Cross-nested logit models

Estimating complex network correlations

Thor Donsby Noe * Kristian Urup Larsen†

May 31, 2018

Abstract

In many cases the Multinomial Logit model produces very biased results as the strict assumption of Independence of Irrelevant Alternatives is violated. This paper shows how to solve this by investigating how the Generalized Extreme Value models and in particular the Cross-nested Logit model can be useful for modeling discrete choices in economics and other fields. We present and extend the theoretical background required to understand substitution patterns and other inner workings of the models and derive marginal effects, identification restrictions, and other useful measures.

GEV models like regular Logit models have closed form likelihood functions, and as such lend themselves to regular maximum likelihood estimation. However, as model complexity grows they require increasingly optimized code to be feasibly evaluated and optimized. Likewise the number of constraints and identification issues rise rapidly with complexity, implying that specialized tools are the only time efficient way to estimate parameters in GEV models.

We simulate data from a GEV process and discuss issues of computation and identification on the basis of these data. We then apply a simple CNL model on data from the Danish DREAM database to test its usefulness in applied econometrics and compare various specifications of GEV models.

Keystrokes: 71.886 **Standard pages:** 30. **Contributions:** Thor Donsby Noe: 3.1, 3.2, 4.3, 4.4, 5.1, 5.2, 6.3, 7.1 Kristian Urup Larsen: 3.3, 3.4, 4.1, 4.2,5.3, 6.1, 6.2, 8.1, 8.2

*Department of Economics, University of Copenhagen, Øster Farimagsgade 5, DK-1353 Copenhagen K, Denmark

†Department of Economics, University of Copenhagen, Øster Farimagsgade 5, DK-1353 Copenhagen K, Denmark (e-mail: kuol@econ.ku.dk)

Contents

List of Figures

List of Tables

1. Introduction

In this paper we study the Generalized Extreme Value (GEV) model with cross-nesting, simply known as the Cross-nested Logit (CNL) model, as a useful solution to the fact that the Multinomial Logit (MNL) model produces simplified and biased results for the many cases where it is unrealistic to assume Independence of Irrelevant Alternatives (IIA). The GEV framework is a flexible framework encompassing a broad range of discrete choice models which are all consistent with the theory of Random Utility Models (RUM's) and have closed form likelihood functions. So far GEV models, including the CNL model have received little attention by economists, perhaps partly because work to understand the models in themselves is still ongoing. In addition to the lack of a full understanding of the workings of GEV models, it requires a significant investment of time to become proficient in the use of either the available software tools able to estimate CNL/GEV models. Despite these drawbacks GEV models seem useful in cases where it is known that some structural effects are important in decision processes, but the exact form of these structural traits is unknown.

To understand substitution patterns and other dynamics behind RUM choice modelling we first provide a rundown of the theory from the simple MNL model to the GEV framework. We then proceed to describe our somewhat successful implementation of the CNL model and discuss challenges in this regard.

Finally we use prebuild software to estimate a range of GEV models on unemployment data from the Danish DREAM database. Here we find evidence of nesting, such that individuals nest together options that involve unemployment benefits in one nest, and ordinary employment in another nest. The driving force behind this nesting is still to be determined.

2. Literature

The literature on Cross-nested logit models (CNL) is so far highly theoretical, compared to the way econometrics often highlight the practical applications of new techniques. This is in part due to the complexity of the models, and in part because many questions on for example general identifiability are still to be addressed for the CNL. So far the CNL has found its uses primarily in traffic research and travel mode theory, where it helps researchers study individuals' choices without the restrictive limitations of the simpler choice models. Compared to econometrics these fields appear to have a stronger tradition for modelling complex systems, instead of relying on natural experiments.

A large part of the existing literature on the CNL model is due to Michel Bierlaire (Bierlaire, 2001; Bierlaire, 2003; Bierlaire, 2006; Bierlaire, 2008; Bierlaire et al., 2009) who

has contributed both by unifying the various formulations of the CNL that have been proposed (Bierlaire, 2006) and by developing the software package *Biogeme* (Bierlaire, 2003) for estimation of discrete choice models. Other work on the CNL model specifically include Papola (2004) but in general the literature due to others than Bierlaire has been superseded by his attempts to unify the work of many others.

Bierlaire builds on the basis of the extensive work in which Daniel McFadden has introduced the *Generalized Extreme Value* (GEV) model class (McFadden, 1977a; McFadden, 1977b; Hausman and McFadden, 1984).

There is some literature on the estimation of complex network models, e.g. by Newman et al. (2018) and Mai et al. (2017) however we do not touch specifically on this in our paper as the techniques involved in efficient estimation of these models are in general to advanced to mention here.

2.1. Software

In terms of software there are essentially two options that are available. A couple of other alternatives are out there (even STATA can estimate nested models) but these are generally expensive and/or unable to estimate the generalized versions such as the CNL model. The oldest of the free software packages is Biogeme (Bierlaire, 2003) which has now been extended to include the more user friendly PythonBiogeme for Python 3. (Bierlaire, 2016). A newer alternative, also made for Python 3.6 is Larch, presented in Newman et al. (2018). While Biogeme relies on separate model files for specification, Larch handles everything from within a single Python session, making it preferable for quick estimation. Furthermore Larch is on PyPi (Python Package Index) and has a significant speed advantage over Biogeme.

3. Theory

In this section we first take a look at a simple multinomial choice model and its assumption about equal competition among all pairs of alternatives, after which we relax this assumption and look at two different choice models that allows for cases where it is expected that some pairs of alternatives compete more closely than others.

By using a recurrent analogy of a commuter's classic choice of transport we throughout complement the theory with a more intuitive examination of the growingly complex substitution patterns for extending the choice structure from the *Multinomial Logit* (MNL) over the *Nested Logit* (NL) to the *Cross-nested Logit* (CNL) model.

3.1. Random Utility Models (RUM's)

Common for all choice models related to the cross-nested logit is that they build on Random Utility Models (RUM's) which forecast discrete choices based on the assumption that all individuals q seek to maximize utility from the choices they make i. While classic utility and consumer theory has been mostly limited to scale predictions of continuous character (e.g. for prices and quantities) up until the formalization of the probabilistic approach by R. Duncan Luce (1957; 1958) which is necessary for discrete choices in general and for multinomial choices especially. Here the "randomness" of the RUM term comes into play as we are aware that we cannot perfectly predict the utility of a choice for each individual, thus, we include a stochastic random error term (McFadden, 2005), ε_{iq}, for each individual q and choice i to correct for the fact that we as analysts lack information and understanding of the utility for each individual. Formally the utility U_{iq} is composed by a deterministic component V_{iq} and noise ε_{iq} so that

$$U_{iq} = V_{iq} + \varepsilon_{iq} \tag{3.1}$$

For now let the subscripts denote choice i and individual q, but note that the individual subscript will soon be taken as given and thus dropped to not get lost in the notation when we add a nest-subscript.

The fundamental assumption of utility-maximization makes RUM's applicable for analyzing a lot of different human-choice-problems and advanced RUM's are heavily used in e.g. traffic research and route choice problems. Because of the similarities between the econometric models and economic theory the random utility models are also used extensively in microeconomics. It can be noted though, that the general necessity of representing preferences by utility functions of observed quantities in economic theory does impose a limit on how the field can develop as preferences that is somehow revealed can be analyzed (Richter, 1966).

A first step in converting the idea of random utility into actual econometrics is to assume some analytical and data driven description of V_{iq}. Although there are alternatives we will keep to the by far most common formulation in this paper, namely that V_{iq} is linear in parameters, such that the deterministic component of the utility V_{iq} is given by

$$V_{iq} = \beta_i^0 + \sum_{k=1}^{K} \beta_{ik} x_{iqk}$$

Note though, that whether a constant β_i^0 is included and for that matter if it is allowed to vary between alternatives (subscript i) is dependent on the model specification. Determining the minimal necessary restrictions on parameters that is required for identification is not trivial in the nested and cross-nested models. To ease the readability we

through this paper takes the individual's q subscripts for given and drops it while we instead of a sum over k parameters and regressors write the second part of the equation vector form, thus, also drop the k subscript and write the deterministic utility as

$$V_i = \beta_i^0 + \beta_i x_i \tag{3.2}$$

Where x_i can simply be x if the regressors are alternative-invariant (see the end of section 3.2.1). The linear deterministic model (3.2) as well as any non-linear form of V_i is unable to fully capture all of the variables that influences the individual's utility gain for a choice. Therefore, a degree of *randomness* is introduced by letting the individual utility U_{iq} for the occurence that the individual q chooses alternative i be given by equation (3.1). Thus, in addition to the deterministic term V_{iq} the stochastic distribution of the random error terms ε_i's in the complete choice set $\mathcal{C}$ have to be assumed. McFadden (1977a) shows that assuming a joint error distribution

$$F_{\epsilon_1,...,\epsilon_J}(y_1, ...y_J) = \exp\big(-G(e^{-y_1}, ..., e^{-y_J})\big), \quad 1, ..., J = \mathcal{C} \tag{3.3}$$

and implementing suitable requirements on the generating function G gives access to a broad class of models which are consistent with the random utility specification.

3.2. Simple discrete choice structures

The discrete choice models taken into account in this paper model the probability that an individual q given a set of k characteristics x_{qk} chooses an alternative i over each of the remaining alternatives in the choice set $\mathcal{C}$.

Discrete choice models are relevant for analyzing effects of policy that would aim at altering the probability that an individual q chooses a specific alternative i by e.g. affecting the determining variables x_{qk}, which options that are in the choice set $\mathcal{C}$, or directly affecting the attractiveness of an alternative i or $l \neq i$. Thus, it is crucial in what way the probability of a choice is estimated in relation to the probability of choosing the remaining alternatives, i.e. the assumptions about the structure of the choice set.

3.2.1. The Multinomial Logit model (MNL)

The Multinomial Logit model (MNL) was introduced by Daniel McFadden (1973) and is the most widely used econometric model for discrete choices between more than two alternatives. It builds on the RUM (3.2) and by letting the random error term in equation (3.1) be independently and identically Gumbel distributed across alternatives (Koppelman and Sethi, 2000) we get the MNL model. Letting $\mathcal{J}$ be a stochastic variable with support on the choice set $\mathcal{C}$ of an utility maximizing agent, and i the realization of

$\mathcal{J}$ we get the probability that alternative i is choosen by individual q

$$\Pr_q(\mathcal{J} = i|\mathcal{C}) = \frac{e^{V_{iq}}}{\sum_{j \in \mathcal{C}} e^{V_{jq}}}$$

Taking the q subscripts as given and thus dropping them we correspondingly get the notation below that we will use onwards, or simply written $\Pr(i|\mathcal{C})$ which will be used interchangeably.

$$\Pr(\mathcal{J} = i|\mathcal{C}) = \frac{e^{V_i}}{\sum_{j \in \mathcal{C}} e^{V_j}} \tag{3.4}$$

The ratio of probabilities is used to compare the probability of a choice i relative to the probability of choosing some base alternative l

$$\frac{\Pr_q(\mathcal{J} = i|\mathcal{C})}{\Pr_q(\mathcal{J} = l|\mathcal{C})} = \frac{e^{V_i} / \sum_{j \in \mathcal{C}} e^{V_j}}{e^{V_l} / \sum_{j \in \mathcal{C}} e^{V_j}} = \frac{e^{V_i}}{e^{V_l}} \tag{3.5}$$

Which is identical to that of the Binary Logit Model (Cameron and Trivedi, 2005).

As the probability functions only incorporate the deterministic component of the RUM function (3.1), V_{iq}, we should think of the probabilities $\forall i \in \mathcal{C} : \Pr_q(i|\mathcal{C})$ that an individual q with a certain set of characteristics x_q chooses each of the alternatives i in the choice set $\mathcal{C}$, such that these probabilities effectively *"reflect the population probabilities that people with the given set of characteristics and facing the same set of alternatives choose each of the alternatives"* (Koppelman and Bhat, 2006).

3.2.2. Independence of Irrelevant Alternatives (IIA)

McFadden (1973) purposedly constructed the MNL model in order to fulfill the axiomatic idea that *"the relative odds of one alternative being chosen over a second should be independent of the presence or absence of unchosen third alternatives"* which is clearly seen by equation 3.5 above. This axiom was first introduced as the *Independence of Irrelevant Alternatives* (IIA) by Kenneth J. Arrow, 1950 for a particular choice context. The axiom of IIA was popularized in 1959 by R. Duncan Luce (2nd ed. 2005) while he found it more precise to rename it *Independence from Irrelevant Anternatives*" to avoid the misinterpretation that two irrelevant alternatives should be independent of one another while it is the ratio of probabilities between a pair of alternatives that should be independent from including or excluding irrelevant alternatives. A slightly more intuitive way to put it is that the IIA-axiom is the assumption that there is equal competition between all pairs of alternatives, thus no pair of alternatives compete more closely than other pairs (Koppelman and Bhat, 2006).

While this assumption turns out to be relevant for many cases, the IIA assumption

is very strict and MNL estimates will be biased and give incorrect predictions when the assumption is not true, i.e. when the inclusion of other alternatives is indeed not irrelevant.

In appendix B we thoroughly show the implications of assuming *Independence of Irrelevant Alternatives* (IIA) by a commuting example.

