

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

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Heteroscedastic: equality condition

- 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: $\Omega = \Omega^{\text{norm}}$

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

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

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

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

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

- 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.

Random parameters

- 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

Example with Swissmetro

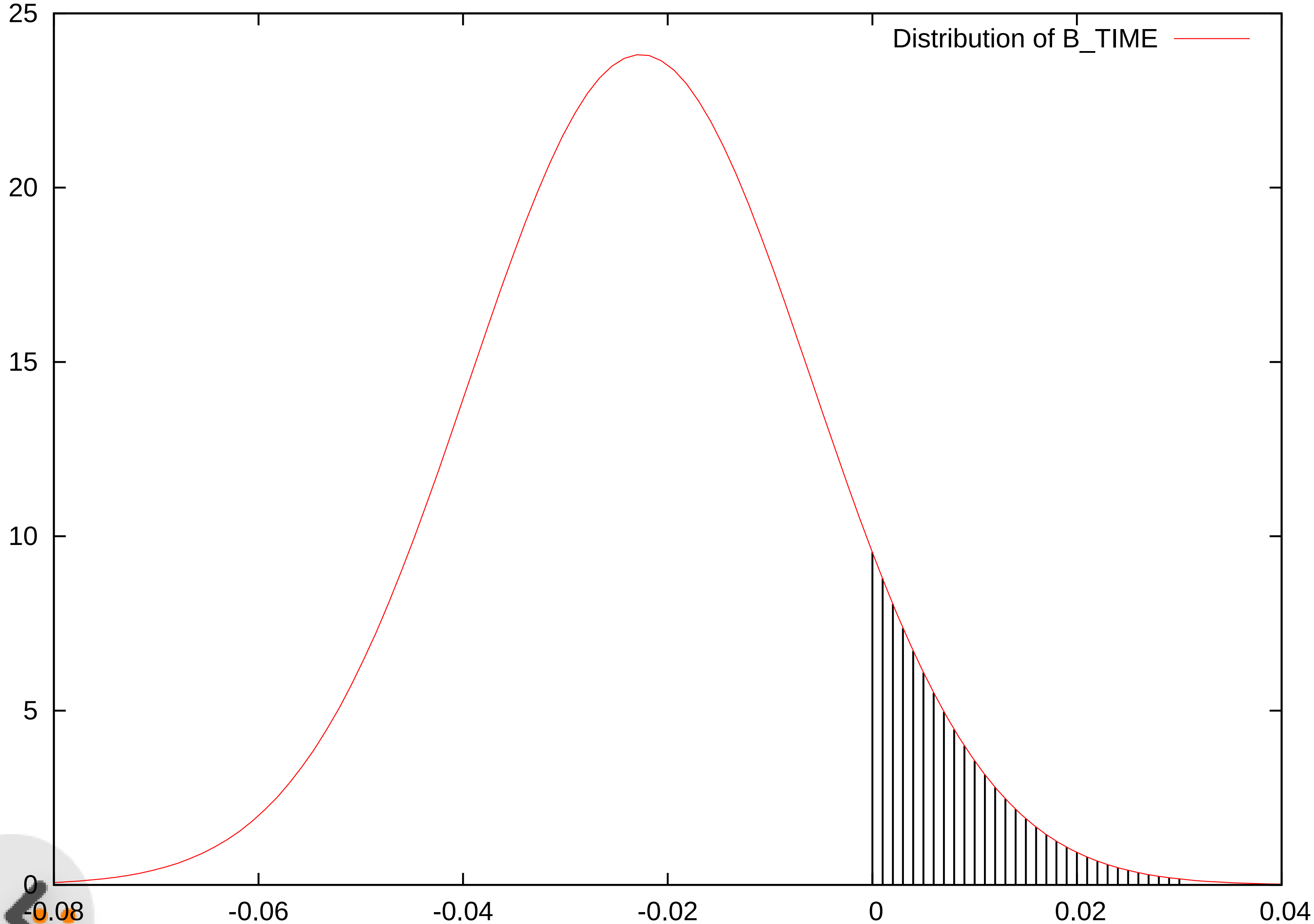
	ASC_CAR	ASC_SBB	ASC_SM	B_COST	B_FR	B_TIME
Car	1	0	0	cost	0	time
Train	0	0	0	cost	freq.	time
Swissmetro	0	0	1	cost	freq.	time

