A nested recursive logit model for route choice analysis

Mai, Tien; Frejinger, Emma; Fosgerau, Mogens

Published in:
Transportation Research Part B: Methodological

Link to article, DOI:
10.1016/j.trb.2015.03.015

Publication date:
2015

Document Version
Peer reviewed version

A nested recursive logit model for route choice analysis

Tien Mai * Mogens Fosgerau† Emma Frejinger *

March 22, 2015

Abstract

We propose a route choice model that relaxes the independence from irrelevant alternatives property of the logit model by allowing scale parameters to be link specific. Similar to the the recursive logit (RL) model proposed by Fosgerau et al. (2013), the choice of path is modelled as a sequence of link choices and the model does not require any sampling of choice sets. Furthermore, the model can be consistently estimated and efficiently used for prediction.

A key challenge lies in the computation of the value functions, i.e. the expected maximum utility from any position in the network to a destination. The value functions are the solution to a system of non-linear equations. We propose an iterative method with dynamic accuracy that allows to efficiently solve these systems.

We report estimation results and a cross-validation study for a real network. The results show that the NRL model yields sensible parameter estimates and the fit is significantly better than the RL model. Moreover, the NRL model outperforms the RL model in terms of prediction.

Keywords: route choice modelling; nested recursive logit; substitution patterns; value iterations; maximum likelihood estimation; cross-validation.

*Department of Computer Science and Operational Research, Université de Montréal and CIRRELT, Canada

†Technical University of Denmark, Denmark, and Royal Institute of Technology, Sweden.

1 Introduction

Discrete choice models are generally used for analyzing path choices in real networks based on revealed preference (RP) data. There are two main modelling issues associated with (i) estimating such models consistently and (ii) subsequently using them for prediction. First, choice sets of paths are unknown to the analyst and the set of all feasible paths for a given origin-destination pair cannot be enumerated. Second, path utilities may be correlated, for instance, due to physical overlap in the network. As we explain below, there is currently no path choice model that can be consistently estimated and used for prediction, while avoiding the specification of choice sets and allowing for correlation due to path overlap. The nested recursive logit (NRL) model, proposed in this paper, fills this gap.

Most of the existing path choice models are based on choice sets of paths that need to be sampled before estimating or applying the model. Many different algorithms exist for sampling choice sets (for reviews, see e.g. Frejinger et al., 2009, Prato, 2009) and they all correspond to importance sampling protocols where paths have non-equal probabilities of being sampled. Frejinger et al. (2009) show that utilities need to be corrected for the sampling of alternatives, which implies that only algorithms that allow computation of the path sampling probabilities can be used. Frejinger et al. (2009) use the logit (MNL) model but recently Guevara and Ben-Akiva (2013a) and Guevara and Ben-Akiva (2013b) have derived results for generalized extreme value (GEV) and mixed logit models, respectively. The sampling approach can be used to consistently estimate a path choice model, but it is still unknown how to use that model for prediction.

A three path example network is often used to illustrate why it is important to allow for correlated utilities (we present this example in more detail in Section 3). At the origin one can take right or left. Going right there are two paths that share one link except for a short distance close to destination where they separate. If all three paths have the same deterministic utility, a logit model assigns the probability 1/3 to each although one would expect a probability 1/2 going left and 1/2 going right. A number of models in the literature allow to model the correlation structure of path utilities. Examples are the link-nested logit (Vovsha and Bekhor, 1998), mixed logit with error components (Bekhor et al., 2001, Frejinger and Bierlaire, 2007) and paired combinatorial logit (Chu, 1989). These models are based on sampled choice sets without correcting the utilities for the sampling protocol. Hence, the parameter estimates are conditional on the choice sets and may have significantly different values if some paths are added or removed from the choice sets. This is problematic since the true choice sets are unknown. As men-

tioned earlier, the MEV models (e.g. link-nested logit) or the mixed logit models can be corrected. Lai and Bierlaire (2014) estimate a link-nested logit model using the results by Guevara and Ben-Akiva (2013a).

Recently, Fosgerau et al. (2013) proposed the recursive logit (RL) model where path choice is modelled as a sequence of link choices using a dynamic discrete choice framework. The RL model can be consistently estimated and used for prediction without sampling choice sets of paths. It is however equivalent to a MNL model over the set of all feasible paths. A correction attribute called link size was proposed that achieves an effect similar to the path size attribute in path choice models (Ben-Akiva and Bierlaire, 1999). These attributes correct the utilities for correlation but the models retain the independence of irrelevant alternatives (IIA) property, unless the path size/link size attributes are updated as utilities change (e.g. changes in link travel times).

In this paper we propose an extension of the RL model that allows path utilities to be correlated in a fashion similar to nested logit (Ben-Akiva, 1973, McFadden, 1978) and where links can have different scale parameters. The key challenge with this extension lies in the computation of the expected maximum utility from a current position in the network until the destination (value functions). A computational advantage of the RL model is that the value functions can be computed by solving a system of linear equations, which is fast and easy to do. In the case of the NRL, the value functions are a solution to a system of *non-linear* equations which is substantially more difficult to deal with. We propose an iterative method with dynamic accuracy to efficiently solve this equation system.

This paper makes a number of contributions. First we propose a model that can be consistently estimated and used for prediction without sampling choice sets *while allowing the random terms to be correlated*. Second, we provide illustrative examples and discuss substitution patterns in order to build an intuition on the properties of the model. Third, we propose an iterative method to solve for the value functions and we derive the analytical gradient of the log-likelihood function for the case that the scales are functions of model parameters so that the NRL model can be efficiently estimated. Fourth, we provide estimation and cross-validation results for a real network using simulated and real observations. Finally, the estimation code is implemented in MATLAB and is freely available upon request.

The paper is structured as follows. Section 2 presents the NRL model. Section 3 discusses substitution patterns by illustrative examples and Section 4 provide a method to compute the value functions. Section 5 derives an analytical formula for the first order derivative of the log-likelihood function. Specifications, estimation and prediction results are presented in Section 6

and finally Section 7 concludes.

2 The nested recursive logit model

In the RL model (Fosgerau et al., 2013) the path choice problem is formulated as a sequence of link choices and modelled in a dynamic discrete choice framework. At each node the decision maker chooses the utility-maximizing outgoing link with link utilities given by the instantaneous utility and the expected maximum utility to the destination. The random terms of the instantaneous utilities are independently and identically distributed (i.i.d.) extreme value type I so that the model is equivalent to MNL. In this section we present the NRL model which relaxes the IIA property of MNL by assuming that the scales of random terms are non-equal across links. We derive the NRL model using the same notation as Fosgerau et al. (2013) (we refer the reader to that paper for a more detailed presentation of the notation). Even though the derivation of NRL is similar to the RL one, the resulting expressions of the value functions and path choice probabilities have important differences.

