Importance sampling of alternatives for route choice models

Emma Frejinger and Michel Bierlaire

Transport and Mobility Laboratory, EPFL,transp-or.epfl.ch

Outline

- Introduction to route choice modeling
  - Modeling framework
  - Estimation
  - Issues
- Stochastic path enumeration approach
- Sampling of alternatives
- Preliminary numerical results

Route choice modeling

Given a transportation network composed of nodes, links, origin and destinations.
For a given transportation mode and origin-destination pair, which is the chosen route?

Route choice modeling

- Deterministic approach: Travelers use the shortest (with regard to any arbitrary generalized cost) route among all
  - Behaviorally unrealistic
- Random utility models (discrete choice models)

Framework

- Utility maximization
- An individual n associates a utility U_{jn} with each path j in his/her choice set $\mathcal{C}_n$ and chooses the alternative with the highest utility

Random Utility Models

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

V_{jn}: Deterministic part $V_{jn} = \beta^T X_{jn}$

β: vector of parameters to be estimated

X_{jn}: attributes

ε_{jn}: random term

Multinomial Logit model

$$P(i|\mathcal{C}_n) = \frac{e^{V_{in}}}{\sum_{j \in \mathcal{C}_n} e^{V_{jn}}}$$

Estimation

- Maximum likelihood estimation

$$\mathcal{L}^*(\hat{\beta}_1, ..., \hat{\beta}_K) = \max_{\beta \in \mathbb{R}} \mathcal{L}(\beta) = \sum_{n=1}^{N} \ln P_n(\beta)$$

- BIOGEME: estimation software

  Bierlaire's **O**ptimization Toolbox for **GE**V **M**odel **E**stimation

Problem characteristics

- Universal choice set very large
- Individual specific choice set unknown
- Correlated alternatives due to overlapping paths
- Data issues

Path Enumeration

- Many heuristics are proposed in the literature
  - Deterministic and stochastic
    Examples: link elimination (Azevedo et al., 1993), labeled paths (Ben-Akiva et al., 1984), simulation (Ramming, 2001) and doubly stochastic (Bovy and Fiorenzo-Catalano, 2006)
  - These approaches assume that generated choice sets include all alternatives considered by the travelers

Importance Sampling Approach

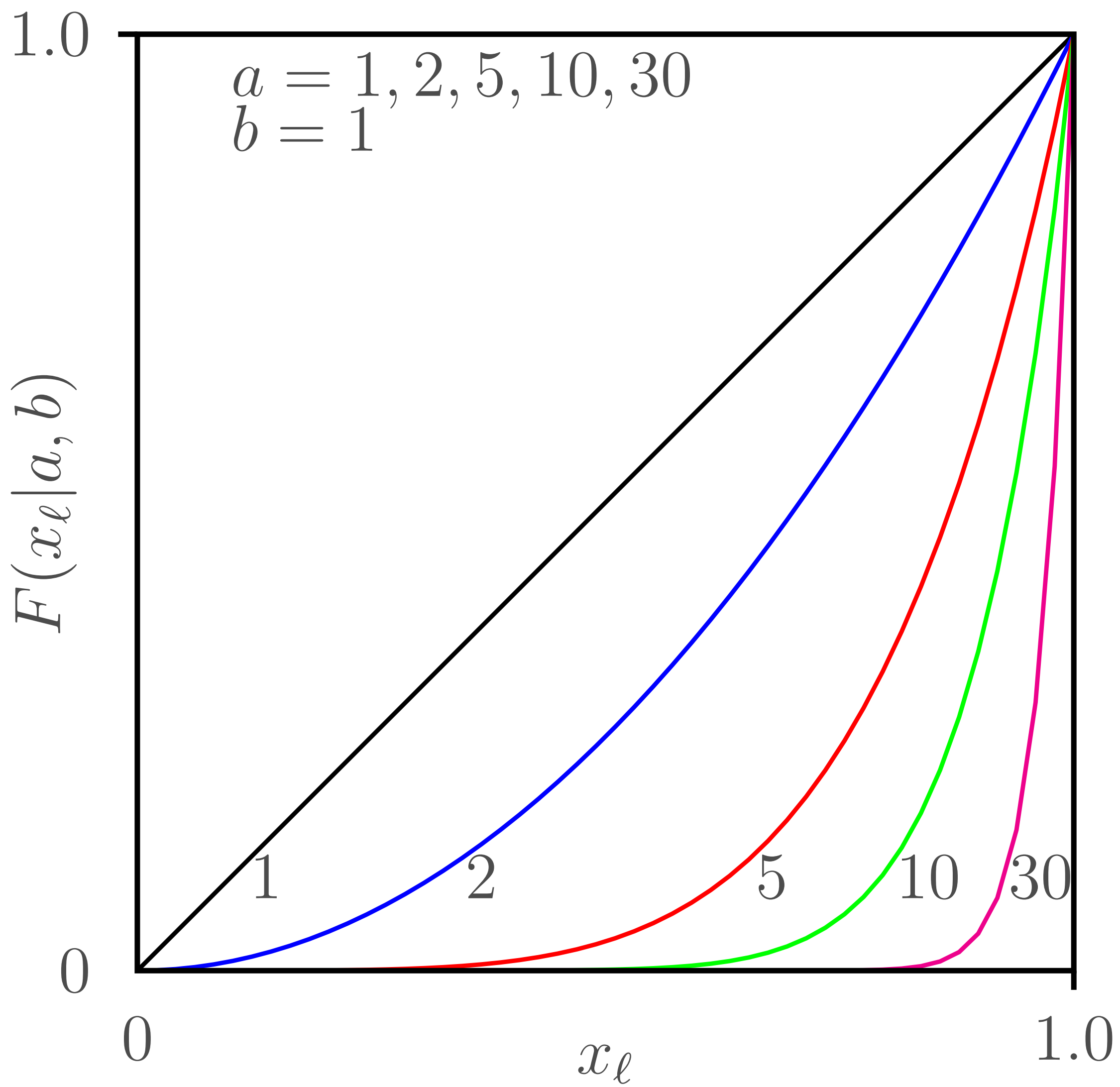
- All paths belong to the true choice set
- Objective: define choice set allowing for unbiased estimation and prediction results
- We view stochastic path enumeration algorithms as importance sampling of alternatives
- In order to obtain unbiased results, path utilities must be corrected
- We propose a stochastic path enumeration algorithm that allows the computation of sampling correction