3.3. The Generalized Extreme Value model class (GEV)

A generalized framework for models similar to the logit is the class of Generalized Extreme Value (GEV) models. The main benefit of this formulation is that it allows a general pattern for choice probabilities while providing closed form solutions for the probability of choosing each alternative (McFadden, 1977a). The core of the class is a generating function G:

$$G(y_1, y_2, ..., y_J) \ \in \ \mathbb{R}^J_+ \tag{3.6}$$

belonging to a class $\mathcal{G}$ of functions that has the following properties:

1. $G \in \mathcal{G}$ is nonnegative, differentiable and homogeneous of degree $\mu > 0$.
2. $\forall j \in \mathcal{C} : \lim_{y_j \to \infty} G(\cdots) = \infty, \quad \mathcal{C} = 1, ..., J$
3. The lth partial derivative of G: $\frac{\partial^l G}{\prod_i \partial y_i}$ is nonnegative if l is odd, and non-positive if l is even. Here J is the number of alternatives, and each y_j a non-negative variable associated with choice j.

McFadden 1977 shows that this specification implies that the GEV models are consistent with utlity maximization in the sense described in the RUM section above, and derives the probability

$$\Pr(\mathcal{J} = i|\mathcal{C}) = \frac{e^{V_i} \frac{\partial G}{\partial y_i}(e^{V_1}, e^{V_2}, ..., e^{V_J})}{\mu G(e^{V_1}, e^{V_2}, ..., e^{V_J})}, \quad i \in \mathcal{C} = 1, ..., J \tag{3.7}$$

Where V_j is the deterministic element of utility which is observed by the researchers and/or parametrized by some known function e.g. given by equation (3.2) or any non-linear functional form. μ is the degree of homogeneity of $G(\cdot)$. How $G(\cdots)$ is then defined gives rise to a variety of models, including the regular *Multinomial Logit* model when $G(y_1, ..., y_J) = \sum_{j=1}^{J} y_j$. However all functions $G \in \mathcal{G}$ are valid, and generalized specifications lead to the *Nested-* or *Cross-nested* logit models among others.

A first thing to notice is that using Eulers theorem of homogeneous functions[1] on

[1] Let $f(\boldsymbol{z})$ be a homogeneous function of degree q such that $f(t\boldsymbol{z}) = t^q \cdot f(\boldsymbol{z})$. $\boldsymbol{z}$ is a vector of i variables denoted z_i. Eulers theorem then simply states that $\sum_i z_i f'_{z_i}(\boldsymbol{z}) = q f(\boldsymbol{z})$

G we have that

$$\mu G(e^{V_1}, e^{V_2}, ..., e^{V_J}) = \sum_{j=1}^{J} e^{V_j} \frac{\partial G}{\partial e^{V_j}}(e^{V_1}, e^{V_2}, ..., e^{V_J}) \tag{3.8}$$

By redefining $z_i = e^{V_i}$ the probability in (3.7) can be re-expressed as

$$\Pr(\mathcal{J} = i|\mathcal{C}) = \frac{z_i \frac{\partial G}{\partial z_i}}{\sum_{j \in \mathcal{C}} z_j \frac{\partial G}{\partial z_j}} \tag{3.9}$$

This expression should remind one of the equivalent expression encountered when deriving the ordinary logit model, and it will in turn be clear that the MNL (3.4) can be expressed as a GEV such that $\partial G/\partial z_i = 1$. Notice further that $z_i \frac{\partial G}{\partial z_i} = e^{V_i} \frac{\partial G}{\partial z_i} = e^{\ln\left(e^{V_i} \frac{\partial G}{\partial z_i}\right)} = e^{V_i + \ln \frac{\partial G}{\partial z_i}}$ why we can also write (3.9) as

$$\Pr(\mathcal{J} = i|\mathcal{C}) = \frac{e^{V_i + \ln \frac{\partial G}{\partial z_i}}}{\sum_{j \in \mathcal{C}} e^{V_j + \ln \frac{\partial G}{\partial z_j}}} \tag{3.10}$$

Like above this expression is immediately similar to the one known from the MNL model when setting the partial derivative og G equal to 1. (Bierlaire, 2006)

3.3.1. IIA and substitution patterns under cross nesting

For nested structures we should first note that whether *Independence of Irrelevant Alternatives* (IIA) holds in general for a pair of alternatives (i, k) is not only related to the two alternatives, as an addition to the choice set $\mathcal{C}$ might preserve the ratio of probabilities between (i, l) if introduced at certain places in the tree while not at other places.

First lets handle IIA in a fixed choiceset $\mathcal{C}$, e.g. in the set shown in figure C.4. In any nested structure we can express the probability of ending in a given choice c_1 as a product of probabilities $\Pr(i|m)$ for the steps needed to reach c_1 - in the NL model this sequence of steps is unique. So we have that

$$\Pr(\text{ending in } i) = \prod_{\{s\}_j^i} P(s|m_s) \tag{3.11}$$

where$\{s\}_j^i$ is the unique sequence of steps leading from the root to i and m_s is the nest containing choice s for each step in the sequence. In a two-level case this could for example be $\Pr(\textit{public transport}|\textit{root}) \cdot \Pr(\textit{train}|\textit{public transport})$. In the nested logit the probability $\Pr(i|m)$ is given by $\frac{e^{V_i \mu_m}}{\sum_{j \in \mathcal{C}_m} e^{V_j \mu_m}}$ as shown in the section on estimation.

IIA is a property by which the fraction of two such products for the probability of ending in i, k does not depend on anything but V_i and V_k, that is we have IIA if each of these fractions cancel out, except for the ones containing $e^{V_i \mu_m}$ and $e^{V_k \mu_m}$ in the numer-

ator. This is only the case if the two alternative i, k are in the same nest (and in CNL if these nests are only reachable from one path). To see this write

$$\frac{\Pr(\text{ending in } i)}{\Pr(\text{ending in } k)} = \frac{\prod_{\{s\}_j^i} P(s|m_s)}{\prod_{\{s\}_j^k} P(s|m_s)} = \frac{\prod_{\{s\}_j^i \notin \{s\}_j^k} P(s|m_s)}{\prod_{\{s\}_j^k \notin \{s\}_j^i} P(s|m_s)} \tag{3.12}$$

Which subject to i, k sharing the same childless nest m collapses to

$$\frac{\Pr(\text{ending in } i)}{\Pr(\text{ending in } k)} = \frac{e^{V_i \mu_m}}{e^{V_k \mu_m}} \tag{3.13}$$

but would otherwise contain the utilities of many other choices. If the alternatives i, k are in different nests, the number of elements in the sequence $\{s\}_j^i$ might be different from the number of elements in $\{s\}_j^k$. In this case it is relevant to ask which of the denominators will cancel out, as this dictates where in the tree changes will break the IIA relation between two choices. By writing out a chain of the probabilities $e_i^V \mu_m / \sum_{j \in \mathcal{C}_m} e^{V_j \mu_m}$ it is clear that cancelling out happens whenever the set $\mathcal{C}_m$ is shared between the two probability chains. In other words two choices, accessible through paths $\{s\}_j^i$ and $\{s\}_j^k$ will maintain IIA when changes are made to the choice-set in any nests, which are visited as a part of both $\{s\}_j^i$ and $\{s\}_j^k$.

From this we also learn that in the CNL model there is only IIA when choices are not connected to multiple nests, that is there is only IIA when there is no cross-nesting. This is because there will be multiple paths to a given node under cross-nesting, meaning

$$\Pr^{CNL}(\text{ending in } i) = \sum_{\text{possible paths to } i} \prod_{\{s\}_j^i} P(s|m_s) \tag{3.14}$$

why nothing cancels out in the relative probabilities. From this we can distil that

1. A pair of elemental nodes (i, l) within the same nest is IIA as their ratio of probabilities will be independent from any existence or modification of other alternatives. A pair of alternatives (i, l) belonging to *different* nests is on the other hand not IIA in general as the ratio of probabilities can depend on the alternatives in their respective nests.

2. For any pair of alternatives i, l their ratio of probabilities is independent from all nodes n that are *at, next to, or prior to* the lowest structural node from which both i and l can be reached as well as from all nodes in branches of nodes n that do not reach i or l.

3. For a pair of alternatives i, l within the same nest m where at least one of them is a *structural node* their ratio of probabilities is independent from other alternatives within that nest, though, is not independent to any alternatives belonging to any branch following i or l.

Property **1.** is a common result for the NL model (Train, 2009). In the CNL model cross-nesting adds the following complexities. In general property 1. and 3. holds only for alternatives i, l that are *not* cross-nested, i.e. $\forall n \in \mathcal{M} : \ \alpha_{n,i}, \alpha_{n,l} \in 0, 1$. Property 3. is violated if the branches following structural nodes i or l contains a node that is crossed to a nest not in the branches following i or l.

Appendix C presents a variation of the proof, and concretize the application through examples.

3.3.2. A first generalization - the Nested Logit model (NL)

Before jumping to the Cross-nested Logit model (CNL) it is worth spending some time studying the simpler Nested Logit model (NL) that was introduced by H.C.W.L. Williams (1977). This is the most commonly used relaxation of the *Multinomial Logit* MNL model (Koppelman and Bhat, 2006), and thus relaxation of the assumption of *Independence of Irrelevant Alternatives* (IIA).

We should separate a pair of alternatives by allocating them into different nests when we imagine the IIA-axiom can be violated, while nesting pairs together for which we assume that the IIA-axiom holds.

A strategy for proposing a nesting structure is to pairwise consider whether each possible pair of alternatives in a choice set $\mathcal{C}$ share unobserved attributes. If this is the case for none of the pairs we assume IIA and should use the MNL. Otherwise if we assume a pair to compete more closely than others (e.g. *bus* and *train* to compete more closely than *drive alone* as shown in the NL-tree for $\mathcal{C}$ in figure 1 below) then the IIA-assumption is violated and we need to create a *public transport* nest for these two alternatives to part them from *drive alone*. Otherwise our model is misspecified as it within the same nest requires independence of errors terms. We consider nests beside the root as alternatives in themselves. *Public transport* is a nest which partly has some utility in itself (e.g. the attributes that are shared by the different means of public transport) and partly depends on the expected utility of subsequently choosing either *bus* or *train* in order to maximize one's utility (Koppelman and Bhat, 2006).

It is notable that the NL structures for the choice sets $\mathcal{C}'$ and $\mathcal{C}''$in figure 1 are only one out of 13 possible two-level and 12 possible 3-level nesting structures respectively. While some are more plausible than others, we as modelers have 25 different nesting structures to choose from for four alternatives. For five and six alternatives the number of possible nesting structures goes up to no less than 235 and 2711 respectively.

The Logsum Utility

As mentioned the RUM class is based on the fundamental idea that individuals derive utility U from each alternative in a choice set, such that an individual's utility of an alternative U_i is additively composed of known values in V_i and noise ϵ_i. In the case of the logit model, the specification of this utility is straight forward as there is no

sequentiality in choices. However, in the nested and cross-nested models, things are different. Here individuals might derive utility along their choice path, and might choose based on utility that is only available later in the choice structure. To account for this, in a way that is consistent with utility maximization, an additional term must be added to the utility of structural nests.

The total utility V_m of a structural nest m consist partly of ordinary linear utility W_m, and partly of the *logsum variable* or *logsum utility*, Γ_m, which represents the expected utility of subsequently choosing between the alternatives j in nest m (e.g. taking into account the potential utility of *bus* and *train* when considering the utility of choosing the structural node *public transport*) in order to maximize the given agent's utility given her observed characteristics in x (Koppelman and Bhat, 2006)

$$V_m = W_m + \frac{1}{\mu_m}\Gamma_m \tag{3.15}$$

Where $0 < \mu_m < 1$ is the *Gumbel scale parameter* for the nest as we below use it to scale the within-nest variance. The inverse $\theta_m = \frac{1}{\mu_m}$ is known as the *logsum parameter* (Koppelman and Bhat, 2006). To comply with utility maximization theory, Γ_m is defined such that the nest utility becomes

$$V_m = W_m + \frac{1}{\mu_m} ln \sum_{j \in \mathcal{C}_m} e^{\mu_m V_j} \tag{3.16}$$

In the *GEV* notation we do not explicitly distinguish between the utility of nests and end nodes, but always denote the derived utility V_j for consistent notation. McFadden has shown how the NL probability function including the logsum utility can be derived by inserting equation (3.17) below into the *GEV* probability function (3.7) (McFadden, 1977b; Train, 2009).

Utilities in the NL model

Error terms are additively composed of higher levels of the choice tree, leading to correlation of the error terms in lower nests. For two elemental nodes i, l both with the structural node m as a parent the individuals' utility of choosing i or l would be given by (Train, 2009)

$$\begin{aligned} U_i &= V_i + \varepsilon_m + \varepsilon_i, \qquad V_i = W_m + Y_i \\ U_l &= V_l + \varepsilon_m + \varepsilon_l, \qquad V_l = W_m + Y_l \end{aligned} \tag{3.17}$$

Where the observed utility of the alternative V_i consist of the underlying utilities Y_i that is the utility that varies between the alternatives in the nest and W_m that is the same nest-specific part of the deterministic utility for the nest as seen in (3.16) above and resembles the utility of the shared attributes of the alternatives within the nest.

