Capturing Correlation in Route Choice Models using Subnetworks

Emma Frejinger and Michel Bierlaire
Transport and Mobility Laboratory (TRANSP-OR), EPFL
Outline

- Issues of route choice analysis
- Modelling correlation with subnetworks
  - Methodology
  - Example
- Empirical results
  - Borlänge GPS dataset
  - Estimation results
  - Forecasting results
- Conclusion and future work
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?

• Issues:
  • Universal choice set very large
  • Correlated alternatives due to overlapping paths
  • Data collection issues
Route Choice Modelling

- Deterministic utility maximisation e.g. shortest path assumption is behaviourally unrealistic
- Random utility models
  Utility $U_{in}$ an individual $n$ associates with alternative $i$:
  \[
  U_{in} = V_{in} + \varepsilon_{in}
  \]
  where $V_{in} = \beta^T X_{in}$ is the deterministic part and $\varepsilon_{in}$ is the random term
Route Choice Models

- Few models explicitly capturing correlation have been used on route choice problems of real size
  - C-Logit (Cascetta et al., 1996)
  - Path Size Logit (Ben-Akiva and Bierlaire, 1999)
  - Link-Nested Logit (Vovsha and Bekhor, 1998)
  - Logit Kernel model adapted to route choice situation (Bekhor et al., 2002)
- Probit model (Daganzo, 1977) permits an arbitrarily covariance structure specification but can rarely be applied in a real size route choice context
Subnetworks

How can we explicitly capture the most important correlation structure without considerably increasing the model complexity?
Subnetworks

How can we explicitly capture the most important correlation structure without considerably increasing the model complexity?

- Which are the behaviourally important decisions?
Subnetworks

How can we explicitly capture the most important correlation structure without considerably increasing the model complexity?

- Which are the behaviourally important decisions?
- Our hypothesis: choice of specific parts of the network (e.g. main roads, city centre)
- Concept: subnetwork
Subnetworks

- Subnetwork approach designed to be behaviourally realistic and convenient for the analyst
- Subnetwork component is a set of links corresponding to a part of the network which can be easily labelled
- Paths sharing a subnetwork component are assumed to be correlated even if they are not physically overlapping
Subnetworks - Methodology

- Factor analytic specification of an error component model (based on model presented in Bekhor et al., 2002)

\[ U_n = \beta^T X_n + F_n T \zeta_n + \nu_n \]

- \( F_n (JxQ) \) : factor loadings matrix
- \( (f_n)_{iq} = \sqrt{l_{niq}} \)
- \( T_{(QxQ)} = \text{diag} (\sigma_1, \sigma_2, \ldots, \sigma_Q) \)
- \( \zeta_n (Qx1) \) : vector of i.i.d. N(0,1) variates
- \( \nu (Jx1) \) : vector of i.i.d. Extreme Value distributed variates
Subnetworks - Example

Path 1
Path 2
Path 3

Path 1
Path 2
Path 3
Subnetworks - Example

\begin{align*}
U_1 &= \beta^T X_1 + \sqrt{l_{1a}} \sigma_a \zeta_a + \sqrt{l_{1b}} \sigma_b \zeta_b + \nu_1 \\
U_2 &= \beta^T X_2 + \sqrt{l_{2a}} \sigma_a \zeta_a + \nu_2 \\
U_3 &= \beta^T X_3 + \sqrt{l_{3b}} \sigma_b \zeta_b + \nu_3
\end{align*}

\[
\mathbf{FTT}^T \mathbf{FT} = \\
\begin{bmatrix}
l_{1a} \sigma_a^2 + l_{1b} \sigma_b^2 & \sqrt{l_{1a}} \sqrt{l_{2a}} \sigma_a^2 & \sqrt{l_{1b}} \sqrt{l_{3b}} \sigma_b^2 \\
\sqrt{l_{1a}} \sqrt{l_{2a}} \sigma_a^2 & l_{2a} \sigma_a^2 & 0 \\
\sqrt{l_{3b}} \sqrt{l_{1b}} \sigma_b^2 & 0 & l_{3b} \sigma_b^2
\end{bmatrix}
\]
Empirical Results

- The approach has been tested on three datasets: Boston (Ramming, 2001), Switzerland, and Borlänge.
- Deterministic choice set generation
  - Link elimination
- GPS data from 24 individuals
  - 2978 observations, 2179 origin-destination pairs
- Borlänge network
  - 3077 nodes and 7459 links
- BIOGEME (biogeme.epfl.ch, Bierlaire, 2003) has been used for all model estimations.
Borlänge Road Network
## Subnetwork Components

