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 $C_n$ and chooses the alternative with the highest utility
Random Utility Models

\[ U_{jn} = V_{jn} + \varepsilon_{jn} \]

\( V_{jn} \): Deterministic part \( V_{jn} = \beta^T X_{jn} \)
\( \beta \): vector of parameters to be estimated
\( X_{jn} \): attributes
\( \varepsilon_{jn} \): random term

Multinomial Logit model

\[ P(i|C_n) = \frac{e^{V_{in}}}{\sum_{j \in C_n} e^{V_{jn}}} \]
Estimation

- Maximum likelihood estimation

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

- BIOGEME: estimation software

Bierlaire’s Optimization Toolbox for GEV Model Estimation
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 $\ell$ 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

- Biased random walk algorithm

\[ q(j) = \prod_{\ell \in \Gamma_j} q(\ell | E_v) \]

- \( \Gamma_j \): set of all links in \( j \)
- \( v \): source node of \( j \)
- \( E_v \): set of all outgoing links from \( v \)
- \( q(\ell | E_v) \) 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 $C_n$ defined by the analyst (e.g. Ben-Akiva and Lerman, 1985)

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

$q(C_n|j)$: probability of sampling $C_n$ given that $j$ is the chosen alternative
Sampling of Alternatives

- Sampling protocol: a set $\tilde{C}_n$ 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: $(\tilde{k}_1, \tilde{k}_2, \ldots, \tilde{k}_J)$ and $\sum_{j \in \mathcal{U}} \tilde{k}_j = R$

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

- Alternative $j$ appears $k_j = \tilde{k}_j + \delta_{cj}$ in $\tilde{C}_n$.
Sampling of Alternatives

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

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

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

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

\[ U_j = \beta_L \text{ length}_j + \beta_{SB} \text{ nbspeedbumps}_j + \varepsilon_j \]

\[ \beta_L = -0.6 \text{ and } \beta_{SB} = -0.3 \]

\( \varepsilon_j \) 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
## Preliminary Numerical Results

<table>
<thead>
<tr>
<th>Sampling correction</th>
<th>MNL without</th>
<th>MNL with</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\beta}_L$</td>
<td>-0.203</td>
<td>-0.286</td>
</tr>
<tr>
<td>Scaled estimate</td>
<td>-0.600</td>
<td>-0.600</td>
</tr>
<tr>
<td>Robust std.</td>
<td>0.0193</td>
<td>0.019</td>
</tr>
<tr>
<td>Robust t-test</td>
<td>-10.53</td>
<td>-15.01</td>
</tr>
<tr>
<td>$\hat{\beta}_{SB}$</td>
<td>-0.0194</td>
<td>-0.143</td>
</tr>
<tr>
<td>Scaled estimate</td>
<td>-0.0573</td>
<td>-0.300</td>
</tr>
<tr>
<td>Robust std.</td>
<td>0.0662</td>
<td>0.0661</td>
</tr>
<tr>
<td>Robust t-test</td>
<td>-0.29</td>
<td>-2.17</td>
</tr>
<tr>
<td>Null log-likelihood</td>
<td>-1069.453</td>
<td>-1633.501</td>
</tr>
<tr>
<td>Final log-likelihood</td>
<td>-788.42</td>
<td>-759.848</td>
</tr>
<tr>
<td>Adjusted $\bar{\rho}^2$</td>
<td>0.261</td>
<td>0.288</td>
</tr>
</tbody>
</table>

BIOGEME has been used for all model estimations.
Preliminary Numerical Results
Preliminary Numerical Results

- True model: probit (Burrell, 1968)

\[ U_\ell = \beta_L \text{length}_\ell + \beta_{SB} \text{nbspeedbumps}_\ell + \sigma \sqrt{L_\ell} \nu_\ell \]

\( \beta_L = -0.6 \) and \( \beta_{SB} = -0.4 \)
\( \nu_\ell \) 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

<table>
<thead>
<tr>
<th>Sampling correction</th>
<th>MNL without</th>
<th>MNL with</th>
<th>PSL without</th>
<th>PSL with</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\beta}_L$</td>
<td>-0.627</td>
<td>-0.978</td>
<td>-0.619</td>
<td>-0.969</td>
</tr>
<tr>
<td>Scaled estimate</td>
<td>-0.600</td>
<td>-0.600</td>
<td>-0.600</td>
<td>-0.600</td>
</tr>
<tr>
<td>Robust std.</td>
<td>0.0397</td>
<td>0.032</td>
<td>0.0407</td>
<td>0.0358</td>
</tr>
<tr>
<td>Robust t-test</td>
<td>-15.79</td>
<td>-30.57</td>
<td>-15.22</td>
<td>-27.04</td>
</tr>
<tr>
<td>$\hat{\beta}_{SB}$</td>
<td>-0.0822</td>
<td>-0.0801</td>
<td>-0.347</td>
<td>-0.461</td>
</tr>
<tr>
<td>Scaled estimate</td>
<td>-0.0787</td>
<td>-0.0491</td>
<td>-0.336</td>
<td>-0.285</td>
</tr>
<tr>
<td>Robust std.</td>
<td>0.052</td>
<td>0.0559</td>
<td>0.182</td>
<td>0.158</td>
</tr>
<tr>
<td>Robust t-test</td>
<td>-1.58</td>
<td>-1.43</td>
<td>-1.90</td>
<td>-2.92</td>
</tr>
<tr>
<td>$\hat{\beta}_{PS}$</td>
<td></td>
<td></td>
<td>1.17</td>
<td>1.74</td>
</tr>
<tr>
<td>Scaled estimate</td>
<td></td>
<td></td>
<td>1.13</td>
<td>1.08</td>
</tr>
<tr>
<td>Robust std.</td>
<td></td>
<td></td>
<td>0.788</td>
<td>0.705</td>
</tr>
<tr>
<td>Robust t-test</td>
<td></td>
<td></td>
<td>1.49</td>
<td>2.47</td>
</tr>
</tbody>
</table>
# Preliminary Numerical Results

<table>
<thead>
<tr>
<th>Sampling correction</th>
<th>MNL without</th>
<th>MNL with</th>
<th>PSL without</th>
<th>PSL with</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null log-likelihood</td>
<td>-988.63</td>
<td>-2769.959</td>
<td>-988.63</td>
<td>-2769.959</td>
</tr>
<tr>
<td>Final log-likelihood</td>
<td>-676.111</td>
<td>-653.396</td>
<td>-674.481</td>
<td>-649.268</td>
</tr>
<tr>
<td>Adjusted $\bar{\rho}^2$</td>
<td>0.314</td>
<td>0.337</td>
<td>0.315</td>
<td>0.340</td>
</tr>
</tbody>
</table>

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 $\sum_{j \in U} q(j) \approx 1$
  - Empirical results on real data
  - Correction in prediction