A sampling of alternatives approach for route choice models

Emma Frejinger

Centre for Transport Studies, KTH, Stockholm

Joint work with Michel Bierlaire and Moshe Ben-Akiva
Short about CTS at KTH

- A collaboration between
  - KTH (Transport and Location Analysis, Transport and Logistics)
  - VTI Transport Economics
  - SIKA, WSP, JIBS
- 10 year financing
  - VINNOVA, Rail adm., Road adm., Rikstrafiken
  - The partners
- Model development, appraisal methodology, applied appraisal and analysis
Outline

- Introduction to route choice modeling and choice set generation
- Sampling of alternatives and derivation of sampling correction
- Stochastic path generation
- Expanded path size attribute
- Numerical results
- Conclusions and future work
Introduction

Given

- a uni-modal transportation network
- an origin-destination pair
- link and path attributes
- socio-economic characteristics of a traveler

identify the route a (given) traveler would select

- Important problem in e.g. intelligent transport systems, GPS navigation and transportation planning
Introduction

- Most simple route choice model: travelers use the shortest path with respect to any generalized cost function
  - Behaviorally unrealistic
- Random utility models
- Here we consider estimation of random utility models based on disaggregate revealed preferences data
Introduction

Network

Trips
Introduction

- Network
- Trips
- Path
  - Observations
Introduction
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

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

- Underlying assumption in existing approaches: the actual choice set is generated
- Empirical results suggest that this is not always true
- In order to avoid bias in the econometric model we propose
  - True choice set = universal set \( \mathcal{U} \)
  - Too large
  - Sampling of alternatives
Sampling of Alternatives

- Multinomial Logit model can be consistently estimated on a subset of alternatives (McFadden, 1978)

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

\(C_n\): 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

- Here we use the Path Size Logit with an Expanded Path Size attribute
Sampling Correction

• Based on Ben-Akiva (1993)

• Sampling protocol

1. A set $\tilde{C}_n$ is generated by drawing $R_n$ paths with replacement from the universal set of paths $\mathcal{U}$

2. Add chosen path to $\tilde{C}_n$

• Outcome of sampling: $(\tilde{k}_{1n}, \tilde{k}_{2n}, \ldots, \tilde{k}_{Jn})$ and

\[
\sum_{j \in \mathcal{U}} \tilde{k}_{jn} = R_n
\]

\[
P(\tilde{k}_{1n}, \tilde{k}_{2n}, \ldots, \tilde{k}_{Jn}) = \frac{R_n!}{\prod_{j \in \mathcal{U}} \tilde{k}_{jn}!} \prod_{j \in \mathcal{U}} q(j)^{\tilde{k}_{jn}}
\]
Sampling Correction

- Alternative $j$ appears $k_{jn} = \tilde{k}_{jn} + \delta_{jc}$ in $\tilde{C}_n$
- Let $C_n = \{j \in U \mid k_j > 0\}$

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

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

\[ P(i | C_n) = \frac{e^{\mu V_{in} + \ln \left( \frac{k_i}{q(i)} \right)}}{\sum_{j \in C_n} e^{\mu V_{jn} + \ln \left( \frac{k_j}{q(j)} \right)}} \]
Sampling Correction

- General methodology valid for any definition of $\mathcal{U}$ if
  - it exists an algorithm generating any path in $\mathcal{U}$ with non zero probability
  - path sampling probabilities can be computed
Stochastic Path Generation

- Biased (towards shortest path) random walk algorithm
- A measure of distance $x_\ell \in [0, 1]$ to the shortest path for link $\ell = (v, w)$

$$x_\ell = \frac{SP(v, s_d)}{C(\ell) + SP(w, s_d)}$$

- Parametrized function mapping $[0, 1]$ into itself (inspired from the Kumaraswamy distribution)

$$\omega(\ell|b_1, b_2) = 1 - \left(1 - x_\ell^{b_1}\right)^{b_2}$$
Stochastic Path Generation

\[ \omega(\ell | b_1, b_2) \]

\[ b_1 = 1, 2, 5, 10, 30 \]

\[ b_2 = 1 \]
Stochastic Path Generation

Initialize \( v = s_o, \Gamma = \emptyset \)

Loop While \( v \neq s_d \) perform the following

Weights For each link \( \ell = (v, w) \in E_v \), where \( E_v \) is the set of outgoing links from \( v \), we compute the weights \( \omega(\ell|b_1, b_2) \)