Stochastic Path Enumeration

- We choose to include in the choice set a link ℓ or a sequence of links in a stochastic way based on its distance to the shortest path
- Paths can be generated using different algorithms
- Kumaraswamy distribution, cumulative distribution function $F(x_\ell|a,b) = 1 - (1 - {x_\ell}^a)^b$ for $x_\ell \in [0,1]$.

$$x_\ell = \frac{SP(o,d)}{SP(o,i) + C(\ell) + SP(j,d)}$$

Stochastic Path Enumeration



Stochastic Path Enumeration

- Biased random walk algorithm

$$q(j) = \prod_{\ell \in \Gamma_j} q(\ell | \mathcal{E}_v)$$

- Γ_j: set of all links in j
- v: source node of j
- E_v: set of all outgoing links from v
- $q(\ell|\mathcal{E}_v)$ is distributed Kumaraswamy
- Issue: the set of all paths $\mathcal{U}$ is unbounded but we assume $\sum_{j \in \mathcal{U}} q(j) \approx 1$ and treat it as bounded

Sampling of Alternatives

- Multinomial Logit model: Probability of i conditional on the choice set $\mathcal{C}_n$ defined by the analyst (e.g. Ben-Akiva and Lerman, 1985)

$$P(i|\mathcal{C}_n) = \frac{q(\mathcal{C}_n|i)P(i)}{\sum_{j\in\mathcal{C}_n} q(\mathcal{C}_n|j)P(j)} = \frac{e^{V_{in}+\ln q(\mathcal{C}_n|i)}}{\sum_{j\in\mathcal{C}_n} e^{V_{jn}+\ln q(\mathcal{C}_n|j)}}$$

$q(\mathcal{C}_n|j)$: probability of sampling $\mathcal{C}_n$ given that j is the chosen alternative

Sampling of Alternatives

- Sampling protocol: a set $\widetilde{\mathcal{C}}_n$ is generated by drawing R paths with replacement from the universal set of paths $\mathcal{U}$ and adding the chosen path to it
  Outcome of sampling: $(\widetilde{k}_1, \widetilde{k}_2, \ldots, \widetilde{k}_J)$ and $\sum_{j\in\mathcal{U}} \widetilde{k}_j = R$

$$P(\widetilde{k}_1, \widetilde{k}_2, \ldots, \widetilde{k}_J) = \frac{R!}{\prod_{j\in\mathcal{U}} \widetilde{k}_j!} \prod_{j\in\mathcal{U}} q(j)^{\widetilde{k}_j}$$