B_TIME randomly distributed across the population, normal distribution

Random parameters

	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, lognormal distribution

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Random parameters

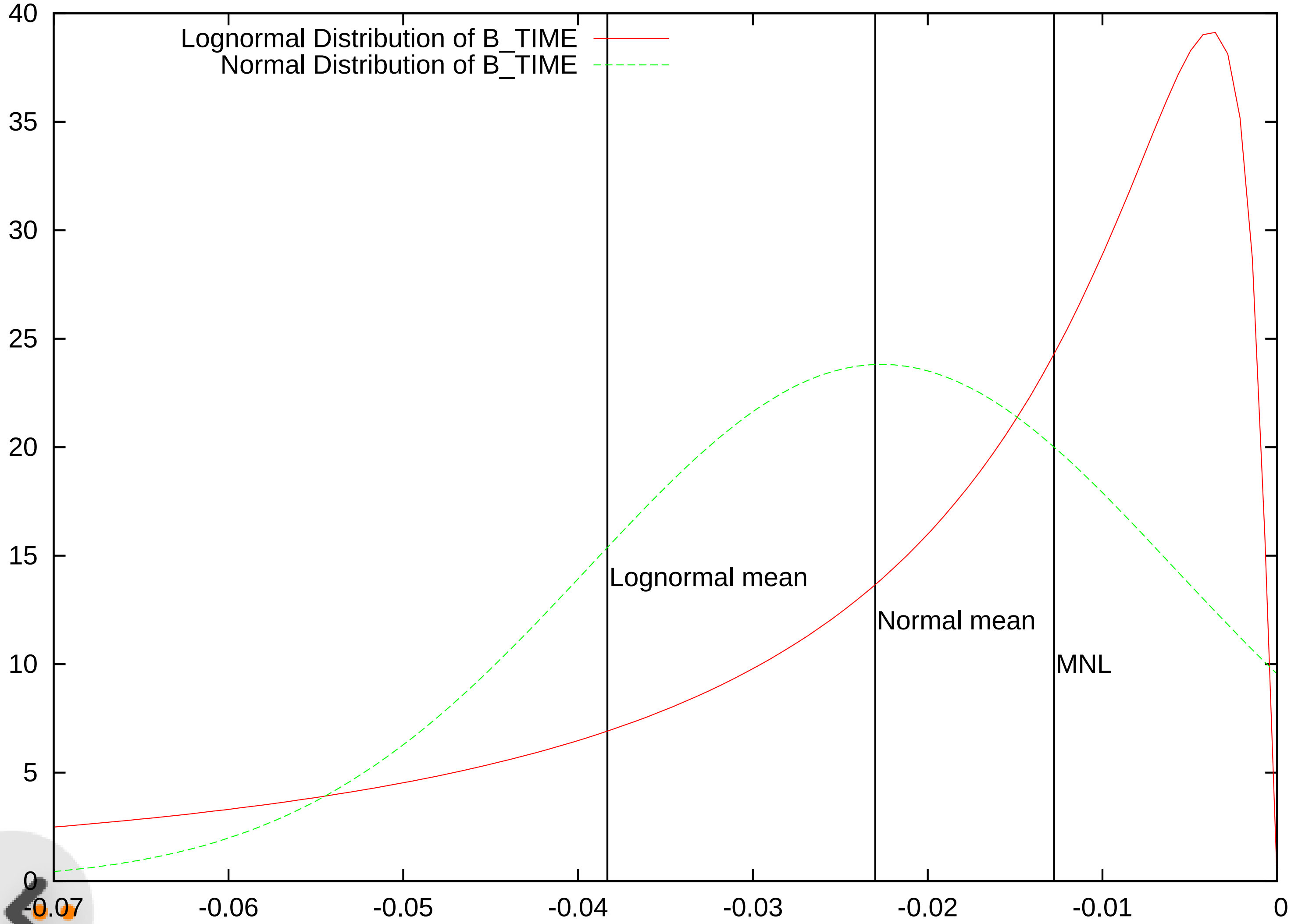
```
[Utilities]
11 SBB_SP TRAIN_AV_SP ASC_SBB_SP * one        +
                      B_COST     * TRAIN_COST +
                      B_FR       * TRAIN_FR
21 SM_SP SM_AV        ASC_SM_SP  * one        +
                      B_COST     * SM_COST    +
                      B_FR * SM_FR
31 Car_SP CAR_AV_SP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

Individual-level parameters

- Random parameters capture heterogeneity of the population
- Distribution of taste across the entire population
- For a given individual, can we have more information about where his taste lies in the distribution?
- Idea: the choice reveals something about the taste.
- Proposed by Revelt and Train (2000)

Individual-level parameters

- Random parameter: β