A directed connected graph (not assumed acyclic) $G = (A, \mathcal{V})$ is considered, where A and $\mathcal{V}$ are the sets of links and nodes, respectively. For each link $k \in A$, we denote the set of outgoing links from the sink node of k by $A(k)$. Moreover we associate an absorbing state with each destination by extending the network with dummy links d (see Figure 1). This is a link without successors so a trip stops once this state is reached. The set of all links is $\tilde{A} = A \cup \{d\}$ and the corresponding deterministic utility is $v(d|k) = 0$ for all k that have destination d as sink node. Given two links $a, k \in \tilde{A}$, the following instantaneous utility is associated with action $a \in A(k)$ of individual n

$$u^n(a|k;\beta) = v^n(a|k;\beta) + \mu_k \epsilon(a) \tag{1}$$

where β is a vector of parameters, $\epsilon(a)$ are i.i.d extreme value type I and μ_k is a strictly positive scale parameter. We ensure that $\epsilon(a)$ have zero mean by subtracting Euler's constant. The deterministic term $v^n(a|k;\beta)$ is assumed negative for all links except the dummy link d. We emphasize the difference with the original RL model where scale parameters are assumed equal ($\mu_k = \mu \ \forall k \in A$). For notational simplicity, we omit an index for individual n but note that the utilities can be individual specific.

The expected maximum utility from the sink node of k to the destination is the value function $V^d(k;\beta)$. The superscript d indicates that the value functions are destination specific and they also depend on parameters β.

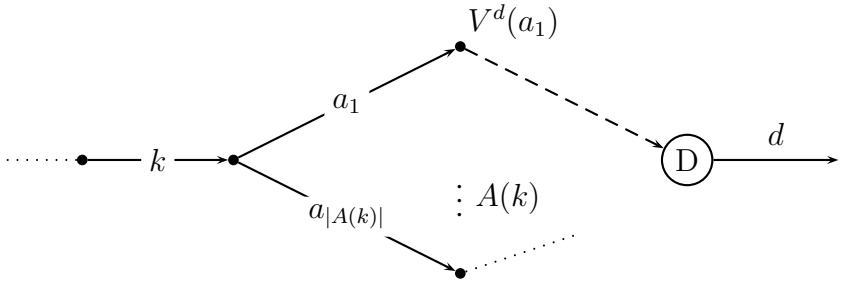

Figure 1: Illustration of notation

$V^d(k;\beta)$ is recursively defined by Bellman's equation

$$V^d(k;\beta) = \mathbb{E}\Big[\max_{a\in A(k)} \big(v(a|k;\beta) + V^d(a;\beta) + \mu_k\epsilon(a)\big)\Big] \quad \forall k \in A \tag{2}$$

or equivalently

$$\frac{1}{\mu_k}V^d(k;\beta) = \mathbb{E}\Big[\max_{a\in A(k)} \big(\frac{1}{\mu_k}(v(a|k;\beta) + V^d(a;\beta)) + \epsilon(a)\big)\Big] \quad \forall k \in A. \tag{3}$$

For notational simplicity we omit from now on β from the value functions $V(.)$ and the utilities $v(.)$.

Given these assumptions the probability of choosing link a given state k is given by the MNL model

$$\begin{aligned} P^d(a|k) &= \delta(a|k)\frac{e^{\frac{1}{\mu_k}(v(a|k)+V^d(a))}}{\sum_{a'\in A(k)} e^{\frac{1}{\mu_k}(v(a'|k)+V^d(a'))}} \\ &= \delta(a|k)e^{\frac{1}{\mu_k}(v(a|k)+V^d(a)-V^d(k))} \quad \forall k, a \in \tilde{A}. \end{aligned} \tag{4}$$

Note that we include $\delta(a|k)$ that equals one if $a \in A(k)$ and zero otherwise so that the probability is defined for all $a, k \in \tilde{A}$ (we recall that $\tilde{A} = A \cup \{d\}$). Since we assume that the random terms in (1) are distributed i.i.d. EV type I, the value functions (2) are given recursively by the logsum

$$\frac{1}{\mu_k}V^d(k) = \ln\Big(\sum_{a\in A(k)} e^{\frac{1}{\mu_k}(v(a|k)+V^d(a))}\Big) \quad \forall k \in A \tag{5}$$

and $V^d(d) = 0$ by assumption. Similar to Fosgerau et al. (2013) we can write (5) as

$$e^{\frac{1}{\mu_k}V^d(k)} = \begin{cases} \sum_{a\in A} \delta(a|k)e^{\frac{(v(a|k)+V^d(a))}{\mu_k}} & \forall k \in A \\ 1 & k = d \end{cases} \tag{6}$$

and define a matrix $M^d(|\tilde{A}| \times |\tilde{A}|)$ and a vector $z^d(|\tilde{A}| \times 1)$ with entries

$$M^d_{ka} = \delta(a|k)e^{\frac{v(a|k)}{\mu_k}},\ z^d_k = e^{\frac{V(k)}{\mu_k}}, \quad k, a \in \tilde{A}. \tag{7}$$

The key issue here compared to the RL model is that we do not end up with a system of linear equations. Indeed, the value functions are the solutions to the following system of non-linear equations

$$z^d_k = \begin{cases} \sum_{a \in A} M^d_{ka} (z^d_a)^{\mu_a/\mu_k} & \forall k \in A \\ 1 & k = d, \end{cases} \tag{8}$$

where the non-linearity arises due to the scale parameters μ_k not being equal.

The probability of a path σ defined by a sequence of links $\sigma = [k_0, k_1, \ldots, k_I]$ is

$$P(\sigma) = \prod_{i=0}^{I-1} e^{\frac{1}{\mu_{k_i}}(v(k_{i+1}|k_i) + V^d(k_{i+1}) - V^d(k_i))}. \tag{9}$$

Unlike the RL model, the link specific value functions do not cancel out due to the scale parameters. This implies that the path choice probabilities are computationally more costly to evaluate.

We note that if the network contains cycles, the RL and NRL model allow for paths to contain loops (Akamatsu, 1996, Fosgerau et al., 2013, discuss this in more detail). The probability of paths with loops depend on the data and network structure. For the data used in this paper, Fosgerau et al. (2013) report that paths with loops have a very small probability.

Finally we note that the IIA property does not hold in the NRL model. Consider the ratio of the choice probabilities of two paths $\sigma_1 = [k_1, \ldots, k_{I_1}]$ and $\sigma_2 = [h_1, \ldots, h_{I_2}]$ connecting just one origin-destination pair

$$\frac{P(\sigma_1)}{P(\sigma_2)} = \frac{\prod_{i=1}^{I_1-1} e^{\frac{1}{\mu_{k_i}}(v(k_{i+1}|k_i) + V^d(k_{i+1}) - V^d(k_i))}}{\prod_{i=1}^{I_2-1} e^{\frac{1}{\mu_{h_i}}(v(h_{i+1}|h_i) + V^d(h_{i+1}) - V^d(h_i))}}. \tag{10}$$

When the scales $\mu_k = \mu\ \forall k \in A$, the value function terms cancel out and the ratio (10) then only depends on the utilities of two considered paths. For the NRL model, the ratio (10) depends on several values functions, which are evaluated based on the whole network and therefore the IIA property does not hold. In the following section we discuss the resulting substitution pattern in more depth using several illustrative examples.

3 Illustrative examples and substitution patterns

Similar to several studies in the literature (e.g. Ben-Akiva and Bierlaire, 1999), we use a simple three path network shown in 2 to illustrate why it is important to allow for correlated utilities. There are three paths from o to d (link o is the origin and link d is the destination dummy link): $[o, a, d]$, $[o, b, e, d]$, $[o, b, f, d]$. We number these paths 1,2 and 3 and the corresponding path probabilities are P_1, P_2 and P_3, respectively. The only attribute in the instantaneous utility is link length and the values are given in the parentheses on each arc. In order to compute path probabilities we choose a length parameter $\tilde{\beta} = -1$.

