Transportation Research Part B 80 (2015) 220–234

Contents lists available at ScienceDirect

Transportation Research Part B

journal homepage: www.elsevier.com/locate/trb

Specification of the cross-nested logit model with sampling of alternatives for route choice models ☆

Xinjun Lai [a,*], Michel Bierlaire [b]

[a] *Sun Yat-sen University, School of Engineering, China*
[b] *École Polytechnique Fédérale de Lausanne (EPFL), School of Architecture, Civil and Environmental Engineering (ENAC), Transport and Mobility Laboratory, Switzerland*

ARTICLE INFO

Article history:
Received 4 June 2014
Received in revised form 12 June 2015
Accepted 4 July 2015
Available online 30 July 2015

Keywords:
Cross-nested logit
Sampling of alternatives
Route choice
Metropolis–Hasting

ABSTRACT

We present an operational estimation procedure for the estimation of route choice multivariate extreme value (MEV) models based on sampling of alternatives. The procedure builds on the state-of-the-art literature, and in particular on recent methodological developments proposed by Flötteröd and Bierlaire (2013) and Guevara and Ben-Akiva (2013b). Case studies on both synthetic data and a real network demonstrate that the new method is valid and practical.

1. Introduction

A route choice model predicts the probability that one traveler chooses a certain route between a given origin and destination (OD). It is a challenging problem in transportation research because of its large choice set and its highly correlated structure.

In order to address the correlation issue in route choice, the path size and C-logit models are the most frequently used methods due to their simplicity. They both add an additional term to the utility of a logit specification to compensate for the correlation of paths (Cascetta et al., 1996; Ben-Akiva and Bierlaire, 1999). However, the additional terms are empirical approximations and previous research shows that they might be too sensitive to the composition of the choice set (Bovy et al., 2008; Li et al., 2013). The cross-nested logit model explicitly captures the correlation among paths, where alternatives are classified into nests if they share common unobserved attributes (Wen and Koppelman, 2001; Bierlaire, 2006). The link-based cross-nested logit (CNL) model has been proposed by Vovsha and Bekhor (1998). It assumes that each link corresponds to a nest and the paths that share the same link belongs to the same nest (Prashker and Bekhor, 1999; Bekhor and Prashker, 2001; Ramming et al., 2002). However it is difficult to estimate because of the large number of parameters. Therefore, it is common to empirically set their values based on the network topology. The Error Component model (Ramming et al., 2002; Frejinger and Bierlaire, 2007) is also used in the route choice context. It is a mixture logit model and requires simulation-based estimation, which is cumbersome for large-scale applications.

☆ Manuscript submitted for possible publication to *Transportation Research Part B* – Revised version.

* Corresponding author.
E-mail addresses: laixinj@mail2.sysu.edu.cn (X. Lai), michel.bierlaire@epfl.ch (M. Bierlaire).

http://dx.doi.org/10.1016/j.trb.2015.07.005

The models described above require the knowledge of the full choice set which is unfeasible in route choice. In a road network with 80 nodes and 0.4 density,[1] between any OD pairs there are approximately 10^{84} possible routes without loop (Roberts and Kroese, 2007). Enumerating the alternatives in such a choice set is impractical, justifying the need to sample a subset of paths with a reasonable size. Frejinger and Bierlaire (2010) analyze several existing algorithms for the generation of the paths. The most desirable features in this framework are that (i) the chosen alternative is included in the generated set, and (ii) the algorithm is able to generate the sampled probabilities that are required to correct for the sampling biases. A random walk algorithm is proposed to first address the issue of paths sampling (Frejinger et al., 2009). Flötteröd and Bierlaire (2013) present a Metropolis–Hasting (MH) algorithm to sample paths between an origin and a destination with pre-specified sampling probabilities. The MH algorithm is an accept/reject algorithm based on an underlying Markov Chain process. It is particularly well suited for complex sampling distributions with very large support, as the exact probability mass function (pmf) is not needed. The weights used to compute the acceptance probability must be proportional to the sampling probability, so that the normalization factor (involving the enumeration of the full choice set) is not needed (Ross, 2012). Recently, Fosgerau et al. (2013) proposed a recursive logit model that does not require the enumeration of paths. It is defined within a dynamic programming framework, that requires solving a system of equations. Although this is a promising avenue for future research, we focus in this paper on models based on a sample of alternatives.