As no utility is allocated to the root-nest iself, the utility function of an elemental node

k that is a child to the root would be equal to those of the MNL model (3.1)

$$U_k = V_k + \varepsilon_k \tag{3.18}$$

Just as for the MNL model all of the alternatives share the same variance of the (total) error term that is Gumbel-distributed with scale parameter set to 1 (Koppelman and Bhat, 2006). Just as above letting i, l be elemental nodes in the nest m and k an elemental node that is a direct child of the root

$$Var(\varepsilon_m + \varepsilon_i) = Var(\varepsilon_m + \varepsilon_l) = Var(\varepsilon_k) = \frac{\pi^2}{6} \tag{3.19}$$

And the variance of the alternative-specific error terms of the alternatives within nest m are scaled by the *Gumbel scale parameter* μ_m

$$Var(\varepsilon_i) = Var(\varepsilon_l) = \frac{\pi^2}{6\mu_m^2} \tag{3.20}$$

Such that the variance of i and l conditional on already being in the nest m is smaller than the total variance for i or l. As a high value of μ_m equals a high correlation of the error terms for which the shared component ε_m and it variance will be big relative to the alternative specific components $(\varepsilon_i, \varepsilon_l)$ and the variance of these (3.20), showing that the alternatives in the nest will have a low degree of independence of each other (Koppelman and Bhat, 2006).

For a nest m with only two alternatives i, l we have that the *Gumbel scale parameter* μ_m can be estimated as the correlation of the total error terms of the nested alternatives (Cameron and Trivedi, 2005)

$$\mu_m = \frac{1}{\sqrt{1 - \text{Cor}[\varepsilon_m + \varepsilon_i,\ \varepsilon_m + \varepsilon_l]}} \tag{3.21}$$

Generating function for the NL

For the NL model the generating function G (3.6) is defined as (Bierlaire, 2006)

$$\begin{aligned} G(z) &= \sum_{m \in \mathcal{M}} \left(\sum_{j \in \mathcal{C}_m} z_j^{\mu_m} \right)^{\frac{\mu}{\mu_m}} \\ &= \sum_{m \in \mathcal{M}} \left(\sum_{j \in \mathcal{C}_m} e^{\mu_m V_j} \right)^{\frac{\mu}{\mu_m}} \end{aligned} \tag{3.22}$$

Where the sum over $j \in \mathcal{C}_m$ are the available choices in nest m and the sum over $m \in \mathcal{M}$ is over all the available nests in the choice set including the root itself. The parameter μ is the degree of homogeneity of G and $0 < \mu_m < 1$ associated to a nest m is the degree of correlation between error terms of the alternatives in nest m as described in detail above.

Taking the 1st derivative of (3.22) and inserting in the general probability similarly to the derivations for the CNL in (4.1-4.4) below the probability probability function for the NL model (Train, 2009) can be derived as

$$\begin{aligned}\Pr(\mathcal{J}=i|\mathcal{C}) &= \frac{\left(\sum_{j\in\mathcal{C}_m} e^{\mu_m V_j}\right)^{\mu/\mu_m}}{\sum_{n\in\mathcal{M}}\left(\sum_{j\in\mathcal{C}_n} e^{\mu_n V_j}\right)^{\mu/\mu_n}} \frac{e^{\mu_m V_i}}{\sum_{j\in\mathcal{C}_m} e^{\mu_m V_j}} \\ &= \Pr(m|\mathcal{C})\Pr(i|\mathcal{C}_m)\end{aligned} \tag{3.23}$$

The second line shows that the probability function represents the probability of choosing a nest m (which depends on the utilities for the alternatives j in the nest m relative to the utility of the alternatives in all of the nests $n \in \mathcal{M}$, i.e. the utility of the alternatives in the complete choice set $\mathcal{C}$) times the probability of choosing alternative i given that the nest m is chosen (the utility of choosing i relative to the other utilities in the nest m). Summing the whole expression over all nests in $\mathcal{M}$ the terms will be zero for nests m where the option of choosing i is not available.

3.3.3. Commuting-examples for the NL model

To exemplify our proposed properties **1.-3.** for substitution patterns in section 3.3.1 we elaborate on the commuting-example for the MNL model in section (B), still inspired by Koppelman and Bhat (2006).

We first consider the MNL-tree in figure 1 where c_1 is taking the *bus*, c_2 is taking the *train*, and c_3 is to *drive alone*. As pointed out in the MNL-example it is likely that *bus* and *train* have some attributes in common which could either be directly observed (by sharing an alternative-variant regressor like having the same cost in a joint ticket-system) or be unobserved while being represented in the error terms that would correlate (e.g. being more environmental friendly which could correlate with having a higher educational level, or not having a start-up cost as opposed to *car* which could correlate with alternative-invariant regressors like having low income, being young, and/or female).

As a baseline for our two NL-examples we allow for correlation between *bus* and *train* by nesting them together in a *public transport* nest, n_1, as shown in the NL-tree for choice set $\mathcal{C}$ in figure 1.

As a first example we analyze the baseline choice set $\mathcal{C}$ and the addition of the option of choosing *shared-ride*, c_4. We get the choice set $\mathcal{C}'$ in figure 1 by making the strict assumption that *shared-ride* is unnested and though does not have observed or unobserved

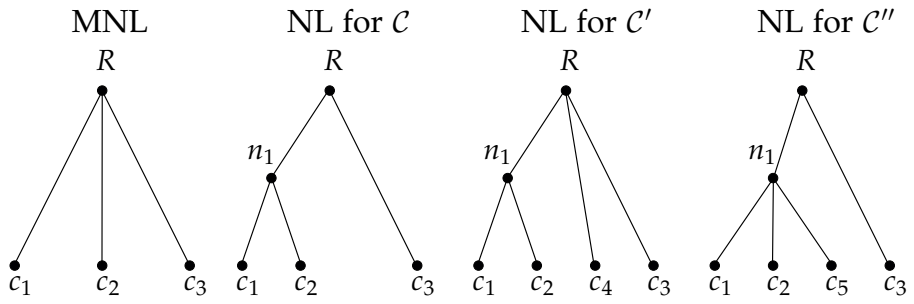


Figure 1: Examples of Nested Logit models for different structures and choice sets. In our analogy c_1 is bus, c_2 is train, c_3 is drive alone (car), c_4 is shared-ride, c_5 is light-rail, and n_1 is a public transport nest.

attributes in common with any other alternative. We will later loosen this restriction by introducing cross-nesting in section 3.4.2.

For a given commuter that with the addition of *shared-ride* would choose this new option of *shared-ride* with positive probability then all of the other probabilities would be decreased to a higher or lesser extent due to substitution, under property **1.-2.** in section 3.3.1 the ratio of probabilities is unchanged for $(bus, train)$ but not for *bus, drive alone*, *bus, drive alone*, or *public transport, drive alone*.

As a second example we take our point of departure in the extended choice set $\mathcal{C}' = c_1, c_2, c_3, c_4, n_1$ in figure 1. Letting a *light-rail*, c_5, enter with positive probability of being chosen we assume equal competition between *light rail*, *bus*, and *train*, thus, allocates *light rail* into the *public transport nest*, n_1, giving us the new choice set $\mathcal{C}''$ in figure 1.

As seen in equation (3.16) above the deterministic utility of choosing the *public transport* nest n_1 contains the *logsum utility* Γ_{n_1} which is affected by the inclusion of *light-rail* within the nest. As long as there actually exists some correlation between each of the pairs for *(bus, train, light-rail)* (i.e. $0 < \mu_{n_1} < 1$) the joint utility of choosing the *public transport* nest will be positively affected by the addition of *light-rail*. To sum up

- Each of the pairs *(bus, train)* and *(shared-ride, drive alone)* are IIA respectively due to property **1.** in section 3.3.1.
- The pairs *(public transport, drive alone)* and *(public transport, drive alone)* are not IIA due to property **3.** in section 3.3.1.

3.4. The Cross-nested Logit model (CNL)

The Cross-nested Logit model (CNL) allows for actual estimation of the cross-elasticities between each pair of alternatives. Intuitively this would reduce the number of choices left for the model builder, as nesting structures can then be estimated, but in reality totally unrestricted estimation of nesting structures becomes infeasible in larger models.

In the CNL model, one additional complication is added compared to the NL model, namely the possibility of each alternative i being in multiple nests, such that they belong to each nest with some weight $0 \leq \alpha_{im} \leq 1$ where for each alternative $\sum_{m \in \mathcal{C}} \alpha_{im} = 1$. Taking a look at the CNL_1 tree in figure 3 in a couple of pages then $\alpha_{c_1 n_1} = 1 \Rightarrow \alpha_{c_1 n_2} = 0$ meaning that alternative c_1 is only a part of the nest n_1 and not of n_2. For the alternative c_2 that is a part of both nests $\alpha_{c_2 n_1} = \kappa \Rightarrow \alpha_{c_2 n_2} = 1 - \kappa$ for $0 < \kappa < 1$.

There are a number of mathematically similar, but notationally varying formulations of the model. These vary in their parametrizations with some implementing restrictions on parameters. Bierlaire (2006) chooses a formulation which he believes to be most general, and throughout this paper we will use his preferred specification. Let $z_i = e^{V_i}$ for each $i \in \{1, ..., J\}$, then the cross nested logit is defined by

$$G(z_1, ..., z_J) = \sum_{m \in \mathcal{M}} \left(\sum_{i \in \mathcal{C}} \alpha_{im} z_i^{\mu_m} \right)^{\frac{\mu}{\mu_m}} \tag{3.24}$$

where m is a nest-index of the set $\mathcal{M}$ of nodes in the graph, $\mathcal{C}$ is the universal choice set in the nesting structure, μ is the degree to which G is homogeneous and μ_m are parameters associated with each nest m. By the *universal* choice set we mean exactly that $\mathcal{C}$ is not in any way conditional on where in the nesting structure m is. Instead setting certain α's equal to zero will determine the nesting structure. This is in opposition to some formulations where instead of summing over $i \in \mathcal{C}$, the summation is restricted to a set $\mathcal{C}_m \in \mathcal{C}$ of nests where $\alpha_{im} \neq 0$.

Train, 2009 and Jong and Kroes, 2014 use a specification that only sum over those choices $\mathcal{C}_m$ that are nested deeper in a decision-tree-like structure. While we keep the general *GEV* formulation, setting the appropriate α's equal to zero allows analysis exactly like the one where the decision space is restricted to go downwards in the tree.

Figure 2 contains an example of a cross nested structures in with cross-nesting such that the nest A_{lr}/B_{ll} can be reached regardless of whether the initial choice is A or B. This is not the simplest thinkable case but illustrates the amount of complexity cross-nesting adds to the problem, as all choices in the highest levels of the tree, now potentially correlate as individuals seek to reach A_{lr}/B_{ll}.

3.4.1. GEV-conditions for the CNL model

The generating function of the CNL needs to satisfy the three criteria for the definition of the generating function G (3.6), thus, a number of a priori restrictions have to be made on the parameters (Bierlaire, 2006). Specifically

- G is non-negative if all $\alpha_{im} > 0$. This should be clear by looking at $G(\cdot)$
- G is homogeneous of degree $\mu > 0$ as long as $\mu > 0$. This is directly visible from $G(tz_1, tz_2, ...tz_J) = \sum_m \left(\sum_j \alpha_{jm} z_j^{\mu_m} t^{\mu_m} \right)^{\frac{\mu}{\mu_m}} = t^{\mu} G(z_1, z_2, ..., z_J)$

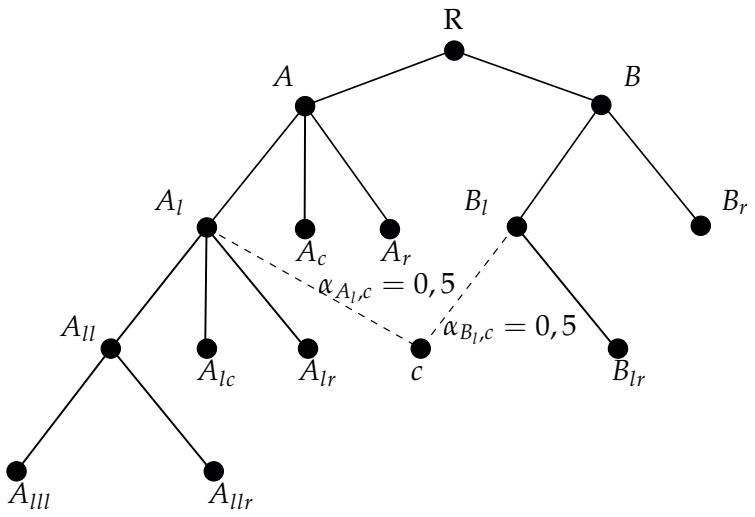


Figure 2: Example of cross-nesting in the actual choice structure. Compared to figure C.4 the alternative c (former B_{ll}) is allowed as an alternative across the nests A_l, B_l with equal weight $\alpha_{A_l,c} = 1 - \alpha_{B_1,c} = \frac{1}{2}$ in each nest.

- $\lim_{z_j \to \infty} G(\cdot) = \infty \ \forall j$ requires that $\sum_m \alpha_{jm} > 0 \ \forall j$, that is all choices have at least some connection to the other nests. If this is not true, we could have only 0-valued α's associated with a nest, causing $\lim_{z_j \to \infty} G(\cdot) = 0$ for that specific j.