When the scales of random terms are equal over links $\mu_k = \mu$, the model corresponds the RL and $P_1 = P_2 = P_3 = 1/3$. When the network has a perfect nested structure as this one (each path in the network belongs to exactly one nest when defined by physical overlap), the NRL model is equivalent to a nested logit model. We can illustrate this by fixing all the scale parameters to 1 except μ_b that we vary over the interval $(0, 1]$. The path probabilities are plotted in the graph on the right hand side in Figure 2.

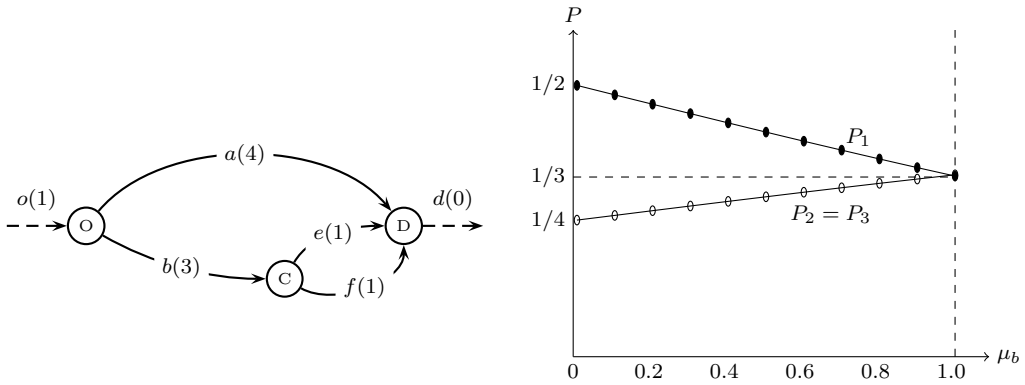

Figure 2: Classic three paths example network

In order to build intuition on the substitution patterns implied by the NRL model, we provide three more examples. The first is shown in Figure 3 which also has a simple nested structure. There are 4 nodes A, B, C, D and 9 links. Moreover, there are 6 possible paths from o to d: $[o, a, a_1, d]$, $[o, a, a_2, d]$, $[o, a, a_3, d]$, $[o, b, b_1, d]$, $[o, b, b_2, d]$ and $[o, b, b_3, d]$ and we number these paths as 1, 2, 3, 4, 5 and 6, respectively.

For the RL model the IIA property holds, meaning that, if we remove

any link in the network, the probabilities of the remaining feasible paths will increase by the same proportion (for example if we remove link a_2, the probabilities of path $[o, a, a_3, d]$ and path $[o, b, b_3, d]$ increase but they are still equal). For the NRL model, the scales of random terms are assigned different values. We assign a scale of 0.5 for links a, a scale of 0.8 for links b and a scale of 1.0 for the others. Similar to an example in Train (2003), we illustrate substitution patterns by removing in turn links a_1, a_2, b_1, b_2 and present changes in probabilities in Table 1.

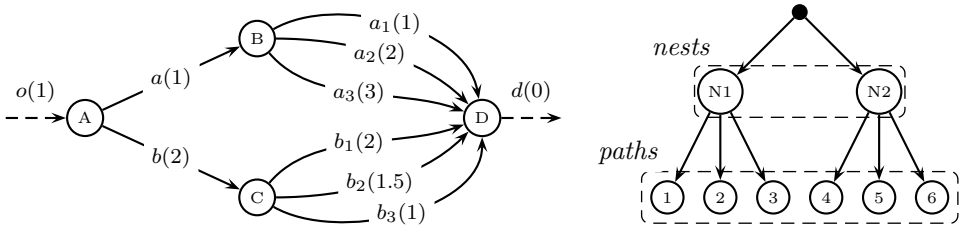

Figure 3: Example network with perfect nested structure

		Probabilities with link removed			
Paths	Original	a_1	a_2	b_1	b_2
$1 : [o, a, a_1, d]$	0.54	-	0.65(+20%)	0.55(+1%)	0.56(+4%)
$2 : [o, a, a_2, d]$	0.15	0.38(+151%)	-	0.16(+1%)	0.16(+4%)
$3 : [o, a, a_3, d]$	0.04	0.11(+151%)	0.05(+20%)	0.05(+1%)	0.05(+4%)
$4 : [o, b, b_1, d]$	0.02	0.05(+93%)	0.03(+15%)	-	0.03(+19%)
$5 : [o, b, b_2, d]$	0.06	0.12(+93%)	0.07(+15%)	0.17(+6%)	-
$6 : [o, b, b_3, d]$	0.17	0.33(+93%)	0.20(+15%)	0.18(+6%)	0.21(+19%)

Table 1: Change in probability when link is removed (example network with perfect nested structure)

We note that the probabilities for paths $[o, a, a_1, d]$, $[o, a, a_2, d]$, $[o, a, a_3, d]$ rise by the same proportions whenever one link is removed from the network. This is also the case for the three paths $[o, b, b_1, d]$, $[o, b, b_2, d]$ and $[o, b, b_2, d]$. As expected, the IIA property holds between paths within the same nest but not for paths in different nests. For example, when link a_1 is removed, the probabilities of the paths in the first nest rise by 151% while the paths in the second nest rise by 93%.

We also consider the case when a link from node B to C is added to the network in Figure 3. This change adds three more paths to nest N1. In Table 2 we report the change in probabilities for the same six paths as

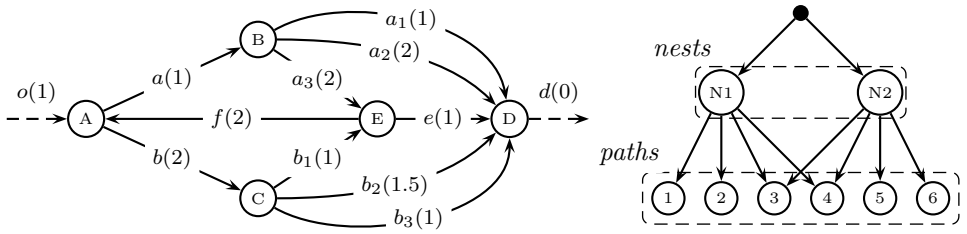

Figure 4: Example network with cross-nested structure

before. We note that the absolute values of choice probabilities change but the substitution pattern remains proportional.

		Probabilities with link removed			
Paths	Original	a_1	a_2	b_1	b_2
$1:[o,a,a_1,d]$	0.487	-	0.572(17.52%)	0.504(3.48%)	0.522(7.27%)
$2:[o,a,a_2,d]$	0.140	0.298(113.38%)	-	0.144(3.48%)	0.150(7.27%)
$3:[o,a,a_3,d]$	0.040	0.085(113.38%)	0.047(17.52%)	0.041(3.48%)	0.043(7.27%)
$4:[o,b,b_1,d]$	0.022	0.038(73.88%)	0.024(12.84%)	-	0.026(22.29%)
$5:[o,b,b_2,d]$	0.059	0.102(73.88%)	0.066(12.84%)	0.063(7.86%)	-
$6:[o,b,b_3,d]$	0.160	0.278(73.88%)	0.180(12.84%)	0.172(7.86%)	0.195(22.29%)

Table 2: Change in probability when link is removed (example network with perfect nested structure with link from B to C)

The network in Figure 3 is designed so that the paths can naturally be divided into separate nests. In the next example shown in Figure 4 we slightly modify the network so that paths have a cross-nested structure. More precisely, we add a node E that splits links a_3 and b_1 into two links. The lengths of the paths in the new network do not change but the structure of the network is different since apart from the origin and destination, two paths $[o,a,a_3,e,d]$ and $[o,a,b_1,e,d]$ share link e. Furthermore, there is a new link f going (backward) from node E to node A so that the expected maximum utilities from link a_3 and b_1 depend on the whole network.