<table>
<thead>
<tr>
<th></th>
<th>R.50 S</th>
<th>R.50 N</th>
<th>R.70 S</th>
<th>R.70 N</th>
<th>R.C.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Component length [m]</td>
<td>5255</td>
<td>4966</td>
<td>11362</td>
<td>7028</td>
<td>1733</td>
</tr>
<tr>
<td>Nb. of Observations</td>
<td>173</td>
<td>153</td>
<td>261</td>
<td>366</td>
<td>209</td>
</tr>
<tr>
<td>Weighted Nb. of Observations</td>
<td>36</td>
<td>88</td>
<td>65</td>
<td>73</td>
<td>116</td>
</tr>
</tbody>
</table>

\[
N_q = \sum_{o \in O} \frac{l_{oq}}{L_q}
\]

Capturing Correlation with Subnetworks in Route Choice Models – p.13/21
Model Specifications

- Six different models: MNL, PSL, EC\(_1\), EC\(_1'\), EC\(_2\) and EC\(_2'\).
- EC\(_1\) and EC\(_1'\) have a simplified correlation structure.
- EC\(_1'\) and EC\(_2'\) do not include a Path Size attribute.
- Deterministic part of the utility:

\[
V_i = \beta_{PS} \ln(PS_i) + \beta_{EstimatedTime}EstimatedTime_i + \\
\beta_{NbSpeedBumps}NbSpeedBumps_i + \beta_{NbLeftTurns}NbLeftTurns_i + \\
\beta_{AvgLinkLength}AvgLinkLength_i
\]
Estimation Results

- Parameter estimates for explanatory variables are stable across the different models.

- Path size parameter estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>PSL</th>
<th>EC&lt;sub&gt;1&lt;/sub&gt;</th>
<th>EC&lt;sub&gt;2&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Path Size</td>
<td>-0.28</td>
<td>-0.49</td>
<td>-0.53</td>
</tr>
<tr>
<td>Scaled estimate</td>
<td>-0.33</td>
<td>-0.53</td>
<td>-0.56</td>
</tr>
<tr>
<td>Rob. T-test 0</td>
<td>-4.05</td>
<td>-5.61</td>
<td>-5.91</td>
</tr>
</tbody>
</table>

- All covariance parameters estimates in the different models are significant except the one associated with R.50 S.
### Estimation Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Nb. σ Estimates</th>
<th>Nb. Estimated Parameters</th>
<th>Final L-L</th>
<th>Adjusted Rho-Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNL</td>
<td>-</td>
<td>12</td>
<td>-4186.07</td>
<td>0.152</td>
</tr>
<tr>
<td>PSL</td>
<td>-</td>
<td>13</td>
<td>-4174.72</td>
<td>0.154</td>
</tr>
<tr>
<td>EC₁ (with PS)</td>
<td>1</td>
<td>14</td>
<td>-4142.40</td>
<td>0.161</td>
</tr>
<tr>
<td>EC'₁</td>
<td>1</td>
<td>13</td>
<td>-4165.59</td>
<td>0.156</td>
</tr>
<tr>
<td>EC₂ (with PS)</td>
<td>5</td>
<td>18</td>
<td>-4136.92</td>
<td>0.161</td>
</tr>
<tr>
<td>EC'₂</td>
<td>5</td>
<td>17</td>
<td>-4162.74</td>
<td>0.156</td>
</tr>
</tbody>
</table>

1000 pseudo-random draws for Maximum Simulated Likelihood estimation
2978 observations
Null log likelihood: -4951.11
BIOGEME (biogeme.epfl.ch) has been used for all model estimations.
Forecasting Results

- Comparison of the different models in terms of their performance of predicting choice probabilities
- Five subsamples of the dataset
  - Observations corresponding to 80% of the origin destination pairs (randomly chosen) are used for estimating the models
  - The models are applied on the observations corresponding to the other 20% of the origin destination pairs
- Comparison of final log-likelihood values
Forecasting Results

- Same specification of deterministic utility function for all models
- Same interpretation of these models as for those estimated on the complete dataset
- Coefficient and covariance parameter values are stable across models
Forecasting Results

![Bar chart showing the Log Likelihood Value for different data sets and models. The models include MNL, PSL, EC, and EC'. The data sets range from Data 1 to Data 5.]
Conclusion

- Models based on subnetworks are designed for route choice modelling of realistic size
- Correlation on subnetwork is explicitly captured within a factor analytic specification of an Error Component model
- Estimation and prediction results clearly shows the superiority of the Error Component models compared to PSL and MNL
Conclusion

- The subnetwork approach is flexible and the trade-off between complexity and behavioural realism can be controlled by the analyst.

- Paper to appear in Transportation Research Part B.

Future work:
- Analysis of the sensitivity of the results regarding the definition of the subnetwork.
- Influence of choice set generation algorithm.