Choice set generation for route choice models using a sampling approach

Emma Frejinger and Michel Bierlaire

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

- Introduction
- Stochastic path enumeration approach
- Sampling of alternatives
- Numerical results
- Conclusions
Introduction

• Route choice problem
  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?

• Discrete choice modeling framework

• Issue
  Universal choice set very large, individual specific choice set unknown
Introduction

Set of all paths $\mathcal{U}$ from $o$ to $d$

Path generation

Deterministic

$\mathcal{M} \subseteq \mathcal{U}$

Stochastic

$\mathcal{M}_n \subseteq \mathcal{U}$

Route choice Choice set formation

Deterministic

Probabilistic

Route choice model

$P(i|C_n) P(i) = \sum_{C_n \in G_n} P(i|C_n) P(C_n)$

Importance sampling of alternatives for route choice models – p.4/23
Introduction

- Choice sets need to be defined prior to the route choice modeling
- Path enumeration algorithms are used for this purpose, many heuristics have been proposed, for example:
  - Deterministic approaches: link elimination (Azevedo et al., 1993), labeled paths (Ben-Akiva et al., 1984)
  - Stochastic approaches: simulation (Ramming, 2001) and doubly stochastic (Bovy and Fiorenzo-Catalano, 2006)
Introduction

- Underlying assumption: the actual choice set is generated
- Empirical results suggest that this is not always true
- Our approach:
  - True choice set = universal set
  - Too large
  - Sampling of alternatives
Sampling of Alternatives

- Multinomial logit model (e.g. Ben-Akiva and Lerman, 1985):

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

\(C_n\): set of sampled alternatives
\(q(C_n|j)\): probability of sampling \(C_n\) given that \(j\) is the chosen alternative
Importance Sampling of Alternatives

- Attractive paths have higher probability of being sampled than unattractive paths
- Path utilities must be corrected in order to obtain unbiased estimation results
MNL Route Choice Models

- Path Size Logit (Ben-Akiva and Ramming, 1998 and Ben-Akiva and Bierlaire, 1999) and C-Logit (Cascetta et al. 1996)
- Additional attribute in the deterministic utilities capturing correlation among alternatives
- These attributes should reflect the true correlation structure
- Hypothesis: attributes should be computed based on all paths (or as many as possible)
Stochastic Path Enumeration

- Flexible approach that can be combined with various algorithms, here a biased random walk approach
- The probability of a link $\ell$ with source node $v$ and sink node $w$ is modeled in a stochastic way based on its distance to the shortest path
- Kumaraswamy distribution, cumulative distribution function $F(x_\ell|a, b) = 1 - (1 - x_\ell^a)^b$ for $x_\ell \in [0, 1]$.

$$x_\ell = \frac{SP(v, d)}{C(\ell) + SP(w, d)}$$
Stochastic Path Enumeration

\[ F(x_\ell | a, b) \]

\[ a = 1, 2, 5, 10, 30 \]
\[ b = 1 \]
Stochastic Path Enumeration

- Probability for path $j$ to be sampled

$$q(j) = \prod_{\ell=(v,w) \in \Gamma_j} q((v, w) | E_v)$$

- $\Gamma_j$: ordered set of all links in $j$
- $v$: source node of $j$
- $E_v$: set of all outgoing links from $v$
- In theory, the set of all paths $\mathcal{U}$ may be unbounded. We treat it as bounded with size $J$. 
Sampling of Alternatives

- Following Ben-Akiva (1993)

- Sampling protocol
  1. A set $\tilde{C}_n$ is generated by drawing $R$ paths with replacement from the universal set of paths $U$
  2. Add chosen path to $\tilde{C}_n$

- Outcome of sampling: $(\tilde{k}_1, \tilde{k}_2, \ldots, \tilde{k}_J)$ and $\sum_{j=1}^{J} \tilde{k}_j = R$

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

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

- Let $C_n = \{ j \in \mathcal{U} \mid k_j > 0 \}$

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

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

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

- Estimation of models based on synthetic data generated with a postulated model
- Evaluation of
  - Sampling correction
  - Path Size attribute
  - Biased random walk algorithm parameters
Numerical Results
Numerical Results

- True model: Path Size Logit
  \[ V_j = \beta_{PS} \text{PS}_j^U + \beta_L \text{Length}_j + \beta_{SB} \text{SpeedBumps}_j \]
  \[ \beta_{PS} = 1, \beta_L = -0.3, \beta_{SB} = -0.1 \]
  \[ \text{PS}_i^U = \sum_{\ell \in \Gamma_i} \frac{L_{\ell}}{L_i} \sum_{j \in U} \delta_{\ell j} \]
  \[ P(i|U) = \frac{e^{V_i}}{\sum_{j \in U} e^{V_j}} \]