We report probabilities for the 6 paths without loops: $[o,a,a_1,d]$, $[o,a,a_2,d]$, $[o,a,a_3,e,d]$, $[o,b,b_1,e,d]$, $[o,b,b_2,d]$, $[o,b,b_3,d]$, which are numbered as 1, 2, 3, 4, 5 and 6, respectively. We keep the same scales as in the first example (i.e. $\mu_a = 0.5$, $\mu_b = 0.8$ and the other scale parameters are equal to one). The changes in probabilities of the six paths when we remove in turn links a_3, b_1 and f are reported in Table 3. We note that the substitution patterns are different than in the previous example since the probabilities of paths 3

and 4 no longer change by the same proportion as the other paths in their respective nest.

		Probabilities with link removed		
Paths	Original	a_1	b_3	f
$1:[o,a,a_1,d]$	0.54	-	0.60(+12%)	0.54(+0.7%)
$2:[o,a,a_2,d]$	0.15	0.38(+150%)	0.17(+12%)	0.15(+0.7%)
$3:[o,a,a_3,e,d]$	0.05	0.11(+148%)	0.05(+11%)	0.04(-1.3%)
$4:[o,b,b_1,e,d]$	0.03	0.05(+86%)	0.05(+90%)	0.02(-6.7%)
$5:[o,b,b_2,d]$	0.06	0.12(+93%)	0.12(+91%)	0.06(+1.4%)
$6:[o,b,b_3,d]$	0.17	0.33(+93%)	-	0.17(+1.4%)

Table 3: Change in probability when link is removed (example network with cross-nested structure)

In order to compare the results with path based models we report probabilities given by the nested logit and link-nested logit (Vovsha and Bekhor, 1998) models in Table 4. The correlation structure given by the link-nested logit model is shown in Figure 5. For the nested models, the nesting parameters take the same values as in the NRL mode, namely 0.8 for nest $N1$ and 0.5 for nest $N2$. The results show that for these examples, the probabilities of the nested model are identical to the NRL model and probabilities of the link-nested logit are slightly different from NRL. We note that the sums of the path probabilities for RL and NRL in the second example are slightly smaller than one, due to the cycle in the network.

In summary, the IIA property can be relaxed by assuming different scales. The resulting substitution pattern depends on the network structure. If the network has a perfect nested structure (e.g Figure 3) the NRL and nested logit models yield the same results.

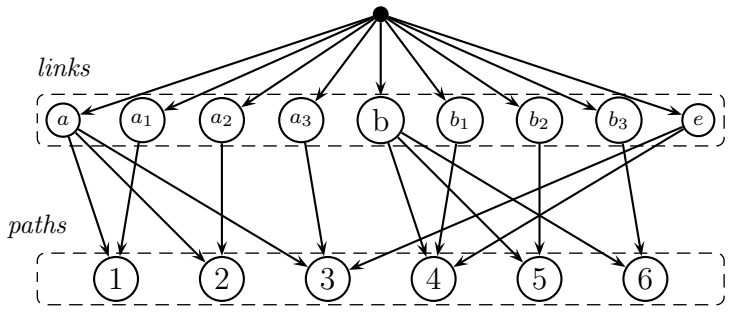

Figure 5: Cross-nested structure from the Link-nested logit model

Paths	Example 1			Example 2		
	MNL	NRL	Nested logit	MNL	NRL	Link nested logit
1	0.449	0.541	0.541	0.443	0.537	0.501
2	0.165	0.155	0.155	0.163	0.154	0.150
3	0.061	0.044	0.044	0.060	0.045	0.051
4	0.061	0.023	0.023	0.060	0.025	0.043
5	0.100	0.064	0.064	0.099	0.063	0.085
6	0.165	0.173	0.173	0.163	0.170	0.171

Table 4: Path probabilities comparison

4 Computation of the value functions

The main challenge associated with the NRL model is to efficiently solve the large-scale system of system of non-linear equations (6). In the following we describe a value iteration approach that is efficient thanks to (i) a good initial solution and (ii) dynamic accuracy.

We define a matrix $X(z)$ with entries

$$X(z)_{ka} = z_a^{\mu_a/\mu_k} \quad \forall k, a \in \tilde{A} \tag{11}$$

so that the Bellman equation (8) can be written as

$$z = [M \circ X(z)]e + b. \tag{12}$$

b is a vector of size $(|\tilde{A}| \times 1)$ with zero values for all states except for the destination that equals 1, e is a vector of size $(|\tilde{A}| \times 1)$ with value one for all states and $\circ$ is the element-by-element product.

Value iterations are based on Equation (12). We start with an initial vector z^0 and then for each iteration i we compute a new vector

$$z^{i+1} \leftarrow [M \circ X(z^i)]e + b. \tag{13}$$

and iterate until a fixed point is found using $||z^{i+1} - z^i||^2 < \gamma$ for a given threshold $\gamma > 0$ as stopping criteria.[1] It can be shown that if the Bellman equation has a solution, this method converges after a finite number of iterations (see for instance Rust, 1987, 1988). The choice of initial vector is however important for the rate of convergence. We use the solution of the system of linear equations corresponding to the RL model ($\mu_k = \mu \ \forall k \in A$) which is fast to compute.

[1]The value functions can also be used in the stopping criteria i.e. the iteration stops when $\sum_{k \in \tilde{A}} (V^{i+1}(k) - V^i(k))^2 < \gamma'$. The value functions have however larger magnitudes than z.

The proofs in the literature establishing the existence and uniqueness of a solution to Bellman's equation use a discount factor less than one. In our case we do not discount future utilities and these proofs do not apply. Fosgerau et al. (2013) discuss this issue in more detail for the RL model. In essence, the existence of a solution depends on the balance between the number of paths connecting the nodes in the network and the size of the scaled instantaneous utilities. It is easy to find a feasible solution by using large enough magnitude of the β parameters.

Since the value functions depend on the parameter values, they need to be solved repeatedly when searching over the parameter space (maximum likelihood estimation). In order to decrease the computational time we use dynamic accuracy. More precisely, we update the threshold γ in the iterations of the non-linear optimization algorithm so that higher accuracy is required close to optimum (γ decreases as the number of iterations of the non-linear optimization algorithm increases).

Before discussing the maximum likelihood estimation in more detail, we note that (12) can be written as $F(z) = 0$, where $F(z) = z - [M \circ X(z)]e + b$. A standard solver can be used e.g. *fsolver* in MATLAB or the Newton-GMRES method (for instance Kelley, 1995). We have tested these methods but found that they are not efficient for our application and that our approach works better.