Regarding the estimation of such models, McFadden (1978) demonstrated that if the underlying model is logit, a correction term should be added to the log likelihood function to achieve consistent parameters estimation with sampled alternatives. There are several applications adopting McFadden's result (Parsons and Kealy, 1992; Sermons and Koppelman, 2001). Nerella and Bhat (2004) examine the effect of the sample size of alternatives on the performance of a mixed logit model. Guevara and Ben-Akiva (2013a) study the conditions needed to achieve consistency, asymptotic normality and relative efficiency for logit mixture models. Bierlaire et al. (2008) suggest a framework for the consistent estimation of MEV (Multivariate Extreme Value) models with samples of alternatives, which has been completed and made operational by Guevara and Ben-Akiva (2013b). The key idea proposed by Guevara and Ben-Akiva (2013b) is to provide an appropriate approximation of the MEV part of the model using only the sampled alternative. They introduce an expansion factor, and prove the desired theoretical properties of the associated estimator. They successfully apply the method to a real data case of residential location choice and a nested logit model.

In this paper, we combine and adapt the recent methodological developments described above in order to obtain a comprehensive and practical methodology for the estimation of route choice models in real networks, based on the estimation of CNL route choice models with sampling of alternatives. The CNL model is appealing because of its closed form. Moreover, Fosgerau et al. (2013) have shown that any random utility model can be approached as close as needed by a cross-nested logit model. The methodology is an extension of Guevara and Ben-Akiva (2013b) adapted for the specific context of route choice. A new expansion factor is provided to achieve consistency and asymptotic normality in model estimation without the need for a full enumeration of the choice set. The methodology is illustrated with synthetic data and real data. Results suggest that the proposed method is practical and outperforms the models without corrections.

The paper is organized as follows. Section 2 describes the methodology, including the sampling of alternatives, the approximation of the cross-nested logit model with the sampled paths, and the expansion factors. The validation on synthetic data is presented in Section 3. The new method is applied on real data in Section 4. Finally conclusions and discussions for future study are given in Section 5.

2. Methodology

The CNL route choice model derived in this paper has a similar structure as proposed by Vovsha and Bekhor (1998). Each link of the network is associated with a nest, and each path (that is each alternative) belongs to some extent to each nest corresponding to the link composing the path. A path belongs to several nests, capturing the correlation among alternatives due to the network topology. If there are M links in the network, the probability that one traveler chooses path i from the universal choice set $\mathcal{C}$ is

$$\Pr(i|\mathcal{C}) = \frac{\exp(V_i + \ln G_i(\mathcal{C}))}{\sum_{j\in\mathcal{C}} \exp(V_j + \ln G_j(\mathcal{C}))} \tag{1}$$

where

$$G_i(\mathcal{C}) = \sum_{m=1}^{M} \left[\mu \alpha_{im} e^{V_i(\mu_m - 1)} \left(\sum_{j\in\mathcal{C}} \alpha_{jm} e^{\mu_m V_j} \right)^{\frac{\mu-\mu_m}{\mu_m}} \right], \tag{2}$$

where V_i is the deterministic utility for path i, $\mu > 0$ (usually normalized to 1) is the scale parameter for the model, μ_m is the scale parameter for nest m, such that $\mu_m \geqslant \mu, \alpha_{im} \in [0, 1]$ is the inclusive parameter capturing the membership of path i to

[1] Density here is the ratio between the actual number of links in the network and the maximum number of possible links, that is $N(N+1)/2$, where N is the number of nodes.