- Distribution of β in the population: $g(\beta|\theta)$
- θ: parameters of the distribution (mean, variance, etc.)
- Choice situation defined by x
- Consider subpopulation of persons choosing alt. y
- Distribution of β in the subpopulation: $h(\beta|y, x, \theta)$

Individual-level parameters

- Joint probability of β and y

$$P(\beta, y|x) = P(y|x,\beta)g(\beta|\theta)$$

or, also,

$$P(\beta, y|x) = h(\beta|y,x,\theta)P(y|x,\theta)$$

- We obtain

$$h(\beta|y,x,\theta) = \frac{P(y|x,\beta)g(\beta|\theta)}{P(y|x,\theta)}$$

Logit

e.g. Normal

Constant: Proba

Individual-level parameters

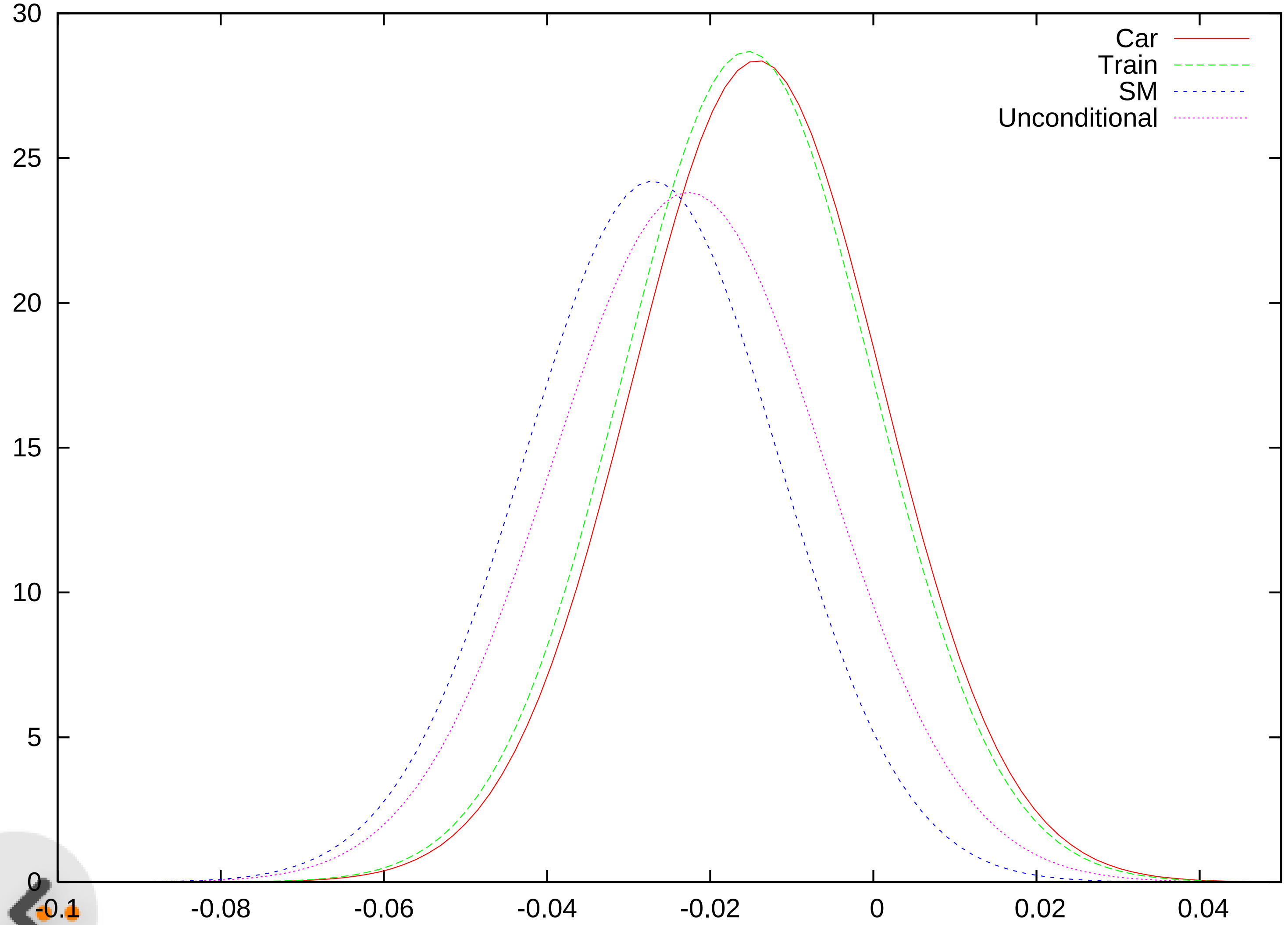
Example: Swissmetro, first observation in the file

	Car	Train	SM
Cost	65	48	52
Time	117	112	63
Frequency		120	20
Proba	21.7%	12.1%	66.3%

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Individual-level parameters



Individual-level parameters

Mean of β in the subpopulation