5 Maximum likelihood estimation

There are several different ways of estimating a dynamic discrete choice model (Aguirregabiria and Mira, 2010), we adopt the nested fixed point algorithm of Rust (1987). This algorithm combines an outer iterative non-linear optimization algorithm for searching over the parameter space with an inner algorithm for solving the value functions.[2] The latter was the focus of the previous section and we now turn our attention to the definition of the log-likelihood (LL) function and the derivation of its gradient which allows us to use classic Hessian approximation such as BHHH and BFGS (see for instance Berndt et al., 1974, Nocedal and Wright, 2006).

The path probabilities are defined by (9) and contain scale parameters

[2]Another option is the algorithm proposed by Aguirregabiria and Mira (2002). The idea is to swap the order of the outer and inner algorithms so that the outer algorithm solves the value functions and the inner algorithm maximizes the pseudo-likelihood function. This is very useful if the value functions are costly to evaluate. In the case of the NRL model, it is more costly to maximize the log-likelihood function than solving the value functions.

μ_k $\forall k \in A$ as well as the parameters β associated with the attributes of the instantaneous utilities. Clearly, it is not possible to estimate all link-specific scale parameters for a real network and therefore we assume that they are a function of parameters β to be estimated $\mu_k(\beta)$. (We refer the reader to the numerical results, Section 6, for an example.)

The LL function defined over the set of path observations $n = 1, \ldots, N$ is

$$LL(\beta) = \sum_{n=1}^{N} \ln P(\sigma_n, \beta) = \sum_{n=1}^{N} \sum_{t=0}^{I_n} \frac{1}{\mu_{k_t}} \left(v^n(k_{t+1}|k_t) + V^n(k_{t+1}) - V^n(k_t) \right) \tag{14}$$

and is very similar to the LL function of the RL model except that the value functions for the states along a path do not cancel out. Assuming a linear-in-parameters formulation of the instantaneous utilities, the gradient with respect to a given parameter β_i is

$$\frac{\partial LL(\beta)}{\beta_i} = \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{I_n - 1} \frac{1}{\mu_{k_t}} \left(\frac{\partial v^n(k_{t+1}|k_t)}{\partial \beta_i} + \frac{\partial V^n(k_{t+1})}{\partial \beta_i} - \frac{\partial V^n(k_t)}{\partial \beta_i} \right) - \frac{\partial \mu_{k_t}}{\mu_{k_t}^2 \partial \beta_i} \left(v^n(k_{t+1}|k_t) + V^n(k_{t+1}) + V^n(k_t) \right)$$

and hence requires the first derivative of the value functions $V^n(k)$, $\forall k \in \tilde{A}$ with respect to β_i. We define $\phi_{ka} = \mu_a / \mu_k$ and take the derivative of a given value function z_k as defined by (8) (without using the superscript for destination d) and obtain

$$\begin{aligned} \frac{\partial z_k}{\partial \beta_i} &= \sum_{a \in A} \left(\frac{\partial M_{ka}}{\partial \beta_i} z_a^{\phi_{ka}} + M_{ka} z_a^{\phi_{ka}} \left(\frac{\phi_{ka}}{z_a} \frac{\partial z_a}{\partial \beta_i} + \frac{\partial \phi_{ka}}{\partial \beta_i} \ln z_a \right) \right) \\ &= \sum_{a \in A} \left(\frac{\partial M_{ka}}{\partial \beta_i} z_a^{\phi_{ka}} + M_{ka} z_a^{\phi_{ka}} \frac{\partial \phi_{ka}}{\partial \beta_i} \ln z_a \right) + \sum_{a \in A} \left(M_{ka} z_a^{\phi_{ka}} \frac{\phi_{ka}}{z_a} \frac{\partial z_a}{\partial \beta_i} \right). \end{aligned} \tag{15}$$

We note that when the scales μ_k contain some model parameters, the derivative of each element of matrix $M(\beta)$ with respect to a given parameter β_i is

$$\frac{\partial M_{ka}}{\partial \beta_i} = \delta(a|k) e^{\frac{v(a|k)}{\mu_k}} \left(\frac{\partial v(a|k)}{\mu_k \partial \beta_i} - v(a|k) \frac{\partial \mu_k}{\mu_k^2 \partial \beta_i} \right), \quad k, a \in \tilde{A}.$$

We introduce two matrices, G^i and K of size $|\tilde{A}| \times |\tilde{A}|$, which have the two sums of (15) as entries

$$G^i_{ka} = \frac{\partial M_{ka}}{\partial \beta_i} z_a^{\phi_{ka}} + M_{ka} z_a^{\phi_{ka}} \frac{\partial \phi_{ka}}{\partial \beta_i} \ln z_a$$

$$K_{ka} = M_{ka} z_a^{\phi_{ka}} \frac{\phi_{ka}}{z_a}, \ \forall k, a \in \tilde{A}. \tag{16}$$

This allows us to define the Jacobian of vector z as a system of linear equations

$$\frac{\partial z}{\partial \beta_i} = G^i e + K \frac{\partial z}{\partial \beta_i} \Rightarrow \frac{\partial z}{\partial \beta_i} = (I - K)^{-1} G^i e, \tag{17}$$

which in theory, can be solved very efficiently. Nevertheless, it is possible to use the fact that $V(k) = \mu_k \ln z_k \ \forall k \in \tilde{A}$ and derive the Jacobian of V instead of z. In this case the gradient of $V(k)$ with respect to a given β_i is

$$\frac{\partial V(k)}{\partial \beta_i} = \frac{\partial \mu_k}{\partial \beta_i} \ln z_k + \frac{\mu_k}{z_k} \frac{\partial z_k}{\partial \beta_i}. \tag{18}$$

Using (15) we get

$$\frac{\partial V(k)}{\partial \beta_i} = \sum_{a \in A} S^i_{ka} + \sum_{a \in A} H_{ka} \frac{\partial V(a)}{\partial \beta_i} + h_k \tag{19}$$

where

$$S^i_{ka} = \mu_k \frac{\partial M_{ka}}{\partial \beta_i} \frac{z_a^{\phi_{ka}}}{z_k} + \mu_k M_{ka} \ln(z_a) \frac{z_a^{\phi_{ka}}}{z_k} \frac{\partial \phi_{ka}}{\partial \beta_i} - M_{ka} \ln(z_a) \frac{z_a^{\phi_{ka}}}{z_k} \frac{\partial \mu_a}{\partial \beta_i}$$

and

$$H_{ka} = M_{ka} \frac{z_a^{\phi_{ka}}}{z_k} \text{ and } h_k = \frac{\partial \mu_k}{\partial \beta_i} \ln z_k.$$