- additionally we need $\mu_m > 0 \ \forall m$ and $\mu \leq \mu_m \ \forall m$ both of which are required to satisfy requirement 3. Bierlaire (2006) show a proof for the general k'th derivative of G and find that

$$\frac{\partial^k G(z)}{\partial z_{i_1}, ..., \partial z_{i_k}} = \sum_m \left(\mu_m^k \prod_{n \in \{i_1, ..., i_k\}} (\alpha_{nm} z_n^{\mu_m - 1}) \prod_{n=0}^{k-1} \left(\frac{\mu}{\mu_m} - n \right) y_m^{\frac{\mu - k\mu_m}{\mu_m}} \right) \tag{3.25}$$

  where $i_1, ..., i_k$ are k arbitrarily selected indices in $\mathcal{J}$. For this to be nonnegative when the order of the derivative is odd, and non-positive when the order of derivative is even, there is then only three cases to consider, namely $k = 1$ in which case $G' > 0$ (this is visible in equation (4.1)). Otherwise if $k > 1$ we might have $\mu = \mu_m$ in which case the derivative is always 0, as $\prod_{n=0}^{k-1}(1 - n)$ gives a 0 when $n = 1$.

  We might also have $k > 1$ and $\mu < \mu_m$ the sign of the entire derivative is given directly from the $\prod_{n=0}^{k-1}(\frac{\mu}{\mu_m} - n)$ term. Clearly in this case μ/μ_m lies between 0 and 1, and thus only the first term where $n = 0$ in the product will be positive. Thus there is 1 positive term in the product and $k - 1$ negative terms, implying

$$\frac{\partial^k G(z)}{\partial z_{i_1}, ..., \partial z_{i_k}} \begin{cases} \geq 0, \text{ if } k \text{ is odd} \\ \leq 0, \text{ if } k \text{ is even} \end{cases} \tag{3.26}$$

  This argument also makes it clear that while $\mu > \mu_m$ might produce a valid GEV

generating function, whether G is a GEV generator depends crucially on the relative size of the parameters.

3.4.2. Commuting-example for the CNL model

Returning to the analogy of commuter's choice between *bus, train, shared-ride,* and *drive alone* (c_1, c_2, c_3, c_4) we go from the NL tree for choice set $\mathcal{C}'$ in figure 1 to the CNL_1 tree in figure 3 by allowing *shared ride*, c_4, to be cross-nested.

The nest n_1 is the *public transport* nest or more generally a *group travel* nest as we imagine *shared-ride* to have some similarities with *bus* and *train* by also being some kind of group travel. On the other hand it also differs by the fact that the means of transport is a private auto wherefore we also assumes that *shared-ride* correlates with *drive alone* and we nest them together in a *private auto* nest, n_2. As c_4 is available across both nests we cannot assume IIA between any pair of alternatives in the tree as a change in any nest would affect this.

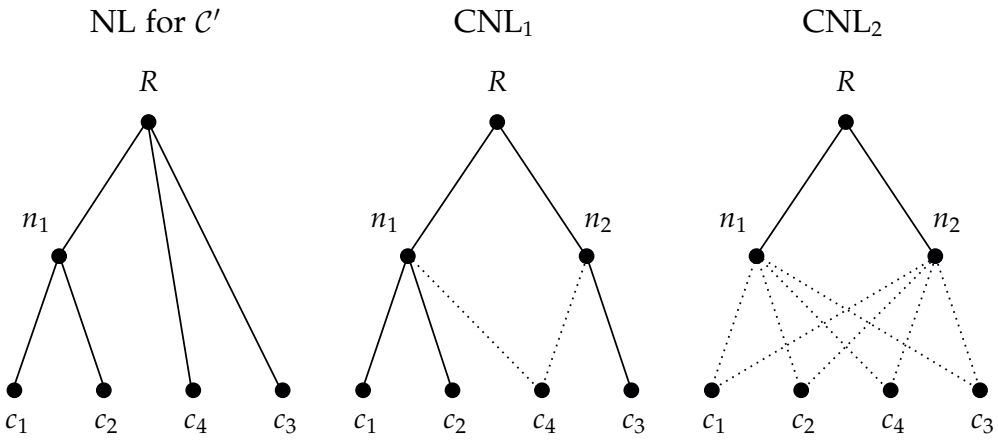


Figure 3: Examples Nested- and Cross-nested Logit models showing three different choice structures for the same choice set.
Through our analogy c_1 is bus, c_2 is train, c_3 is drive alone (car), c_4 is shared-ride, n_1 is a public transport or *group travel* nest, and n_2 is a *private auto* nest.

A common starting point for estimation is to define the tree structure by letting c_4 take equal part in both nests $\alpha_{c_4 n_1} = \alpha_{c_4 n_2} = \frac{1}{2}$ and also for the CNL_1 case restrict the other alternatives to not be cross-nested $\alpha_{c_1 n_1} = \alpha_{c_2 n_1} = \alpha_{c_3 n_2} = 1$. By identifying the system and applying appropriate estimation methods (see section 4 and 6) the idea is to actually estimate the degree to which *shared-ride* belongs to each nest. If $\alpha_{c_4 n_1} = 1 - \alpha_{c_4 n_2}$ is found to be sufficiently close to 0 or 1 we have shown *shared-ride* is indeed not cross-nested and our model collapses to a NL-structure where c_4 only belongs to the nest m for which we estimate $\alpha_{c_4 m} \to 1$.

Other 2-level nesting structures could be assumed such as a *road* nest for *(bus, shared-ride, drive alone)* or a *long-distance* nest for *(train, shared-ride, drive alone)*. The CNL_2 tree in

figure 1 allows each possible pair of the choices c_1, c_2, c_3, c_4 to be correlated as a starting point, thus, in theory lets us estimate any 2-level nesting structure with two nests. Unfortunately, as one could imagine, the objective function that we would then seek to minimize is high-dimensional as there is an α-parameter for each cross-nest and far from convex, so that we are not guaranteed to have convergence nor be sure that our result is the global maximum.

4. Estimation

4.1. Likelihood function

In order to derive the proper likelihood of observing data generated by a cross-nested Data Generating Process (DGP), one need the derivative of (3.24). Fortunately we can easily derive this as

$$\begin{aligned}\frac{\partial G}{\partial z_i} &= \sum_{m\in\mathcal{M}} \frac{\partial}{\partial z_i}\left(\sum_{j\in\mathcal{C}} \alpha_{jm} z_j^{\mu_m}\right)^{\frac{\mu}{\mu_m}} \\ &= \sum_{m\in\mathcal{M}} \left[\frac{\mu}{\mu_m}\left(\sum_{j\in\mathcal{C}} \alpha_{jm} z_j^{\mu_m}\right)^{\frac{\mu}{\mu_m}-1} \cdot \alpha_{im}\mu_m z_i^{\mu_m-1}\right]\end{aligned} \tag{4.1}$$

This gives us an analytical gradient of G and by inserting it in the GEV definition $\Pr(\mathcal{J} = i|\mathcal{C})$ (3.10) (Bierlaire, 2006). Rearranging terms we get

$$\begin{aligned}\Pr(\mathcal{J}=i|\mathcal{C}) &= \frac{z_i\left(\mu\sum_m \alpha_{im} z_i^{\mu_m-1}\left(\sum_j \alpha_{jm} z_j^{\mu_m}\right)^{\frac{\mu}{\mu_m}-1}\right)}{\mu\sum_n\left(\sum_j \alpha_{jn} z_j^{\mu_n}\right)^{\frac{\mu}{\mu_n}}} \\ &= \frac{\sum_m \alpha_{im} z_i^{\mu_m}\left(\sum_j \alpha_{jm} z_j^{\mu_m}\right)^{\frac{\mu}{\mu_m}-1}}{\sum_n\left(\sum_j \alpha_{jn} z_j^{\mu_n}\right)^{\frac{\mu}{\mu_n}}}\end{aligned} \tag{4.2}$$

where n is a secondary nest index, as both $G(\cdot)$ and $G_i(\cdot)$ sums over all nests. Now adding one count of $\sum_j \alpha_{jm} z_j^{\mu_m}$ to the exponent in the numerator and using that the sums over m and n can be rearranged without concern as they are independent of each others indices, we arrive at

$$\Pr(\mathcal{J}=i|\mathcal{C}) = \sum_m \frac{\left(\sum_j \alpha_{jm} z_j^{\mu_m}\right)^{\frac{\mu}{\mu_m}}}{\sum_n\left(\sum_j \alpha_{jn} z_j^{\mu_n}\right)^{\frac{\mu}{\mu_n}}} \times \frac{\alpha_{im} z_i^{\mu_m}}{\sum_j \alpha_{jm} z_j^{\mu_m}} \tag{4.3}$$

This expression of the probability has a convenient interpretation as the summed probabilities of being in nest m conditional on the choice set, times the probability of choosing

option i given that one is in nest m, that is

$$\Pr(\mathcal{J} = i|\mathcal{C}) = \sum_m \Pr(m|\mathcal{C})\Pr(i|m) \tag{4.4}$$

Which simply states that the probability of making any choice i is the probability of choosing i from nest m, summed over all nests with access to option i (Bierlaire, 2006).

Compared to the probability function for the NL (3.23) the CNL (4.3) is different as the sum over j is over the whole choice set $\mathcal{C}$ and not only the within nest sets $\{\mathcal{C}_m\}_{m=1}^M$. As mentioned this is countered simply by setting $\alpha_{im} = 0$ whenever choice i is not in nest m.

Using the expression of probability given in (4.3) it is relatively simple to derive the likelihood. Letting (4.3) be the probability of observing an observation c_i, consequentially the product of these probabilities over the sample *of observed outcomes* will represent the probability of observing the entire dataset within a CNL model parametrized with some parameters $\beta, \alpha, \mu_m, \mu$. Letting d_{kj} be a dummy with value 1 if an individual k chooses choice j and 0 otherwise, we can write the likelihood as

$$\mathcal{L}(\beta, \alpha, \mu_m, \mu|z) = \prod_{k=1}^K \Pr(\mathcal{J} = i|\mathcal{C})^{d_{ki}} \tag{4.5}$$

and as a direct extention thereof

$$\begin{aligned} \ln \mathcal{L}(\beta, \alpha, \mu_m, \mu|z) &= \sum_{k=1}^K d_{ki} \ln \Pr(\mathcal{J} = i|\mathcal{C}) \\ &= \sum_{k=1}^K d_{ki} \ln \left(\sum_m \frac{\left(\sum_j \alpha_{jm} z_j^{\mu_m}\right)^{\frac{\mu}{\mu_m}}}{\sum_n \left(\sum_j \alpha_{jn} z_j^{\mu_n}\right)^{\frac{\mu}{\mu_n}}} \times \frac{\alpha_{im} z_i^{\mu_m}}{\sum_j \alpha_{jm} z_j^{\mu_m}} \right) \end{aligned} \tag{4.6}$$

Here also we see that the multinomial logit is fully contained in the CNL framework as when $\text{size}(\mathcal{C}) = 1$, $\mu_m = \mu = 1$ and $\forall j, m : \ \alpha_{jm} \in \{0, 1\}$ the sums over nests can be dropped, resulting in the first fraction collapsing to a one. Due to the restrictions of parameters the second fraction will then exactly be the probability from the multinomial logit, and the entire expression is then the likelihood of a multinomial logit.

As noted by Newman et al. (2018) the expression of the likelihood highlights that $\Pr(\mathcal{J} = i|\mathcal{C})$ does not depend on individual k's actual choice, that is it does not depend on d_{ki}. This has computational benefits as it allows the calculation of the vector $\boldsymbol{P}(\mathcal{J} = i|\mathcal{C})$ of probabilities over all alternatives, and summing the subset of this vector where a dummy vector $\boldsymbol{d}_k$ is activated. By doing this the computation of P can be vectorized, allowing dynamically typed code to approach the speeds achievable in more low-level code.

4.2. Standard errors

In this paper we only calculate standard errors when using pre-build software which is able to do so. In Biogeme standard errors are calculated as the Ramer-Crao lower bound $-E[\nabla^2 \mathcal{L}(\theta)]^{-1}$ (Bierlaire, 2016). Since this is dependent on the second derivatives of $\mathcal{L}$ it is infeasible to compute in python, without heavy optimization. There is no information available as to how Larch calculates standard errors, but considering that bootstrapping will be an extremely costly method, and that there are no easy ways to implement the delta method, we suspect Larch uses the same method as in Biogeme. As a sidenote the availability of relatively strong theoretical results on the CNL should make implementing standard errors using the Ramer-Crao lower bound relatively straight forward in faster languages.

After completing this initial step of calculating $\sigma_k = -E[\nabla^2 \mathcal{L}(\theta)]^{-1}$ it should then be straight forward to calculate t-statistics as $t_k = \beta_k / \sigma_k$ and p-values as $p = 2(1 - \Phi(t_k))$. Standard errors on the marginal effects are probably easiest to calculate using the delta method.