- Alternative j appears $k_j = \widetilde{k}_j + \delta_{cj}$ in $\widetilde{\mathcal{C}}_n$

Sampling of Alternatives

- Let $\mathcal{C}_n = \{j \in \mathcal{U} \mid k_j > 0\}$
- Following Ben-Akiva (1993)

$$q(\widetilde{\mathcal{C}}_n|i) = \frac{R!}{(k_i-1)!\prod_{\substack{j\in\mathcal{C}_n\\ j\neq i}} k_j!} q(i)^{k_i-1} \prod_{\substack{j\in\mathcal{C}_n\\ j\neq i}} q(j)^{k_j} = K_{\mathcal{C}_n}\frac{k_i}{q(i)}$$

$$K_{\mathcal{C}_n} = \frac{R!}{\prod_{j\in\mathcal{C}_n} k_j!}\prod_{j\in\mathcal{C}_n} q(j)^{k_j}$$

$$P(i|\widetilde{\mathcal{C}}_n) = \frac{e^{V_{in}+\ln\left(\frac{k_i}{q(i)}\right)}}{\sum_{j\in\mathcal{C}_n} e^{V_{jn}+\ln\left(\frac{k_j}{q(j)}\right)}}$$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Preliminary Numerical Results

- Estimation of models based on synthetic data generated with postulated models
  - Non-correlated paths
  - Correlated paths in a “grid-like” network
- True parameter values are compared to estimates

Preliminary Numerical Results

- True model: multinomial logit
  $U_j = \beta_{\mathsf{L}}\ \text{length}_j + \beta_{\mathsf{SB}}\ \text{nbspeedbumps}_j + \varepsilon_j$
  $\beta_{\mathsf{L}} = -0.6$ and $\beta_{\mathsf{SB}} = -0.3$
  ε_j is distributed Gumbel with location parameter 0 and scale 1
- 500 observations
- Biased random walk using 40 draws with $a = 2$ and $b = 1$
  Generated choice sets include at least 7, maximum 18 and on average 11.9 paths

TRANSP-OR

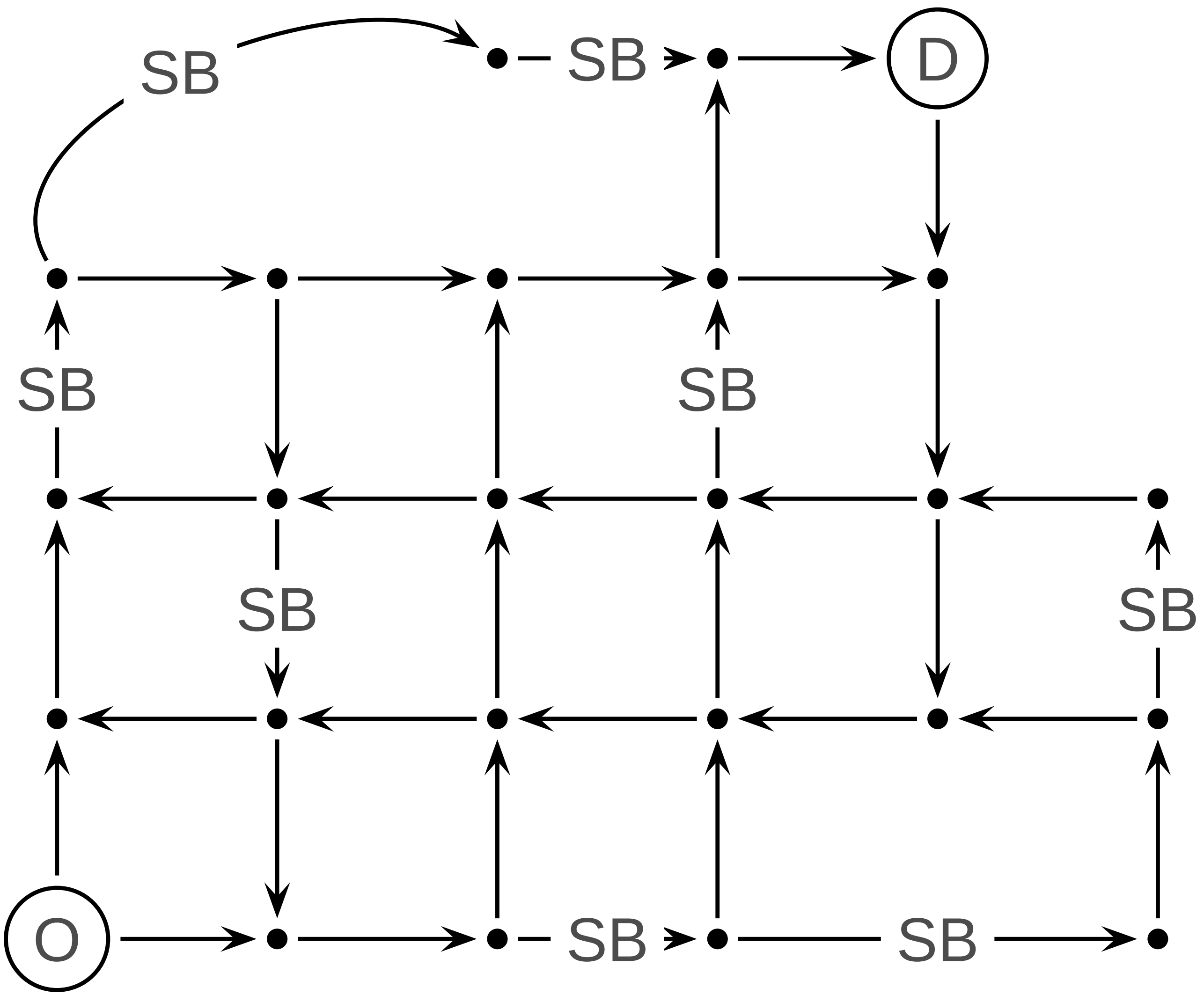
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Preliminary Numerical Results