We denote S^i, H be two matrices of size $|\tilde{A}| \times |\tilde{A}|$ and h, V be two vectors of size $|\tilde{A}| \times 1$ with entries S^i_{ka}, H_{ka}, h_k, $V(k)$ for all $k, a \in \tilde{A}$, respectively. The Jacobian of vector V can then be written as a system of linear equations

$$\frac{\partial V}{\partial \beta_i} = (I - H)^{-1} (S^i e + h). \tag{20}$$

Although theoretically equivalent, we now discuss the numerical differences between the two formulas (17) and (20) for computing the gradient of the value functions. We consider the definitions of the matrix K and H. z_a, $a \in \tilde{A}$ are exponential functions of the value functions which are negative by assumption. The value of z_a may therefore be very close to zero. Since the elements of matrix K can be written as $K_{ka} = \phi_{ka} M_{ka} z_a^{\phi_{ka}-1}$ $(\forall k, a \in \tilde{A})$ if $\phi_{ka} < 1$, the value of K_{ka} can be very large, and if $\phi_{ka} > 1$, K_{ka} can be very close to zero. These wide range of values in the elements of matrix K (and

also in matrix $I - K$) can lead to numerical issues when solving the system (17). Based on equation (8), each element of matrix H can be written as

$$\begin{aligned} H_{ka} &= \frac{M_{ka} z_a^{\phi_{ka}}}{\sum_{a' \in A(k)} M_{ka'} z_{a'}^{\phi_{ka'}}} \\ &= \frac{1}{1 + \sum_{a' \in A(k), a' \neq a} \frac{M_{ka'} z_{a'}^{\phi_{ka'}}}{M_{ka} z_a^{\phi_{ka}}}} \\ &\forall k, a \in \tilde{A}, a \in A(k) \end{aligned}$$

so that $0 < H_{ka} < 1$, meaning that the elements of matrix H are closer in value, compared to matrix K. Therefore, using (20) to compute the gradient of LL function is better than (17) for numerical reasons. In summary, the analytical gradient of the LL function has a complicated form but can be efficiently computed by solving systems of linear equations.

6 Numerical results

In this section we present estimation and prediction results for four different models: the RL model with and without link size (LS) attribute and the NRL model, also with and without LS attribute. We use the same data as Fosgerau et al. (2013) (also used in Frejinger and Bierlaire, 2007, Mai et al., 2014) which has been collected in Borlänge, Sweden. The network is composed of 3077 nodes and 7459 links and is uncongested so travel times can be assumed static and deterministic. The sample consists of 1832 trips corresponding to simple paths with a minimum of five links. Moreover, there are 466 destinations, 1420 different origin-destination (OD) pairs and more than 37,000 link choices in this sample.

6.1 Model specifications

The same five attributes as Fosgerau et al. (2013) are used in the instantaneous utilities. First, link travel time $TT(a)$ of action a. Second, a left turn dummy $LT(a|k)$ that equals one if the turn angle from k to a is larger than 40 degrees and less than 177 degrees. Third, a u-turn dummy $UT(a|k)$ that equals one if the turn angle is larger than 177. Fourth, a link constant $LC(a)$. The fifth attribute is $LS(a)$ (for a detailed description see Fosgerau et al., 2013) and it has been computed using a linear-in-parameters formulation of the aforementioned four attributes using parameters $\tilde{\beta}_{TT} = -2.5$, $\tilde{\beta}_{LT} = -1$, $\tilde{\beta}_{LC} = 0.4$, $\tilde{\beta}_{UT} = -4$.

Even in this fairly small network there are more than 7000 links, so it is not possible to estimate link specific parameters. We therefore impose a constraint on the scale parameters $\mu_k > 0$ by defining them as a function of link attributes. More precisely, $\mu_k = e^{\lambda_k}$ where $\lambda_k = \omega x_k$, ω is a vector of parameters and x_k a vector of attributes associated with link k. This assumption ensures that (i) the estimation problem is unconstrained and (ii) we can use the analytical gradient (18). Note that if all the parameters in λ_k are zero, the scales are equal to one for all links $k \in \tilde{A}$, meaning that the NRL model becomes the RL model. As much as data allows, it is possible to elaborate on the specification of the scale parameters. For example, by including different attributes in the exponential function or by estimating link specific scales parameters for some links in the network.

For the numerical results presented in this paper we use three link specific attributes: travel time, LS and the number of outgoing links $OL(k) = |A(k)|$. Accordingly, λ_k is

$$\lambda_k = \omega_{TT} TT(k) + \omega_{LS} LS(k) + \omega_{OL} OL(k). \tag{21}$$

We do not use a link constant since it has the same value for all links, the rationale behind using it in the instantaneous utilities is to penalize paths with many crossings (links). Note that this is not a regression model, it is simply a specification of the scale parameters μ_k that enter the instantaneous utility functions.

To summarize, the deterministic utilities for four different model specifications with respect to link a given link k are

$$\begin{aligned} v^{\text{RL}}(a|k;\beta) = v^{\text{NRL}}(a|k;\beta) &= \beta_{TT} TT(a) + \beta_{LT} LT(a|k) + \beta_{LC} LC(a) \\ &\quad + \beta_{UT} UT(a|k) \\ v^{\text{RL-LS}}(a|k;\beta) = v^{\text{NRL-LS}}(a|k;\beta) &= \beta_{TT} TT(a) + \beta_{LT} LT(a|k) + \beta_{LC} LC(a) \\ &\quad + \beta_{UT} UT(a|k) + \beta_{LS} LS(a) \end{aligned}$$

and the instantaneous utilities are

$$\begin{aligned} u^{\text{RL}}(a|k;\beta) &= v^{\text{RL}}(a|k;\beta) + \mu\epsilon(a) \\ u^{\text{RL-LS}}(a|k;\beta) &= v^{\text{RL-LS}}(a|k;\beta) + \mu\epsilon(a) \\ u^{\text{NRL}}(a|k;\beta,\omega) &= v^{\text{NRL}}(a|k;\beta) + e^{\lambda_k}\epsilon(a) \\ u^{\text{NRL-LS}}(a|k;\beta,\omega) &= v^{\text{NRL-LS}}(a|k;\beta) + e^{\lambda_k}\epsilon(a). \end{aligned}$$

6.2 Estimation results

We report the estimation results for the four specifications in Table 5. The results are comparable to those previously published using the same data.

The β estimates have their expected signs and are highly significant. $\widehat{\omega}_{LS}$ and $\widehat{\omega}_{OL}$ are significant and negative while $\widehat{\omega}_{TT}$ is not significantly different from zero when the LS attribute is included in the instantaneous utilities. The LS attribute corresponds to expected normalized flows and takes positive values but is numerically close to zero for a majority of the links in the network. $\widehat{\omega}_{LS}$ and $\widehat{\omega}_{OL}$ indicate that the scales are inversely related to flow and number of outgoing links; links with more flow and more outgoing links have smaller variance of the error terms than links with less flow and fewer outgoing links.

It is not straightforward to analyze the resulting scale parameters based on $\widehat{\omega}$. We therefore provide two histograms in Figure 6 showing the distribution of μ_k and $\ln \phi_{ka} = \ln \frac{\mu_a}{\mu_k}$ over the links in the network for the NRL-LS model. The graph on the left shows that the values of μ_k vary over the links in the network which ensures that IIA does not hold (the average value of μ_k is 0.78). The peaks in the distribution are due to the attribute number of outgoing links $OL(k)$ which take discrete values. We note that a few links have values larger than one: this is consistent with utility maximization and does not imply counter intuitive path probabilities. The graph on the right shows the distribution of $\ln \phi_{ka}$ which is quite symmetric around 0 (the average value of ϕ_{ka} is 1.03). The symmetry can be explained by the attribute $OL(k)$. Consider the u-turn link a' of link $a \in A(k)$. Since link k and a' have the same sink node we have $OL(k) = OL(a')$. For our data this results in values of μ_k numerically close to $\mu_{a'}$ and thus

$$\phi_{ka}\phi_{aa'} = \frac{\mu_a}{\mu_k}\frac{\mu_{a'}}{\mu_a} \approx 1$$

or equivalently

$$\ln \phi_{ka} + \ln \phi_{aa'} \approx 0.$$

The LS attribute was designed to correct the utilities of overlapping paths in a way similar to the path size attribute. Moreover, if the values of these attributes are updated in case of a change in any attribute in the network, they relax the IIA property. Several studies in the literature (e.g. Bekhor et al., 2001, Frejinger and Bierlaire, 2007) report a better model fit and prediction results if these attributes are included in the deterministic utilities in addition to correlated random terms. It is also the case in this study: we observe a significant improvement in final log-likelihood values when we add the LS attribute (the likelihood ratio test are reported in Table 6, when applicable). The best model in terms of in-sample fit is NRL-LS. Since the scale parameters and the link size parameter are estimated off the same variation in the data, it is important to note that an identification issue may occur. It is however not the case for this data set.