Probability For each link \( \ell = (v, w) \in E_v \), we compute

\[
q(\ell|E_v, b_1, b_2) = \frac{\omega(\ell|b_1, b_2)}{\sum_{m \in E_v} \omega(m|b_1, b_2)}
\]

Draw Randomly select a link \( (v, w^*) \) in \( E_v \) based on the above probability distribution

Update path \( \Gamma = \Gamma \cup (v, w^*) \)

Next node \( v = w^* \).
Stochastic Path Generation

- Probability $q(j)$ of generating path $j$ is

$$q(j) = \prod_{\ell \in \Gamma_j} q(\ell | \mathcal{E}_v, b_1, b_2)$$
Expanded Path Size

- Path size attribute should reflect the “true” correlation structure

- Original formulation

\[
PS^C_{in} = \sum_{a \in \Gamma_i} \frac{L_a}{L_i} \frac{1}{M_{an}}, \quad M_{an} = \sum_{j \in C_n} \delta_{aj}
\]

- Expanded formulation

\[
EPS_{in} = \sum_{a \in \Gamma_i} \frac{L_a}{L_i} \frac{1}{M_{an}^{EPS}}, \quad M_{an}^{EPS} = \sum_{j \in C_n} \delta_{aj} \Phi_{jn}
\]
Expanded Path Size

- Expansion factor

\[ \Phi_{jn} = \begin{cases} 
1 & \text{if } \delta_{jc} = 1 \text{ or } q(j)R_n \geq 1 \\
\frac{1}{q(j)R_n} & \text{otherwise}
\end{cases} \]

- Chosen alternative always included
- Duplicates are ignored
- Asymptotically valid

if \( R_n \to +\infty \) then \( q(j)R_n \geq 1 \ \forall \ j \in U \) and

\[ M_{an}^{\text{EPS}} \approx \sum_{j \in U} \delta_{aj} \]
Numerical Results

- Evaluation of the impact on estimation results of
  - the sampling correction,
  - the definition of the PS attribute,
  - the biased random walk algorithm parameters and
  - the definition of the postulated model used for generating the data
Numerical Results – Data
Numerical Results – Data

• Two sets of observations with two postulated Path Size Logit models

\[ U_{in} = \beta_{PS} \ln PS^U_i + \beta_L \text{Length}_i + \beta_{SB} \text{NbSB}_i + \epsilon_{in} \]

• \( \beta_{PS} = 1, \beta_{SB} = -0.1, \epsilon_{in} \) i.i.d. EV(0,1)

• \( \beta_L = -0.3 \) and \( \beta_L = -1 \)

• Path size

\[ PS^U_i = \sum_{a \in \Gamma_i} \frac{L_a}{L_i} \sum_{j \in U} \delta_{aj} \]

• 3000 observations per dataset
Numerical Results

Path number vs. Probability

- $\beta_L = -0.3$
- $\beta_L = -1$

Seminar, Namur, February 2009, A sampling of alternatives approach for route choice models – p. 24/34
Numerical Results – Models

\[ \begin{align*}
M_{\text{NoCorr}}^{PS(C)} \quad V_{in} &= \mu \left( \beta_{PS} \ln PS_{in}^C + \beta_L \text{Length}_i + \beta_{SB} \text{NbSB}_i \right) \\
M_{\text{Corr}}^{PS(C)} \quad V_{in} &= \mu \left( \beta_{PS} \ln PS_{in}^C + \beta_L \text{Length}_i + \beta_{SB} \text{NbSB}_i \right) + \ln \left( \frac{k_{in}}{q(i)} \right) \\
M_{\text{NoCorr}}^{PS(U)} \quad V_i &= \mu \left( \beta_{PS} \ln PS_{i}^U + \beta_L \text{Length}_i + \beta_{SB} \text{NbSB}_i \right) \\
M_{\text{Corr}}^{PS(U)} \quad V_{in} &= \mu \left( \beta_{PS} \ln PS_{i}^U + \beta_L \text{Length}_i + \beta_{SB} \text{NbSB}_i \right) + \ln \left( \frac{k_{in}}{q(i)} \right) \\
M_{\text{Corr}}^{EPS} \quad V_{in} &= \mu \left( \beta_{PS} \ln EPS_{in} + \beta_L \text{Length}_i + \beta_{SB} \text{NbSB}_i \right) + \ln \left( \frac{k_{in}}{q(i)} \right)
\end{align*} \]
Numerical Results – Sampling