4.3. Marginal effects

To calculate marginal effects with respect to data x the second derivative $\frac{\partial G_i}{\partial x}$ of (3.24) is needed, as the first derivative is present in the expression of $\Pr(i|\mathcal{C})$. This is found by (tedious) application of the chain rule to be

$$\begin{aligned}\frac{\partial^2 G}{\partial z_i \partial x} = \sum_m \frac{\mu}{\mu_m} \alpha_{im} \mu_m \Bigg[\left(\frac{\mu}{\mu_m} - 1\right) \left(\sum_j \alpha_{jm} z_j^{\mu_m}\right)^{\frac{\mu}{\mu_m}-2} \cdot \left(\sum_j \alpha_{jm} z_j^{\mu_m - 1} \mu_m \beta_j\right) \cdot z_i^{\mu_m - 1} \\ + \left(\sum_j \alpha_{jm} z_j^{\mu}\right)^{\frac{\mu}{\mu_m}-1} \cdot (\mu_m - 1) z_i^{\mu_m - 2} \beta_i \Bigg]\end{aligned} \tag{4.7}$$

With this result it is then possible to derive the marginal effects as

$$\begin{aligned}\frac{\partial \Pr(i|\mathcal{C})}{\partial x} &= \frac{\frac{\partial}{\partial x}\left[e^{x\beta_i}\frac{\partial G}{\partial z_i}\right] \cdot \left(\sum_j e^{x\beta_j}\frac{\partial G}{\partial z_j}\right) - e^{x\beta_i}\frac{\partial G}{\partial z_i}\frac{\partial}{\partial x}\left(\sum_j e^{x\beta_j}\frac{\partial G}{\partial z_j}\right)}{\left(\sum_j e^{x\beta_j}\frac{\partial G}{\partial z_j}\right)^2} \\ &= \frac{e^{x\beta_i}\left(\beta_i \frac{\partial G}{\partial z_i} + \frac{\partial^2 G}{\partial z_i \partial x}\right) \cdot \left(\sum_j e^{x\beta_j}\frac{\partial G}{\partial z_j}\right) - e^{x\beta_i}\frac{\partial G}{\partial z_i}\sum_j e^{x\beta_j}\left(\beta_j \frac{\partial G}{\partial z_j} + \frac{\partial^2 G}{\partial z_j \partial x}\right)}{\left(\sum_j e^{x\beta_j}\frac{\partial G}{\partial z_j}\right)^2} \\ &= \beta_i \Pr(i|\mathcal{C}) + \frac{e^{x\beta_i}\frac{\partial^2 G}{\partial z_i \partial x}}{\sum_j e^{x\beta_j}\frac{\partial G}{\partial z_j}} - \Pr(i|\mathcal{C}) \cdot \sum_j \beta_j \Pr(j|\mathcal{C}) + \frac{e^{x\beta_j}\frac{\partial^2 G}{\partial z_j \partial x}}{\sum_{j'} e^{x\beta_{j'}}\frac{\partial G}{\partial z_{j'}}}\end{aligned} \tag{4.8}$$

Recall that in the simply multinomial case these marginal effects can be derived to be $p_i\left(\beta_i - \sum_j p_j \beta_j\right)$, and that for the multinomial model, the first derivative of G is 1, to

see that this expression will collapse to the multinomial marginal effects under suitable restrictions. In general however we then have that

$$\frac{\partial \Pr(i|\mathcal{C})}{\partial x} = \Pr(i|\mathcal{C}) \left(\beta_i - \sum_j \left[\beta_j \Pr(j|\mathcal{C}) + \frac{e^{x\beta_j} \frac{\partial^2 G}{\partial z_j \partial x}}{\sum_{j'} e^{x\beta_{j'}} \frac{\partial G}{\partial z_{j'}}} \right] \right) + \frac{e^{x\beta_i} \frac{\partial^2 G}{\partial z_i \partial x}}{\sum_j e^{x\beta_j} \frac{\partial G}{\partial z_j}} \tag{4.9}$$

4.4. Identification

Due to the complexity of the generating function $G(z_1, ..., z_J)$ (eq.3.24) the exact set of necessary constraints for model identification is not known. Though, in the literature there are proposed several different sufficient restrictions for model estimation that are either complementary or substitutes.

The sufficient conditions in section 3.4.1 for G to be a GEV generating function should be applied. For an alternative j belonging to l nests it is common in application to fix the weights to $\alpha_{jm} = 1/l$ (Jong and Kroes, 2014). Among others this approach is used by Stephane Hess et al (2012) and motivated by the complexity of actually estimating the α-coefficients.

Inserting the definition of $z_i = e^{V_i}$ into the probability function (4.3) and specifying $V_i = (\beta_i^0 + \psi) + x(\beta_i + \delta)$ gives a version of the likelihood where parameters are directly accesible, and where we can study some identification problems:

$$\begin{aligned}
\Pr(i|\mathcal{C})|_{V_i=(\beta_i^0+\psi)+x(\beta_i+\delta)} &= \sum_m \frac{\left(e^{\mu_m x\delta}\right)^{\frac{\mu}{\mu_m}} \left(e^{\mu_m \psi}\right)^{\frac{\mu}{\mu_m}} \left(\sum_j \alpha_{jm} e^{\mu_m(\beta_j^0 + x\beta_j)}\right)^{\frac{\mu}{\mu_m}}}{\sum_n \left(e^{\mu_n x\delta}\right)^{\frac{\mu}{\mu_n}} \left(e^{\mu_n \psi}\right)^{\frac{\mu}{\mu_n}} \left(\sum_j \alpha_{jn} e^{\mu_n(\beta_j^0 + x\beta_j)}\right)^{\frac{\mu}{\mu_n}}} \times \frac{e^{\mu_m x\delta} e^{\mu_m \psi} \alpha_{im} e^{\mu_m(\beta_i^0 + x\beta_i)}}{e^{\mu_m x\delta} e^{\mu_m \psi} \sum_j \alpha_{jm} e^{\mu_m(\beta_j^0 + x\beta_j)}} \\
&= \frac{e^{\mu x\delta}}{e^{\mu x\delta}} \frac{e^{\mu \psi}}{e^{\mu \psi}} \sum_m \frac{\left(\sum_j \alpha_{jm} e^{\mu_m(\beta_j^0 + x\beta_j)}\right)^{\frac{\mu}{\mu_m}}}{\sum_n \left(\sum_j \alpha_{jn} e^{\mu_n(\beta_j^0 + x\beta_j)}\right)^{\frac{\mu}{\mu_n}}} \times \frac{\alpha_{im} e^{\mu_m(\beta_i^0 + x\beta_i)}}{\sum_j \alpha_{jm} e^{\mu_m(\beta_j^0 + x\beta_j)}} \\
&= \Pr(i|\mathcal{C})|_{V_i=\beta_i^0+x\beta_i}
\end{aligned} \tag{4.10}$$

Thus there is both a scale, and a location problem with identification. Regarding the MNL model there exists a common strategy for solving both by arbitrarily fixing $\beta_j^0 = 0$ for one of the alternatives j, and fixing one parameter of β_i in each choice (Ben-Akiva et al., 1985).

For the NL model there exist two traditional approaches for solving the location-issue. The one known as the "elemental strategy" fixes $\beta_j^0 = 0$ for one of the elemental nodes (arbitrarily choosing one of the "dead ends") as well as for all of the structural nodes. The "structural strategy" on the other hand imposes $\beta_j^0 = 0$ for one arbitrary "child" of the root and each structural node of the tree. Bierlaire (1997) introduces the "ortogonal strategy" where the "constraint subspace" contains all the necessary constraints and is chosen in order to be orthogonal to the "invariant subspace" that contains all the

overspecification due to ASCs. The main benefit of the orthogonal strategy is that is reduces the chance that a maximization algorithm will stop too early and that it will arrive at different results depending on the arbitrary choices of j for which $\beta_j^0 = 0$ (Bierlaire, 1997).

A separate issue is that of "utility-shifting" where adding and/or subtracting to utility respectively in nests and nest-children does not affect the overall utility of the nest because of the logsum term. In the NL models is a sufficient condition to fix β_j^0 and one of the parameters in the β_j vector for each of the structural nodes. A simple proof of this can be found in appendix D, a more general approach is available in (Bierlaire, 1997).

Knowledge of the exact restrictions required for identification of the CNL model is limited. With regards to the issue of utility shifting, the non-pure nesting structure makes the shifting many dimensional, as there is no longer one nest for each sub-choiceset $\mathcal{C}_m$. Thus one need to solve as many interdependent equations as there are cross-nested nests, to show identification. For the case of ACS, mathematically one needs to show that a parametrization satisfies

$$\begin{bmatrix} V_s^\psi \\ \vdots \\ V_d^\psi \end{bmatrix} = \begin{bmatrix} H_s + \psi_s + \frac{1}{\mu_s} \ln \sum_j e^{V_j \mu_s - \psi_j} \\ \vdots \\ H_d + \psi_d + \frac{1}{\mu_d} \ln \sum_j e^{V_j \mu_d - \psi_j} \end{bmatrix} \neq \begin{bmatrix} V_s \\ \vdots \\ V_d \end{bmatrix} \tag{4.11}$$

where $s...d$ are indices for all the nests. Unlike in NL where these can be solved separately, the cross nesting implies that the sum over j will count the same ψ_j in multiple equations, meaning some equations must be solved in pairs, triplets etc. It is this highly context-dependent nature of the problem that makes it hard.

5. Data

Generating data as if it was derived from a CNL model is not straight forward because of the complicated correlations across alternatives and nests. We employ a shortcut which allows us to sample sequential multinomial trials and convert these into proper nested and cross-nested trials using sampling based on the probabilities derived above. As an example notice in (4.3) that the last term $\frac{\alpha_{im} z_i^{\mu_m}}{\sum_j \alpha_{jm} z_j^{\mu_m}}$ is essentially the ordinary multinomial choice probability with a bit of added complexity in the form of α's and μ's, and that for a child-node with only one parent-node the first term will provide only one hit in the numerator as all other α_{jm} are 0. Similarly the denominator will produce one hit for each value of m. In the case of two nested binary outcome choices, (4.3) thus collapses to the product of the probabilities of two individual multinomial trials weighted by μ_m's. This pattern extends nicely to larger nested structures, making it easy to generate datasets of the nested kind.

In the cross nested model there's the added difficulty of partial node-membership, that is $\alpha_{im} \neq \{0,1\}$ in general. Our approach handles this simply by drawing from the probability density $\Pr(\mathcal{J} = i|m)$ as derived above. In this way we simulate full paths through the tree, with each step modelled as a logit-like trial, now indcluding α parameters. In this way we get one choice per individual per nest. This gives us one to many choices as we observe a choice $(c_1|c_2)$ at n_1 and a choice $(c_2|c_3)$ at n_2. Naturally these cannot co-occur, and thus we drop observations according to the individuals choice at R. Appendix E contains a discussion of the computational issues this double counting gives.

5.1. DGP structure in simulated data

For the sake of simplicity we restrict out estimator testing to datasets with simple nesting structures. We construct both a dataset of nested and cross-nested data, each with three observed outcomes c_1, c_2, c_3. A visual impression of the DGP's is given in figure figure 4. In the nested case we create two nests m_0 and m_1 each with binary outcomes, and linked by one option in m_0 to the root of m_1. To accommodate cross-nesting we have to add a third nest, making both of the roots child-nodes n_1, n_2 unobserved.

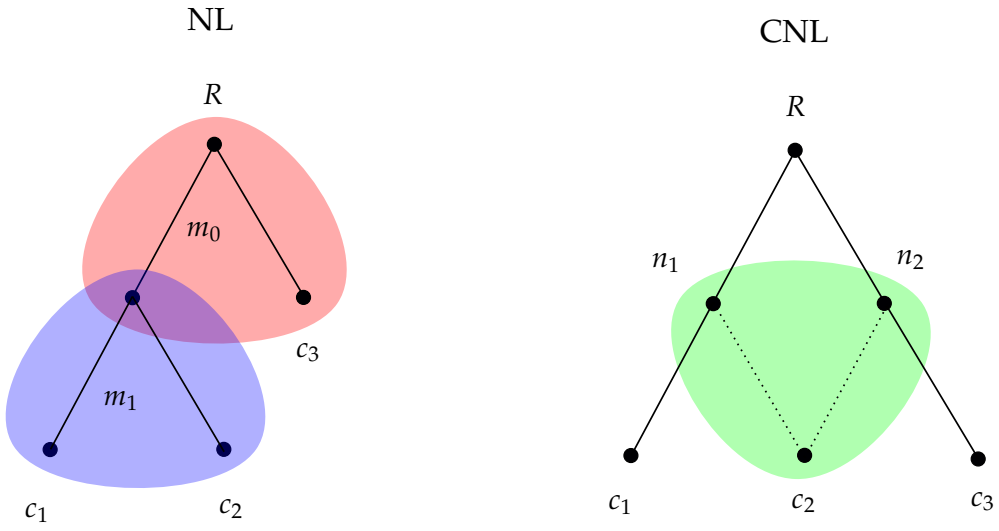


Figure 4: The nesting structures employed as simulated data.

This kind of cross nesting is indeed very simple compared to the capabilites of CNL, but are computationally less demanding while maintaining the important trait that choosing branch (n_1 or n_2) does not restrict the choice-set to proper sub-branches.

As the data is multinomial one need to specify a vector of β-parameters associated with each choice in the dataset excluding the root node R. In doing so researchers might choose to include different variables at different nodes. To maximize the amount of information gained from each observation we choose not to do this, and instead

include the two independent vectors x_1, x_2 as regressors in all utility equations. Since there are five choices in the CNL DGP, we get a total of 5 equations (4 in the NL case) of the form

$$V_j = [x_1 \; x_2]\beta_j \tag{5.1}$$

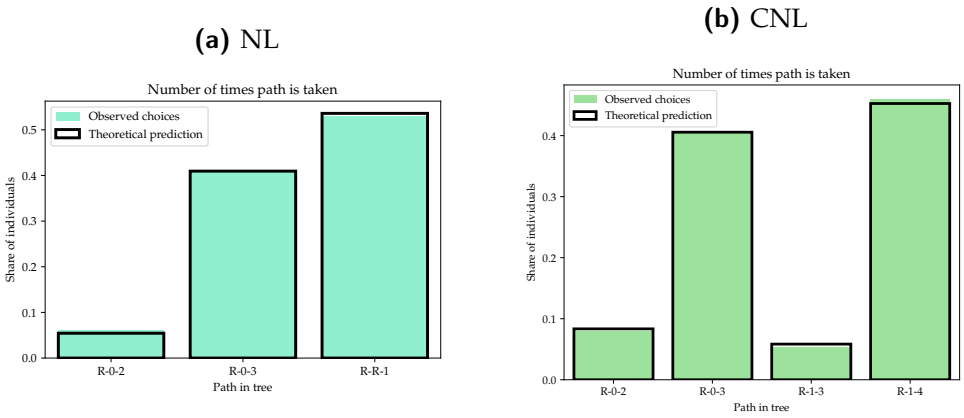


Figure 5: Actual and theoretical choice probabilities. The x-axis is read as follows. From left to right is the path through the tree, with R being the root, and apart from that the mapping is $(n_1, n_2, c_1, c_2, c_3) \to (0, 1, 2, 3, 4)$.