$$\begin{aligned}\bar{\beta} &= \int \beta \, h(\beta|y,x,\theta) d\beta \\ &= \frac{\int \beta P(y|x,\beta) g(\beta|\theta) d\beta}{P(y|x,\theta)} \\ &= \frac{\int \beta P(y|x,\beta) g(\beta|\theta) d\beta}{\int P(y|x,\beta) g(\beta|\theta) d\beta}\end{aligned}$$

Individual-level parameters

Simulation:

1. Generate draws β_r from $g(\beta|\theta)$
2. Compute weights
$$w_r = \frac{P(y|x,\beta_r)}{\sum_s P(y|x,\beta_s)}$$
3. The simulated subpopulation mean is
$$\tilde{\beta} = \sum_r w_r \beta_r.$$

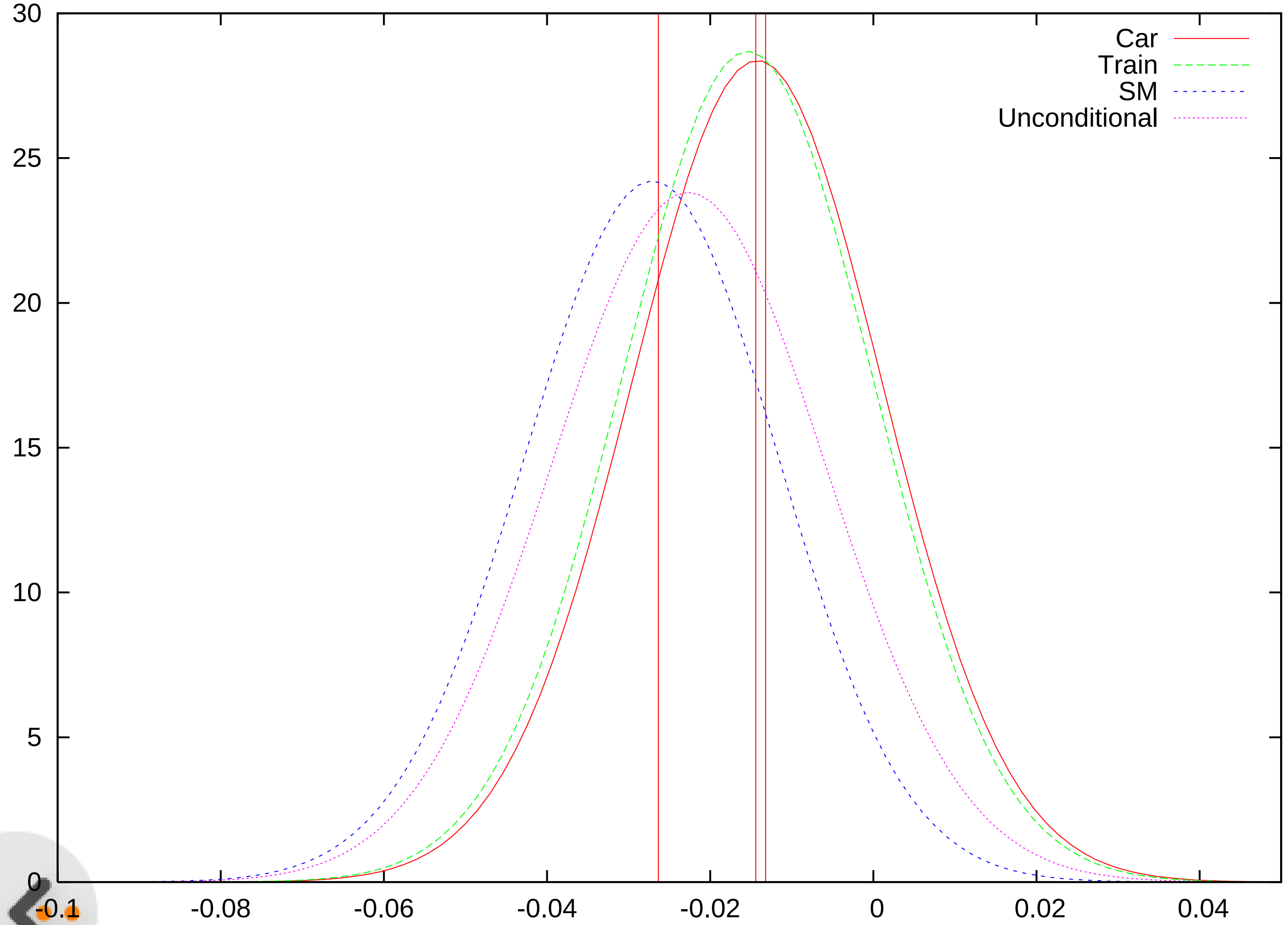
TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Individual-level parameters

$$\hat{\beta}_{\text{Car}} = -0.01320197$$
$$\hat{\beta}_{\text{Train}} = -0.014408413$$
$$\hat{\beta}_{\text{SM}} = -0.026355011$$

Individual-level parameters



Latent classes

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

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*

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*	

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

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

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