- Number of draws: 10, 20, 40, 80, 120, 170, 250
- Parameters of the distribution: $b_1 = 1, 2, 3, 5, 10, 15, 20$ with $b_2$ always fixed to one
Numerical Results

![Graph showing numerical results](image.png)

- $b_1 = 1$
- $b_1 = 2$
- $b_1 = 3$

Number of draws vs. Average choice set size
Numerical Results – Estimation

- Over 300 models have been estimated
  Two datasets, five model specifications, different settings of the sampling algorithm
- Detailed results for one setting: $\beta_L = -1$ dataset using 40 draws and $b_1 = 1$
  - Average size of sampled choice sets: 30.92
  - BIOGEME (Bierlaire, 2003, and Bierlaire, 2007)
### Numerical Results

<table>
<thead>
<tr>
<th></th>
<th>$M_{PS(U)}^{NoCorr}$</th>
<th>$M_{PS(U)}^{Corr}$</th>
<th>$M_{PS(C)}^{NoCorr}$</th>
<th>$M_{PS(C)}^{Corr}$</th>
<th>$M_{EPS}^{Corr}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\beta}_{PS}$</td>
<td>-0.108</td>
<td>0.969</td>
<td>0.285</td>
<td>0.397</td>
<td>1.09</td>
</tr>
<tr>
<td>Rob. std</td>
<td>0.045</td>
<td>0.0541</td>
<td>0.0785</td>
<td>0.074</td>
<td>0.052</td>
</tr>
<tr>
<td>Rob. $t$-test 1</td>
<td>-24.62</td>
<td>-0.57</td>
<td>-9.11</td>
<td>-8.15</td>
<td>1.73</td>
</tr>
<tr>
<td>$\hat{\beta}_{SB}$</td>
<td>-0.547</td>
<td>-0.0849</td>
<td>-0.52</td>
<td>-0.00941</td>
<td>-0.109</td>
</tr>
<tr>
<td>Rob. std</td>
<td>0.0322</td>
<td>0.0262</td>
<td>0.0331</td>
<td>0.0261</td>
<td>0.0281</td>
</tr>
<tr>
<td>Rob. $t$-test -0.1</td>
<td>-13.88</td>
<td>0.58</td>
<td>-12.69</td>
<td>3.47</td>
<td>-0.32</td>
</tr>
<tr>
<td>$\hat{\mu}$</td>
<td>1.04</td>
<td>0.983</td>
<td>1.05</td>
<td>0.945</td>
<td>1.05</td>
</tr>
<tr>
<td>Rob. std</td>
<td>0.0314</td>
<td>0.028</td>
<td>0.0316</td>
<td>0.0264</td>
<td>0.0314</td>
</tr>
<tr>
<td>Rob. $t$-test 1</td>
<td>1.27</td>
<td>-0.61</td>
<td>1.58</td>
<td>-2.08</td>
<td>1.59</td>
</tr>
<tr>
<td>Final L-L</td>
<td>-7284.711</td>
<td>-6966.668</td>
<td>-7281.035</td>
<td>-7160.154</td>
<td>-6704.515</td>
</tr>
<tr>
<td>Adj. rho bar sq.</td>
<td>0.291</td>
<td>0.322</td>
<td>0.292</td>
<td>0.303</td>
<td>0.348</td>
</tr>
</tbody>
</table>
Numerical Results

\[ \beta_L = -1, \ M_{Corr}^{EPS} \]

\[ \beta_L = -1, \ M_{Corr}^{PS(C)} \]
Numerical Results

\[ \beta_L = -0.3, M_{\text{EPS}}^{\text{Corr}} \]

\[ \beta_L = -0.3, M_{\text{PS(C)}}^{\text{Corr}} \]

---

**Speed Bump Coef. \( \hat{\beta}_{SB} \)**
- \( b_1 = 1 \)
- \( +b_1 = 2 \)
- \( \circ b_1 = 3 \)

**Scale Param. \( \hat{\mu} \)**

**PS Coef. \( \hat{\beta}_{PS} \)**
Conclusions

- Network
- Sampling of paths
- Expanded Path Size
- Choice sets
- Sampling correction
- Route choice model
- Path Observations
- Trips

Description of correlation
Conclusions

- New paradigm for choice set generation and route choice model estimation
- Aim: avoid bias in the econometric model
- Path generation is importance sampling of alternatives
- Sampling correction of path utilities and Path Size attribute
- Numerical results show that true parameter values can be retrieved
Future Work

- Results based on real data
- Forecasting