By simulating and selecting data as described above we construct a dataset of 100 "individuals", and their hypothetical decision at every node of the CNL tree depicted in figure 4. The expected probability of a given path (i.e. $R \to n_1 \to c_2$) is then given by the product of nest-conditional probabilities at each step, and we can quite easily calculate these. Since we simulate the entire structure and not just the lowest nodes, we can trace out otherwise un-separable paths in the tree. Figure 5 shows the observed share of choosers following a given path in the CNL tree and compares it to the theoretical number. First of all, the apparent dissymmetry of the CNL path shares compared to the ones in the NL model is useful in understanding the flexibility of the CNL model. Secondly this figure serves as a way of validating that our code functions as anticipated.

5.2. Data validation

We use a number of tests to ensure the DGP is correctly implemented. These all rely on the various probabilities debated in the section on estimation. Specifically we know that the following identities must hold true

$$\begin{aligned} \sum_{i \in \mathcal{C}} \Pr(i|\mathcal{C}) &= 1 \\ \sum_{m \in \mathcal{M}} \Pr(m|\mathcal{C}) &= 1 \end{aligned} \tag{5.2}$$

And finally we also know $\forall m : \; \sum_{i \in m} \Pr(\mathcal{J} = i|m) = 1$. That is 1) the summed probability of making choice i over the entire choice set must be 1, 2) the summed probability of

	c_1	c_2	c_3
All	0.12	0.35	0.53
Male	0.11	0.34	0.55
AC	0.11	0.27	0.62
Age 15-30	0.14	0.26	0.60
Age 30-40	0.13	0.32	0.55
Age 40-50	0.11	0.35	0.54
Age 50-60	0.11	0.40	0.49

Table 1: Summary statistics: distribution of choices for subgroups in data.

entering nest m summed over all nests must be 1 and 3) Within all nests, the summed probability of choosing $i \in m$ must be 1 when conditioning on being in the specific nest.

Knowing that our code passes all of these criteria we are confident that we simulate data from the correct DGP.

5.3. Real data (the DREAM dataset)

In order to model probabilities regarding the course of those that are out of the labour market on sick leave, we use weekly data from the semi-governmental institution *The Danish Institute for Economic Modelling and Forecasting*, the DREAM Group, based on their agreement with LO, The Danish Confederation of Trade Unions. More specifically the december 2017-version of the microeconometric database covering week-by-week socioeconomic status of all 5,6 million Danish citizens subdivided into 36 labour market categories. The registry lays the foundation for their socioeconomic projection model (DREAM, 2018). Table 1 show summary statistics for the sample.

For simplicity we model the one-year course for the group of 84.123 persons on sick leave that received unemployment benefits ("sygedagpenge") in week 48 of 2016 (last week of november). Thus, the selection besides from being those hit by prolonged illness is limitied to those that have been a member of an unemployment insurance fund and have been in work recently in order to fulfill the conditions for receiving unemployment benefits.

Our dataset is then reduced to 73.864 individuals by dropping the 10.259 above 60 years olds ultimo november 2017 in order to avoid interference with ordinary retirement schemes. Furthermore 6.348 are dropped who in week 48 of 2017 (last week of november) where on either paternity leave, no longer living in Denmark, dead, or neither employed nor receiving unemployment benefits (i.e. no information on one's livelihood).

We end up with 67.516 observations for which the following regressors are included:

$$x = \left[\text{male, ac, age, age}^2\right] \tag{5.3}$$

Where *ac* is a dummy for the 7 pct. of individuals in the dataset who are members of one of the four unemployment insurance funds for college graduates (*AJKS*, *AKA*, *CA* or *MA*).

The choices are defined based on their socioeconomic status in week 48 in 2017 (one year after having been on sick leave).

c_1 (*"Ordinary unemployment"*): Receiving unemployment benefits while actively seeking jobs.

c_2 (*"Sick"*): On sick leave ("sygedagpenge"), early retirement due to sickness ("førtidspension"), subsidized work on reduced hours due to sickness ("fleksjob"), or being in one of several courses for partially sick where one's amount of work-ability is examined.

c_3 Ordinary employment or under education.

As the linear RUM setup (equation 3.1) is based on utility maximizing agents, our estimation model is to assume that the choice between c_1 and c_3 depends on the opportunity cost of being out of labour, i.e. as men and college graduates in general earn above the median wage we use these dummies as proxies for having a higher utility from choosing c_3 relative to c_1, besides from the other characteristics associated with ticking theese boxes which also for the most part drag in the direction of c_3. As a higher age is also correlated with a higher wage we similarly can expect older people to have a higher probability of choosing c_3 while a squared term for age allows for nonlinearity such that there could exist a maximum point for age from which on the probability of choosing c_3 over c_1 could be decreasing instead. Leaving the world of utility maximizing individuals for a second, the employer's side of an imperfect labour market might also have preferences regarding an employer's ideal age that can be described by a similar 2nd order polynomial which would further increase the effect. Given expectations about different probabilities for c_1 and c_3 based on regressors we assume that the error terms tend towards being uncorrelated and that there should not exist cross-nesting between this pair of alternatives.

c_2 on the other hand is difficult to fit into the RUM-setup. Thus we assume that being too sick to be available for the ordinary labour market is mainly correlated with a higher age while correlation with the other regressors is less significant. Assuming a near random draw of those hit by a more permanent sickness, we expect the hidden attributes in the error term to be both correlated with both those who choose c_1 and c_3 and thus we allow for cross-nesting with both nest n_1 and n_2, see the CNL-structure in 4 above.

6. Application to simulated data

Simple econometric models can be estimated in a reasonable amount of time on even very modest computers, but as complexity grows estimation becomes increasingly challenging. While there are optimization routines that can optimize the CNL likelihood in reasonable time, implementing these routines is beyond the scope of this paper. For properly efficient estimation the only real option is one of the two specialized tools *Biogeme* and *Larch*. In all of the following we restrict one parameter in each nest to it's true value before estimation, we also assume α and μ_m's to be fixed at their true values.

We implement reasonably fast versions of the probability equations in 4.2, as well as implementations of $P(m|C)$ and $P(i|m)$ using Numpy in python 3.6. Evaluating the likelihood function in a single point takes approximately 6-7 seconds when using 1.000 observations. As the number of observations increase this process slows further as evident in table 2. This is a major obstacle for estimation, as the large number of parameters, and the complicated effects from parameters on the likelihood, means many observations are generally required to get accurate results. Optimizing the $J \times K$ parameters jointly fails to converge when using the standard SciPy minimization routine. We do show some results from joint estimation, but one should note that these are from non-converging optimizers.

Table 2: Speed comparison five evaluations of $\log \mathcal{L}_K$ at different data sizes

	Obs=10	Obs=50	Obs=100	Obs=1000	Obs=5000
0	1.0	5.02	10.23	124.15	1497.02

Numbers are summed time over five iterations of $\log \mathcal{L}_K(\beta, \Theta; x)$ over the dataset with α set to the CNL structure and β as described. Units is the running time relative the running time of the $K = 10$ dataset. The running time when $Obs = 10$ is approximately 0.35 seconds.

In the following we attempt to estimate model parameters in the CNL model through two very weak, but conceptually simple mechanisms. Especially the first attempt is very unlikely to produce unbiased estimates, let alone estimates that converge. The purpose of showing these attempts should thus not be seen as trying to argue that these methods are in general useable, but to explore the options for estimation in models which are so complex it requires specialist knowledge to code up estimators.

6.1. Iterative estimation

The perhaps simplest way of estimating the model parameters is the following scheme. We assume that both α's, μ and μ_m are known and fixed to their true value. Recall that

there is a β for each x for every choice (with some set to 0), so we can arrange β as a matrix of size $J \times K$.

$$\beta = \begin{array}{c} \\ n_1 \\ n_2 \\ c_1 \\ c_2 \\ c_3 \end{array} \begin{array}{c} \begin{array}{cc} \boldsymbol{\beta_0} & \boldsymbol{\beta_1} \end{array} \\ \begin{pmatrix} \beta_{00} & \beta_{01} \\ \beta_{10} & \beta_{11} \\ \beta_{20} & \beta_{21} \\ \beta_{30} & \beta_{31} \\ \beta_{40} & \beta_{41} \end{pmatrix} \end{array} \hookrightarrow \begin{array}{c} \\ n_1 \\ n_2 \\ c_1 \\ c_2 \\ c_3 \end{array} \begin{array}{c} \begin{array}{cc} \boldsymbol{\beta_0} & \boldsymbol{\beta_1} \end{array} \\ \begin{pmatrix} \beta^*_{00} & \beta^*_{01} \\ \beta^*_{10} & \beta_{11} \\ \beta^*_{20} & \beta_{21} \\ \beta^*_{30} & \beta^*_{31} \\ \beta^*_{40} & \beta_{41} \end{pmatrix} \end{array} \tag{6.1}$$

Our iterative optimization scheme then initialized $\beta^{\text{init}} = \boldsymbol{\beta^*}$, sets β^{init}_{i0} to one as initial value, except for the parameters of choices c_0 (which is structural) and c_3 which are set to their true values. This is in accordance with the identification stategies outlines in section 4 and illustrated by the right hook arrow in equation (6.1). Even though this significantly reduces the number of parameters the Scipy minimizer is still not able to converge when maximizing the likelihood w.r.t the full set of free parameters. We then iteratively optimize each element of β assuming all other elements fixed. After each iteration β is updated to include the found optimal value. This process is then carried out in a loop until the change in the entire β matrix from t to $t+1$ is sufficiently small. In pseudocode a single iteration over the matrix would look like

Algorithm 1: Iterative Optimizer

Input : β: A matrix of parameter vectors
Input : C: A vector of choice indices
foreach *parameter* β_i **in** β **do**
  foreach *choice* c_i **in** C **do**
    $\beta^*_{[b][c]}$ = `OptimizeScalar`($\log \mathcal{L}^{i|\mathcal{C}}_K(\beta, x, \Theta)$) w.r.t $\beta_{[b][c]}$
    $\beta_{[b][c]} \leftarrow \beta^*_{[b][c]}$

As mentioned this process is carried out repeatedly over the continuously updating matrix. Of course this estimation approach is not ideal, but computationally it is feasible to implement within the scope of this paper. For each iteration over the matrix we update the parameter space to be searched, to a slightly narrower segment recentered around the current value of $\beta_{[b][c]}$.

An intuitive argument for why this method will at least in some cases converge is that if each single-dimensional optimization moves the estimated β closer to it's true value, it improves the conditions for estimating the other β's in the following iteration bringing them closer to their optimum as well. Thus in the end parameters should converge on their true values. For this to be the case it is essential that the sub-routines consequently move estimates closer to their true values, which will only be the case if the likelihood function is dimensionwise concave, as it is otherwise possible for estimates to move away from their true values in the subroutines. More formally if $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}$ is a likelihood function dependent on n parameters β, then if every cross-section of $\ell = \mathcal{L}(\beta_i|\beta_{-i}, \Theta, x)$, $\ell : \mathbb{R} \to \mathbb{R}$ has a single optimum (implying the function is strictly concave for some sign)

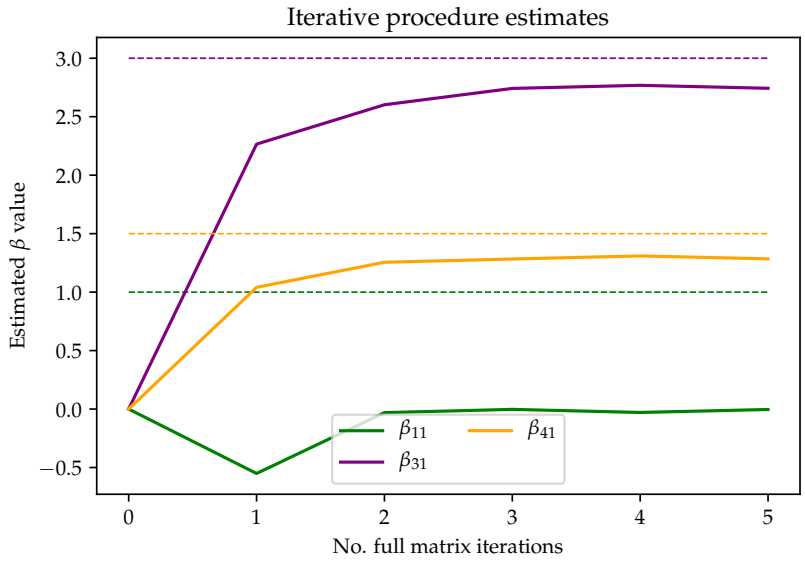


Figure 6: Parameter development using iteration. Dashed lines are true parameter values. Each step on the first axis corresponds to a full loop of algorithm 1 over the β-matrix.

and the full function $\mathcal{L}$ has a single optimum as well, then the optima found in ℓ after each iteration will converge to the global optimum. This is because 1) the optima in ℓ cannot be lower than the global optimum in $\mathcal{L}$, 2) there must be at least one dimension in which a non-global optimum is deviated from (otherwise the function would have local optima) and 3) this deviation will always happen in the direction of the global optimum, as otherwise it would constitute a local optimum. On the other hand in the absence of these rather strict requirements for $\mathcal{L}$ it is intuitive that this kind of optimization will fail.