- 3000 observations
Numerical Results

- Four model specifications
  - Model $M_{PS(C)}^{NoCorr}$:
    \[ V_{in} = \beta_{PS} PS_{in}^C + \beta_L \text{Length}_i + \beta_{SB} \text{SpeedBumps}_i \]
  - Model $M_{PS(C)}^{Corr}$:
    \[ V_{in} = \beta_{PS} PS_{in}^C + \beta_L \text{Length}_i + \beta_{SB} \text{SpeedBumps}_i + \ln\left(\frac{k_i}{q(i)}\right) \]
  - Model $M_{PS(U)}^{NoCorr}$:
    \[ V_i = \beta_{PS} PS_{i}^U + \beta_L \text{Length}_i + \beta_{SB} \text{SpeedBumps}_i \]
  - Model $M_{PS(U)}^{Corr}$:
    \[ V_j = \beta_{PS} PS_{i}^U + \beta_L \text{Length}_i + \beta_{SB} \text{SpeedBumps}_i + \ln\left(\frac{k_i}{q(i)}\right) \]

\[ PS_{in}^C = \sum_{\ell \in \Gamma_i} \frac{l_{\ell}}{L_i} \frac{1}{\sum_{j \in C_n} \delta_{\ell j}} \]
## Numerical Results

<table>
<thead>
<tr>
<th></th>
<th>True PSL</th>
<th>$M_{PS(C)}^{NoCorr}$ PSL</th>
<th>$M_{PS(C)}^{Corr}$ PSL</th>
<th>$M_{PS(U)}^{NoCorr}$ PSL</th>
<th>$M_{PS(U)}^{Corr}$ PSL</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\beta}_{PS}$ Standard error t-test w.r.t. 1</td>
<td>1</td>
<td>0.363</td>
<td>0.443</td>
<td>-0.203</td>
<td>1.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0729</td>
<td>0.086</td>
<td>0.0487</td>
<td>0.0465</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-8.74</td>
<td>-6.48</td>
<td>-24.70</td>
<td>0.65</td>
</tr>
<tr>
<td>$\hat{\beta}_{L}$ Standard error t-test w.r.t. -0.3</td>
<td>-0.3</td>
<td>-0.0529</td>
<td>-0.326</td>
<td>-0.0453</td>
<td>-0.291</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.00864</td>
<td>0.0085</td>
<td>0.00828</td>
<td>0.00788</td>
</tr>
<tr>
<td></td>
<td></td>
<td>28.60</td>
<td>-3.06</td>
<td>30.76</td>
<td>1.14</td>
</tr>
<tr>
<td>$\hat{\beta}_{SB}$ Standard error t-test w.r.t. -0.1</td>
<td>-0.1</td>
<td>-0.345</td>
<td>-0.134</td>
<td>-0.404</td>
<td>-0.0773</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0315</td>
<td>0.0259</td>
<td>0.0298</td>
<td>0.0258</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-7.78</td>
<td>-1.31</td>
<td>-10.20</td>
<td>0.88</td>
</tr>
</tbody>
</table>
## Numerical Results

<table>
<thead>
<tr>
<th>True PSL</th>
<th>$M_{PS(c)}^{NoCorr}$ PSL</th>
<th>$M_{PS(c)}^{Corr}$ PSL</th>
<th>$M_{PS(\ell)}^{NoCorr}$ PSL</th>
<th>$M_{PS(\ell)}^{Corr}$ PSL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Final Log-likelihood</td>
<td>-6596.22</td>
<td>-6047.14</td>
<td>-6598.46</td>
<td>-5840.80</td>
</tr>
<tr>
<td>Adj. rho square</td>
<td>0.02</td>
<td>0.10</td>
<td>0.02</td>
<td>0.13</td>
</tr>
</tbody>
</table>

Null Log-likelihood: -6719.733, 3000 observations

Algorithm parameters: 10 draws, $a = 5$, $b = 1$, $C(\ell) = L_\ell$

Average size of sampled choice sets: 9.43

BIOGEME (biogeme.epfl.ch) has been used for all model estimations
Numerical Results

Length Coefficient

Speed Bump Coefficient

Path Size Coefficient
Conclusions

- New point of view on choice set generation and route choice modeling
- Path generation is considered an importance sampling approach
- We present a path generation algorithm and derive the corresponding sampling correction
- Path Size should be computed based on true correlation structure
- Numerical results are very promising