Sampling correction	MNL without	MNL with
$\widehat{\beta}_{\text{L}}$	-0.203	-0.286
Scaled estimate	-0.600	-0.600
Robust std.	0.0193	0.019
Robust t-test	-10.53	-15.01
$\widehat{\beta}_{\text{SB}}$	-0.0194	-0.143
Scaled estimate	-0.0573	-0.300
Robust std.	0.0662	0.0661
Robust t-test	-0.29	-2.17

Null log-likelihood	-1069.453	-1633.501
Final log-likelihood	-788.42	-759.848
Adjusted $\bar{\rho}^2$	0.261	0.288
BIOGEME has been used for all model estimations.		

Preliminary Numerical Results



Preliminary Numerical Results

- True model: probit (Burrell, 1968)
  $U_\ell = \beta_{\text{L}}\ \text{length}_\ell + \beta_{\text{SB}}\ \text{nbspeedbumps}_\ell + \sigma\sqrt{L_\ell}\nu_\ell$
  $\beta_{\text{L}} = -0.6$ and $\beta_{\text{SB}} = -0.4$
  ν_ℓ is distributed standard Normal
  Link utility variance assumed proportional to length with parameter $\sigma = 0.8$
- Path utilities are link additive
- 382 observations are generated after 500 realizations of the link utilities

Preliminary Numerical Results

- Biased random walk using 30 draws with $a = 2$ and $b = 1$

  Generated choice sets include at least 7, maximum 19 and on average 13.5 paths

Preliminary Numerical Results

Sampling correction	MNL without	MNL with	PSL without	PSL with
$\widehat{\beta}_{\mathrm{L}}$	-0.627	-0.978	-0.619	-0.969
Scaled estimate	-0.600	-0.600	-0.600	-0.600
Robust std.	0.0397	0.032	0.0407	0.0358
Robust t-test	-15.79	-30.57	-15.22	-27.04
$\widehat{\beta}_{\mathrm{SB}}$	-0.0822	-0.0801	-0.347	-0.461
Scaled estimate	-0.0787	-0.0491	-0.336	-0.285
Robust std.	0.052	0.0559	0.182	0.158
Robust t-test	-1.58	-1.43	-1.90	-2.92
$\widehat{\beta}_{\mathrm{PS}}$			1.17	1.74
Scaled estimate			1.13	1.08
Robust std.			0.788	0.705
Robust t-test			1.49	2.47

Preliminary Numerical Results

Sampling correction	MNL without	MNL with	PSL without	PSL with
Null log-likelihood	-988.63	-2769.959	-988.63	-2769.959
Final log-likelihood	-676.111	-653.396	-674.481	-649.268
Adjusted $\bar{\rho}^2$	0.314	0.337	0.315	0.340
BIOGEME has been used for all model estimations.				

Conclusions and Future Work

- Ongoing research
- Modeling path enumeration as importance sampling of alternatives is promising however some work remain
  - Implications of $\sum_{j \in \mathcal{U}} q(j) \approx 1$
  - Empirical results on real data
  - Correction in prediction