From figure 6 it is clear that this simple method suffers from a variety of biases and divergence issues. However this was to be expected. The quick stabilization of estimates might be a sign that the algorithm very quickly approaches a local minimum, which works as an attractor, preventing convergence to the true value. A more formal proof that this is not purely coincidental is needed, but our preliminary testing does provide some circumstantial evidence of (biased) convergence.

6.2. Joint and nest based optimization

As an alternative to the naive method implemented above we also attempt a nest based solution, in which we formulate a likelihood based on $\Pr(i|m)$ instead of using $\Pr(i|C)$ as above. This is in essence the same as considering each choice as it's own α-augmented multinomial trial, and therefore requires information on the structural choices. Like with the iterative optimizer this approach is not ideal as coefficients are not restricted across nests, meaning we can get different β estimates for cross nested choices, depending on

the step of the optimization problem. In exchange the likelihood is significantly simpler, meaning estimation can happen in feasible time.

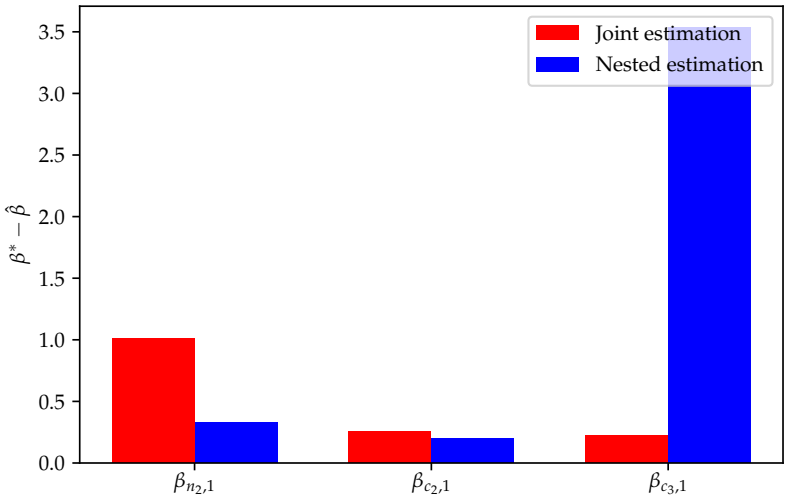


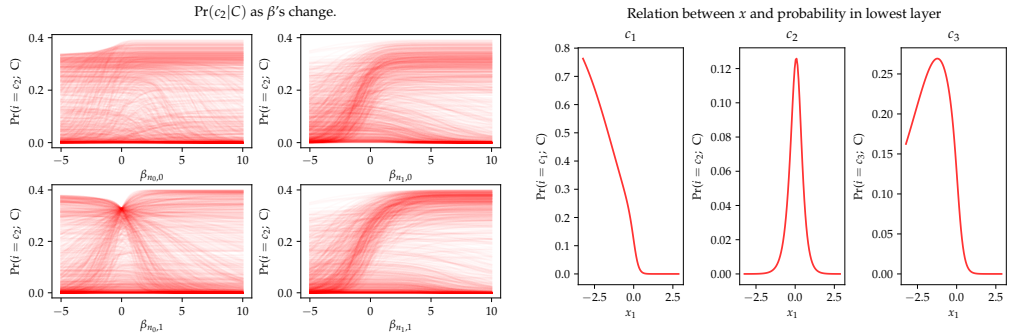
Figure 7: Estimate deviations $\beta^* - \hat{\beta}$ for the nest-based optimization, as well as a joint estimation using the full likelihood.

Figure 7 show parameters in deviations from their true values when applying the nest-based optimization, as well as (non-convergent) optimization results from attempting to optimize the full likelihood. In both cases we implement the same parameter restrictions as in section 6.1 for comparison. We take the results as yet another indication that these iterative optimizers are not ideal, but also note that even though the build in Scipy optimizer fails to converge, the results it find are not far from the true values of β. Actually the Scipy optimizer seems to be about as accurate as our iterative procedure, and the nest based method good only for some parameters. Thus slight improvements in the code might provide good estimates, granted the right parameter restrictions are implemented for identification.

6.3. Marginal effects in simulated data

Marginal effects are usually studied in relation to the regressors - that is the question is, if we alter a specific independent variable slightly, what will the resulting change in choice probabilities be. Small changes in specific β's are however perhaps even more interesting in a simulation setting, as it is the parameters that vary across choices in the multinomial CNL model. Panel (a) in 8 show for each of 1000 simulated individuals their probability of choosing the cross nested alternative c_3 as the β's related to the structural nodes n_1 and n_2 are varied (left and right column respectively). These naturally vary between individuals as the x-value is specific to them, meaning a change in β affects them differently. The apparent diversity in these curves for different values of β is due

to the complex changes in $P(i|\mathcal{C})$ when x is either positive or negative, as well as the changes occurring when $|x| \lesseqgtr 1$.



(a) Individual probabilities as function of β's

(b) Population effects of altering x_1 by a constant

Figure 8: Variations in $P(i|\mathcal{C})$ when altering β's and x's

The right panel on the other hand shows choice probabilities as a function of data, specifically x_1. These probabilities are naturally identical across the population for a given x. The normally considered marginal effects, $\partial p_i / \partial x_i$ is the derivative of these curves. Figure A.2 show a plot similar to the one in figure 8 panel (b), but calculated in the nested logit depicted in figure 4.

Naturally there are many more cross-effects than those shown here. These can easily be constructed by refering to the supplied code.[2] Figure A.1 show similar figures but for $\Pr(i|m)$, this shows clearly how when conditioning on the specific nest, the probability of making any of the choices are essentially inversely related as in the regular logit.

7. Application to real data

7.1. Estimation on DREAM data

In the following we estimate MNL, NL and CNL models on the DREAM data described in section 5.3. For the estimation we use Larch, which uses two optimization routines SLSQP and BHHH to estimate utility parameters as well as a number of nesting related parameters. There is no real consensus about whether structural nodes should be assigned linear utility or simply represent components of the error terms. Larch does not directly model structural utility, but instead lets the utility of choosing a specific nest

[2]The code will be made available on `https://github.com/Kristianuruplarsen` after the paper has been graded.

be[3]

$$V_m = \mu_m \left(\sum_j \alpha_{im} \exp\left(\frac{V_i}{\mu_m} \right) \right) \tag{7.1}$$

Where μ_m is then estimated. That is the independent utility from structural nodes, is simply the weighted sum of utilities available from children of the given nest. Furthermore to estimate the cross nested α parameter, Larch uses a logit-like link function[4] like

$$\alpha_{im} = \frac{\exp(\phi_i Z)}{\sum_j \exp(\phi_j Z)} \tag{7.2}$$

to avoid dealing with optimization constraints, while still ensuring $\alpha_{im} \in [0, 1]$. naturally this defaults to 0.5 when not specified, equivalent to setting $Z = 0$. In 3 we report a CNL model with the default setting of $Z = 0$ to simplify the results as much as possible. Results are however robust to the inclusion of this specification.

For identification we attempt a strategy similar to the one used in the iterative optimization routine, setting all parameters related to c_1 to 0 (we used the true values when simulating, but this is only to help our less sophisticated optimizers), and one parameter in each $c_i = 0$. This effectively normalize the utility of each choice, with a baseline of one, as well as normalizing the utility of each nest, also with a baseline of one.

Table 3 show estimated parameters in MNL, NL and CNL models on the DREAM data described in section 5.3. First note that for the NL model μ_m is estimated to 1, as would be expected since there is only one nest. In the CNL model we do not estimate any α's but instead estimate the logsum parameters to be 0.6 and 0.96 respectively. In all three models, the square of age is very close to 0, while age in itself is only really important in the NL model. In general it is difficult to compare these estimates directly. Like in the MNL all of these estimates are best interpreted by calculating marginal effects at the mean, but for both the NL and CNL models, these are complicated functions dependent on G's derivatives as we have shown in section 4.3.

Some authors (Hausman and McFadden, 1984; Koppelman and Bhat, 2006)) propose using likelihood ratio (LR), Wald, or Lagrange multiplier tests for model selection by comparing an estimated NL model to the MNL model that can be specified as a special case of the more general NL model. On the surface this might seem reasonable to apply to CNL models as well. We have shown the GEV framework nests both MNL, NL and CNL models, and that one way of representing this is trough the α matrix. However, using LR tests does not account for the fact that α's are co-linearly dependent, nor for the fact that the choice set $\mathcal{C}$ is altered when moving from e.g. a NL model to a MNL model. This alteration of the choice set, means the models are not algebraically nested as is required in the LR test. Instead MNL is only nested when we allow for set

[3] http://larch.readthedocs.io/en/latest/math/aggregate-choice.html
[4] http://larch.readthedocs.io/en/latest/example/111_cnl.html

	CNL		NL		MNL	
parameter	β	t-value	β	t-value	β	t-value
ASC_2	-0.412	-2.344	13.053	33.174	-0.578	-2.661
ASC_3	1.012	5.648	-88.537	-61.640	1.053	5.249
age_2	0.048	5.660	-0.581	-32.905	0.058	5.419
age_3	0.014	1.561	5.833	66.572	0.018	1.859
$agesq_2$	-0.000	-3.712	0.007	36.513	-0.000	-3.420
$agesq_3$	-0.000	-1.754	-0.091	-69.004	-0.000	-1.729
AC_2	-0.163	-3.739	-3.318	-55.419	-0.209	-3.838
AC_3	0.262	5.945	11.220	52.640	0.238	4.782
$Male_2$	0.016	0.769	-0.331	-10.520	0.020	0.761
$Male_3$	0.108	4.826	6.618	61.748	0.113	4.419
μ_1	0.755	32.108	1.000	-	-	-
μ_2	0.962	-1.738	-	-	-	-
Iterations	64		61		4	
AIC	-2.137		-4.882		1.866	
Optimizer	SLSQP		SLSQP, BHHH		BHHH	

Table 3: Optimization results, DREAM data. Subscripts denote the choice $c_i : i \in 2,3$ for which the parameter is calculated, except for μ where they identify nests.

operations.

The justification for using LR tests is weak as it does not take these considerations into account. Instead of LR values, we therefore calculate the Akaike Information Criterion[5], which does not require models to be nested.

We find that while AIC's are close, the lowest AIC is achieved by the NL model indicating that nesting c_1 ordinary unemployment benefits and c_2 being on sick leave, or various other benefits is the best of the three models. This is intuitive as the third category c_3, ordinary employment is a choice probably made by a different type of individuals than those who end up in either c_1 or c_2. One interpretation of these results is that there is selection between c_1, c_2 and c_3. In other word there seems to be some selection into the nest of unemployment among those who were on sick leave one year prior. Why this might be the case is left for future research, as this question is probably not suitable for answering within the GEV framework. The model does not give an answer as to whether this selection is driven by the individuals, as an alternative explanation could be that it is a sampling effect, where those in the nest share traits such as high age, making it difficult for them to get into employment and overcome severe sickness respectively. A bias might also arise from the effect that it is not solely up to the individual to decide on which benefits to receive, producing a high degree of substitution within the social benefit nest if some individuals are misplaced in the benefits system.

[5] $AIC = 2k - 2\ln(\hat{\mathcal{L}})$ with k being the number of parameters estimated and $\hat{\mathcal{L}}$ the maximum of the likelihood.

8. Perspectives

8.1. Results

We manage with some success to estimate CNL parameters even with low-performance estimation techniques, however the convergence of these methods is highly dependent on their initiation state, and the data at hand. For future work estimators must be implemented in a faster language, which is more suitable for heavy computation. Given that one does implement CNL in a better framework however, there are plenty opportunities for research, either into the interpretation of model output, or in the behavior in the model under various types of biases and data.

When applying the model to real data, we're faced with the issue of interpretability, and resort to the AIC for our conclusions. Using the Akaike Information Criterion we find some evidence of nesting in the choices of individuals who have been on sick leave, and in particular find that there is an "unemployment nest" in the choice structure. How robust these results are is difficult to say, given the complexity of the model, and future work could investigate the validity of conclusion drawn on the basis of NL and CNL models.

8.2. Usability and interpretability of CNL

While CNL will surely have applications in some areas of economics and other sciences, it is unlikely to gain widespread use, as estimation is restricted to a few software packages, which are relatively poorly documented and extremely difficult to modify. Implementing a maximum likelihood estimator of the likelihood is feasible for experts, but doing so will sacrifice transparency of methodology. A recent paper by Mai et al. (2017) suggests borrowing methodology from the field of dynamic programming for estimation, and development of pre-coded solutions for these methods can possibly increase accessibility to the GEV models.

Advances in computation will however not make interpretation of results easier, and from the figures 8 and from the marginal effects in equation (4.9) it is clear that marginal effects are not guaranteed to be monotone in neither parameters or regressors, making truthful reporting very difficult.