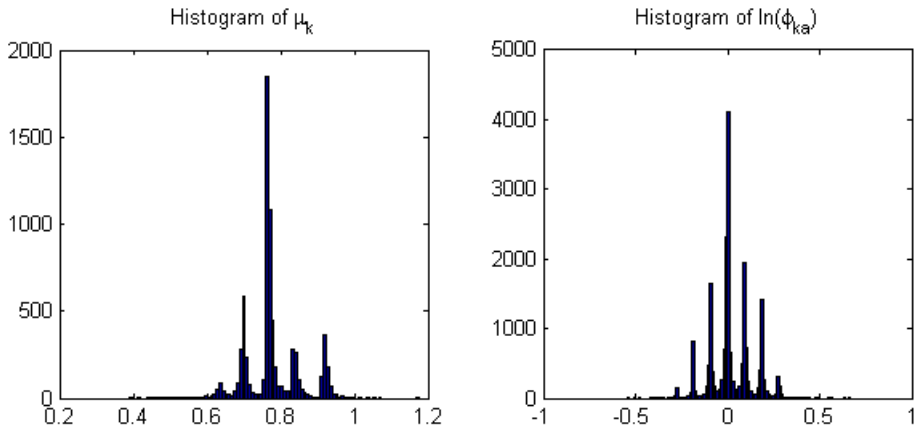

Figure 6: Histogram of μ_k and $\ln \phi_{ka}$ for NRL-LS

Before comparing prediction results in the following section we make some remarks concerning the estimation. We use a basic trust region algorithm with the BHHH method for approximating the Hessian and the code is implemented in MATLAB (and available upon request). We use the iterative method with dynamic accuracy for the computation of the value functions (see Section 4). We note that if we use an initial vector as a solution of the system of linear equations, about 100 iterations is enough for a high precision ($\gamma' = 10^{-8}$) but we need about 200 iterations for the same precision when the initial vector is the unit vector (all the elements are equal to one). Moreover, using only 50 iterations in the beginning of the optimization (corresponding to a precision $\gamma' \in [1, 10]$) and switching to the high precision $\gamma' = 10^{-8}$ when the norm of the gradient of the log-likelihood function is less than 10^{-3} we were able to double the speed of the estimation.

6.3 Prediction results

In this section we focus on comparing the prediction performance of the different models. We use a cross validation approach where the sample of observations is divided into two sets by drawing observations at random with a fixed probability: one set is used for estimation (80% of the observations) and the other is used as holdout (20% of the observations) to evaluate the predicted probabilities by applying the estimated model. We generate 40 holdout samples of the same size by reshuffling the real sample. The log-likelihood loss is used as the loss function to evaluate the prediction performance. More precisely, for each holdout sample i, $0 \leq i \leq 40$ we estimate the parameters $\hat{\beta}_i$ off the corresponding training sample and this vector of

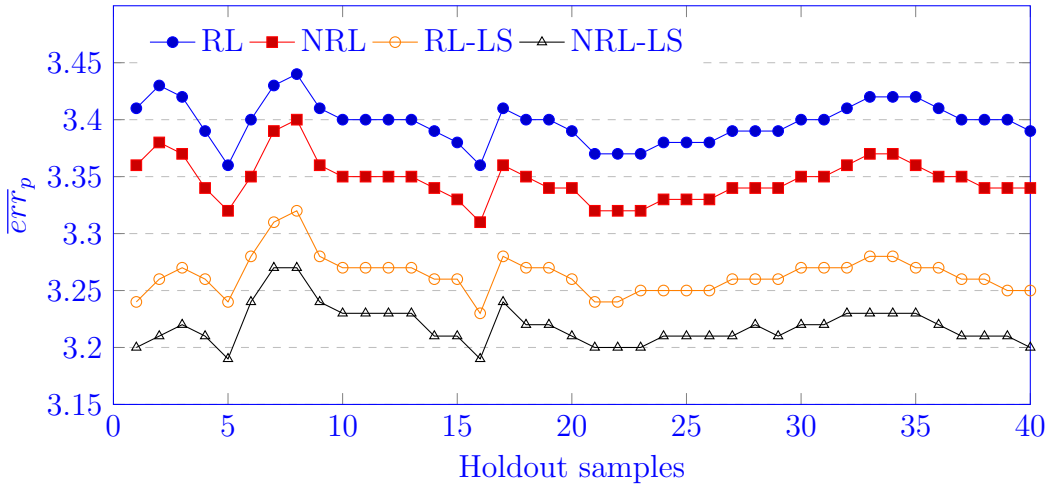

Figure 7: Average of the test error values over holdout samples

parameters is used to compute the test errors err_i

$$err_i = -\frac{1}{|PS_i|} \sum_{\sigma_j \in PS_i} \ln P(\sigma_j, \hat{\beta}_i)$$

where PS_i is the size of prediction sample i. Then err_i is a random variable that depends on the holdout sample i. In order to have unconditional test error values we compute the average of err_i values over samples as follows

$$\overline{err}_p = \frac{1}{p} \sum_{i=1}^{p} err_i \quad \forall 1 \leq p \leq 40. \tag{22}$$

The values of $\overline{err}_p$, $1 \leq p \leq 40$ are plotted in Figure 7 and Table 7 reports the average of the test error values over 40 samples given by the RL, RL-LS, NRL, NRL-LS models. For each model the value of $\overline{err}_p$ becomes more stable as p increases. The prediction results show that models including the LS attribute perform better than those without. The NRL-LS model has a significantly better fit and also a better prediction performance than RL-LS. We note that the differences between the models' test error values are quite constant over the holdout samples. This is due to the fact that (i) the same holdout sample is used across models, and (ii) the number of observations used for estimation and the size of the holdout samples are large, so the parameter estimates are stable and so are the predicted log-likelihood values.

7 Conclusion

This paper has presented the NRL model which avoids the IIA property of the RL model by allowing scale parameters to be link specific while keeping the advantages of the RL model. We have proposed an efficient approach to estimate the model, solving the value functions using an iterative method with dynamic accuracy. Moreover, we have derived the gradient of the log-likelihood function which can be computed by solving systems of linear equations.

We have provided numerical results using real data. The parameter estimates are sensible and the NRL model has significantly better fit than the RL model. The LS attribute plays an important role and the best models including this attribute have significantly better model fit than those without. We have also provided a cross-validation study suggesting that NRL-LS and NRL are better than the RL-LS and RL model, respectively.

In future research we plan to investigate further the importance of the LS attribute and its definition. Moreover, there are only few attributes available in the data set used in this paper. We would like to test the model on other data sets that allows us to test other possible functional forms of the scale parameters.

In this paper we use a unimodal network and observations of trips made by car. We emphasize that the model is not restricted to this type of network. More precisely, by adapting the state space, the model can be used in e.g. dynamic networks (state is time and location) and multi-modal networks (state is location and mode) as long as link attributes are deterministic. The dynamic network is suitable for modeling congested networks, the RL model has been used for this purpose by Ramos et al. (2012). The challenge lies in the size of the state space, which is considerably larger than a static network since it is the number of links multiplied by the number of time intervals.

As a final remark we note that since the RL and NRL models are based on the universal choice set (including all path even those with loops), they avoid having to consider choice set formation. They can therefore be seen as alternatives to the approach proposed by Manski (1977). The RL and NRL may be relevant to other contexts than route choice where there is an issue associated with large choice sets.

Acknowledgement

This research was partly funded by the National Sciences and Engineering Research Council of Canada, discovery grant 435678-2013.

Parameters	RL	NRL	RL-LS	NRL-LS
$\widehat{\beta}_{TT}$	-2.494	-1.854	-3.060	-2.139
Rob. Std. Err.	0.098	0.132	0.103	0.145
Rob. t-test(0)	-25.45	-14.05	- 27.709	-14.75
$\widehat{\beta}_{LT}$	-0.933	-0.679	-1.057	-0.748
Rob. Std. Err.	0.030	0.043	0.029	0.047
Rob. t-test(0)	-31.10	-15.79	-36.448	-15.91
$\widehat{\beta}_{LC}$	-0.411	-0.258	-0.353	-0.224
Rob. Std. Err.	0.013	0.016	0.011	0.015
Rob. t-test(0)	-31.62	-16.13	-32.091	-14.93
$\widehat{\beta}_{UT}$	-4.459	-3.340	-4.531	-3.301
Rob. Std. Err.	0.114	0.200	0.126	0.207
Rob. t-test(0)	-39.11	-16.7	- 35.960	-15.95
$\widehat{\beta}_{LS}$	-	-	-0.227	-0.155
Rob. Std. Err.	-	-	0.013	0.013
Rob. t-test(0)	-	-	-17.462	-11.92
$\widehat{\omega}_{TT}$	-	0.515	-	0.341
Rob. Std. Err.	-	0.255	-	0.288
Rob. t-test(0)	-	2.02	-	1.18
$\widehat{\omega}_{LS}$	-	-0.674	-	-0.581
Rob. Std. Err.	-	0.093	-	0.090
Rob. t-test(0)	-	-7.25	-	-6.46
$\widehat{\omega}_{OL}$	-	-0.086	-	-0.092
Rob. Std. Err.	-	0.015	-	0.016
Rob. t-test(0)	-	-5.73	-	-5.75
$LL(\widehat{\beta})$	-6303.9	-6187.9	-6045.6	-5952.0

Table 5: Estimation results

Models	χ^2	p-value
RL & NRL	232	5.11e-50
RL-LS & NRL-LS	187.2	2.46e-40
NRL & NRL-LS	471.8	1.30e-104

Table 6: Likelihood ratio test results

RL	NRL	RL-LS	NRL-LS
3.392	3.336	3.252	3.204

Table 7: Average of test error values over 40 holdout samples

References

Aguirregabiria, V. and Mira, P. Swapping the nested fixed point algorithm: A class of estimators for discrete markov decision models. *Econometrica*, 70(4):1519–1543, 2002.

Aguirregabiria, V. and Mira, P. Dynamic discrete choice structural models: A survey. *Journal of Econometrics*, 156(1):38–67, 2010.

Akamatsu, T. Cyclic flows, markov process and stochastic traffic assignment. *Transportation Research Part B*, 30(5):369–386, 1996.

Bekhor, S., Ben-Akiva, M., and Ramming, M. Estimating route choice models for large urban networks. 9th World Conference on Transport Research, Seoul, Korea, 2001.

Ben-Akiva, M. and Bierlaire, M. Discrete choice methods and their applications to short-term travel decisions. In Hall, R., editor, *Handbook of Transportation Science*, pages 5–34. Kluwer, 1999.

Ben-Akiva, M. *The structure of travel demand models*. PhD thesis, MIT, 1973.

Berndt, E. K., Hall, B. H., Hall, R. E., and Hausman, J. A. Estimation and inference in nonlinear structural models. *Annals of Economic and Social Measurement*, 3/4:653–665, 1974.

Chu, C. A paired combinatorial logit model for travel demand analysis. In *Proceedings of the fifth World Conference on Transportation Research*, volume 4, pages 295–309, Ventura, CA, 1989.

Fosgerau, M., Frejinger, E., and Karlström, A. A link based network route choice model with unrestricted choice set. *Transportation Research Part B*, 56(1):70–80, 2013.

Frejinger, E., Bierlaire, M., and Ben-Akiva, M. Sampling of alternatives for route choice modeling. *Transportation Research Part B*, 43(10):984–994, 2009.

Frejinger, E. and Bierlaire, M. Capturing correlation with subnetworks in route choice models. *Transportation Research Part B*, 41(3):363–378, 2007.

Guevara, C. A. and Ben-Akiva, M. E. Sampling of alternatives in multivariate extreme value (MEV) models. *Transportation Research Part B*, 48(1):31 – 52, 2013a. ISSN 0191-2615.

Guevara, C. A. and Ben-Akiva, M. E. Sampling of alternatives in logit mixture models. *Transportation Research Part B: Methodological*, 58(1): 185 – 198, 2013b. ISSN 0191-2615.

Kelley, C. T. *Iterative Methods for Linear and Nonlinear Equations*. Society for Industrial and Applied Mathematics, 1995.

Lai, X. and Bierlaire, M. Specification of the cross nested logit model with sampling of alternatives for route choice models. *Technical report, TRANSP-OR, EPFL*, 2014.

Mai, T., Frejinger, E., and Bastin, F. A misspecification test for logit based route choice models. *Technical report, CIRRELT - 32*, 2014.

Manski, C. F. The structure of random utility models. *Theory and decision*, 8(3):229–254, 1977.

McFadden, D. Modelling the choice of residential location. In Karlqvist, A., Lundqvist, L., Snickars, F., and Weibull, J., editors, *Spatial Interaction Theory and Residential Location*, pages 75–96. North-Holland, Amsterdam, 1978.

Nocedal, J. and Wright, S. J. *Numerical Optimization*. Springer, New York, NY, USA, 2nd edition, 2006.

Prato, C. G. Route choice modeling: past, present and future research directions. *Journal of Choice Modelling*, 2:65–100, 2009.

Ramos, G., Frejinger, E., Daamen, W., and Hoogendoorn, S. Route choice model estimation in dynamic network based on GPS data. In *Presented at the 1st European Symposium on Quantitative Methods in Transportation Systems*, Lausanne, Switzerland, 2012.

Rust, J. Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. *Econometrica*, 55(5):999–1033, 1987.

Rust, J. Maximum likelihood estimation of discrete control processes. *SIAM Journal on Control and Optimization*, 26(5):1006–1024, 1988.

Train, K. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2003.

Vovsha, P. and Bekhor, S. Link-nested logit model of route choice Overcoming route overlapping problem. *Transportation Research Record*, 1645: 133–142, 1998.