9. Conclusion

In this paper we have presented the Cross-nested Logit (CNL), through a thorough review of less complex models in the GEV family, and a GEV-oriented description of the features and limitations of the cross nested logit model. In this regard we contribute primarily known material. We also show a number of derivations, e.g. the marginal

effects in equation (4.9) and the properties for substitution patterns in section 3.3.1, which are perhaps to concrete to have been of interest in previous work, but has a very direct link to the ordinary logit models.

We simulate data from both nested and cross-nested structures, and implement two relatively naive estimation techniques as computational difficulties hinder a full likelihood estimation.

The paper also contains discussions on the usability of the cross nested models, and implements a simple model on real data from the Danish DREAM database. In this regard we find that while there is with no doubt lessons to be learnt from complex structural modelling. However the estimates suffer in interpretability and transparency, why we foresee some waiting before the CNL gains widespread use. We find evidence that nesting is a better model specification for the data (see section 5.3), than MNL suggesting that those on sick leave at time n will one year later act in accordance with a nested structure with a node for employment, and a nest covering various states of unemployment.

References

Arrow, Kenneth J. (1950). "A Difficulty in the Concept of Social Welfare". en. In: *Journal of Political Economy* 58.4, pp. 328–346. ISSN: 0022-3808, 1537-534X. DOI: 10.1086/256963. URL: https://www.journals.uchicago.edu/doi/10.1086/256963 (visited on 05/11/2018).

Ben-Akiva, Moshe E., Steven R. Lerman, and Steven R. Lerman (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*. en. Google-Books-ID: oLC6ZYPs9UoC. MIT Press. ISBN: 978-0-262-02217-0.

Bierlaire, Michel (1997). "On the overspecification of multinomial and nested logit models due to alternative specific constants". fre. In: *Transportation Science* 31.4, pp. 363–371. ISSN: 0041-1655. DOI: 10.1287/trsc.31.4.363.

– (2001). "A general formulation of the cross-nested logit model". In: vol. 1st.

– (2003). "BIOGEME: a free package for the estimation of discrete choice models". en. In: vol. 3rd, p. 27.

– (2006). "A theoretical analysis of the cross-nested logit model". en. In: *Annals of Operations Research* 144.1, pp. 287–300. ISSN: 0254-5330, 1572-9338. DOI: 10.1007/s10479-006-0015-x. URL: https://link.springer.com/article/10.1007/s10479-006-0015-x (visited on 03/22/2018).

– (2008). *Estimation of discrete choice models with BIOGEME 1.6*.

– (2016). "PythonBiogeme: a short introduction". In: 2016.

Bierlaire, Michel et al. (2009). "Estimation of discrete choice models: extending BIOGEME". In: *In Swiss Transport Research Conference (STRC*.

Cameron, A. Colin and Pravin K. Trivedi (2005). *Microeconometrics: Methods and Applications*. en. Google-Books-ID: TdlKAgAAQBAJ. Cambridge University Press. ISBN: 978-1-139-44486-6.

DREAM (2018). *The Danish Institute for Economic Modelling and Forecasting, DREAM*. URL: http://www.dreammodel.dk/default_en.html (visited on 05/17/2018).

Hausman, Jerry and Daniel McFadden (1984). "Specification Tests for the Multinomial Logit Model". In: *Econometrica* 52.5, pp. 1219–1240. ISSN: 0012-9682. DOI: 10.2307/1910997. URL: http://www.jstor.org/stable/1910997 (visited on 04/10/2018).

Hess, Stephane (2012). "A joint model for vehicle type and fuel type choice: evidence from a cross-nested logit study". fre. In: *Transportation* 39.3, pp. 593–625. ISSN: 0049-4488. DOI: 10.1007/s11116-011-9366-5.

Jong, Gerard de and Eric Kroes (2014). "Discrete Choice Analysis". en. In: *Analytical Decision-Making Methods for Evaluating Sustainable Transport in European Corridors*. SxI - Springer for Innovation / SxI - Springer per l'Innovazione. Springer, Cham, pp. 121–142. ISBN: 978-3-319-04785-0 978-3-319-04786-7. DOI: 10.1007/978-3-319-04786-7_8. URL: http://link.springer.com/chapter/10.1007/978-3-319-04786-7_8 (visited on 04/24/2018).

Koppelman, Frank S. and Chandra Bhat (2006). "Self Instructing Course in Mode Choice Modeling: Multinomial and Nested Logit Models". In:

Koppelman, Frank S. and Vaneet Sethi (2000). "Closed form discrete-choice models". In: *Handbook of transport modelling*, pp. 211–227. ISBN: 978-0-08-043594-7. URL: https://trid.trb.org/view/677899 (visited on 05/09/2018).

Luce, R. Duncan (1957). "A theory of individual choice behavior". PhD thesis. Columbia University. URL: http://www.dtic.mil/get-tr-doc/pdf?AD=AD0130718.

– (1958). "A Probabilistic Theory of Utility". In: *Econometrica* 26.2, pp. 193–224. ISSN: 0012-9682. DOI: 10.2307/1907587. URL: http://www.jstor.org/stable/1907587 (visited on 05/29/2018).

– (2005). *Individual Choice Behavior: A Theoretical Analysis*. en. Google-Books-ID: D74qAwAAQBAJ. Courier Corporation. ISBN: 978-0-486-44136-8.

Mai, Tien et al. (2017). "A dynamic programming approach for quickly estimating large network-based MEV models". In: *Transportation Research Part B: Methodological* 98, pp. 179–197. ISSN: 0191-2615. DOI: 10.1016/j.trb.2016.12.017. URL: http://www.sciencedirect.com/science/article/pii/S0191261515302216 (visited on 04/10/2018).

McFadden, Daniel L. (1973). "Conditional Logit Analysis of Qualitative Choice Behavior". In: *Frontiers in Econometrics*. New York: Academic Press, pp. 105–142.

– (1977a). *Modelling the Choice of Residential Location*. en. Tech. rep. 477. Cowles Foundation for Research in Economics, Yale University. URL: https://ideas.repec.org/p/cwl/cwldpp/477.html (visited on 04/03/2018).

– (2005). "Revealed Stochastic Preference: A Synthesis". In: *Economic Theory* 26.2, pp. 245–264. ISSN: 0938-2259. URL: http://www.jstor.org.ep.fjernadgang.kb.dk/stable/25055949 (visited on 04/10/2018).

McFadden, Daniel (1977b). *Quantitative Methods for Analyzing Travel Behaviour of Individuals: Some Recent Developments*. Cowles Foundation Discussion Paper 474. Cowles Foundation for Research in Economics, Yale University. URL: https://econpapers.repec.org/paper/cwlcwldpp/474.htm (visited on 05/30/2018).

Newman, Jeffrey P., Virginie Lurkin, and Laurie A. Garrow (2018). "Computational methods for estimating multinomial, nested, and cross-nested logit models that account for semi-aggregate data". en. In: *Journal of choice modelling* 26.C, pp. 28–40. URL: https://ideas.repec.org/a/eee/eejocm/v26y2018icp28-40.html (visited on 05/10/2018).

Papola, Andrea (2004). "Some developments on the cross-nested logit model". fre. In: *Transportation research. Part E, Logistics and transportation review* 38.9, pp. 833–851. ISSN: 0191-2615. DOI: 10.1016/j.trb.2003.11.001.

Richter, Marcel K. (1966). "Revealed Preference Theory". In: *Econometrica* 34.3, pp. 635–645. ISSN: 0012-9682. DOI: 10.2307/1909773. URL: http://www.jstor.org/stable/1909773 (visited on 04/10/2018).

Train, Kenneth E. (2009). *Discrete Choice Methods with Simulation*. en. Google-Books-ID: 4yHaAgAAQBAJ. Cambridge University Press. ISBN: 978-1-139-48037-6.

Williams, H. C. W. L. (1977). "On the Formation of Travel Demand Models and Economic Evaluation Measures of User Benefit". en. In: *Environment and Planning A: Economy and Space* 9.3, pp. 285–344. ISSN: 0308-518X. DOI: 10.1068/a090285. URL: https://doi.org/10.1068/a090285 (visited on 05/09/2018).

A. Figures

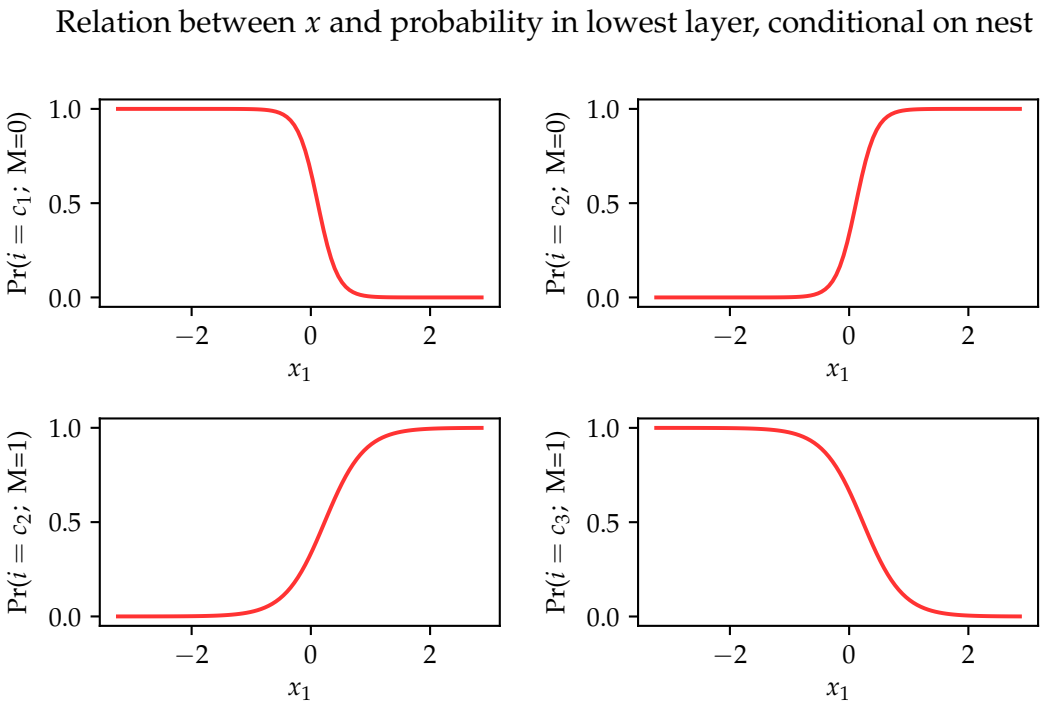


Figure A.1: Individual probabilities as function of β's, conditional on m

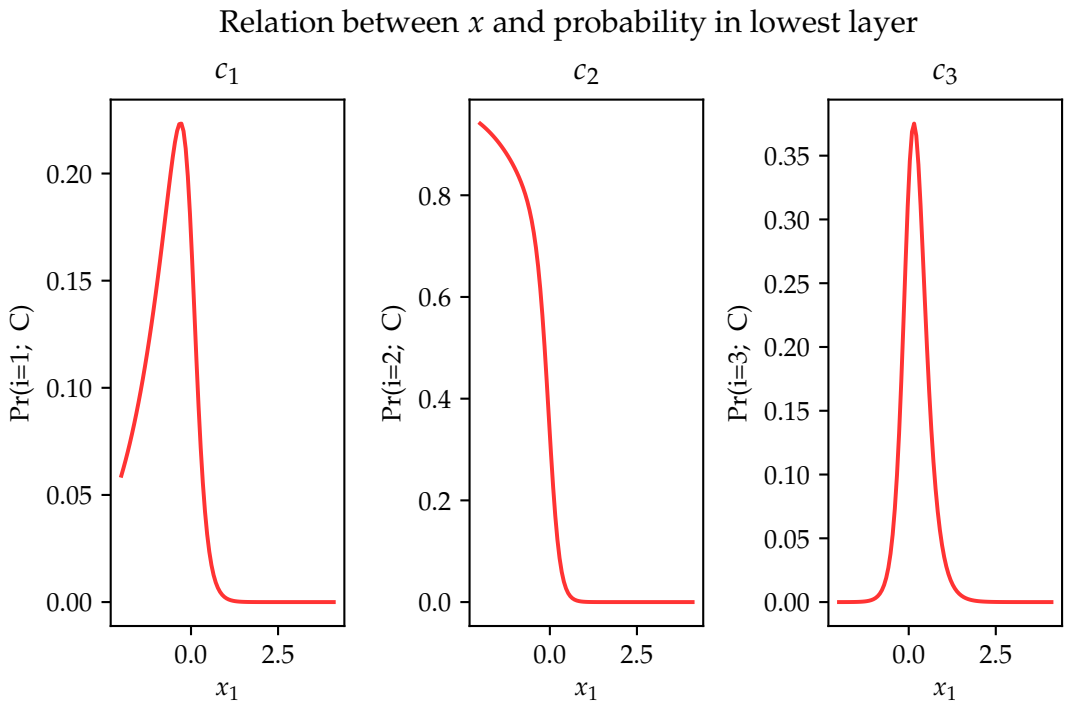


Figure A.2: Relation between x_1 and choice probabilities in the Nested Logit model

B. Commuting-example for the MNL model

Inspired by Koppelman and Bhat (2006) the starting point is the commuters choice between *car* and *bus* as the two possible means of transport in a choice set $\mathcal{C}^0$ for everyday commuting to work. Say that the commuter given her characteristics would choose to drive alone in a car with probability $\frac{2}{3}$ and to take the bus with probability $\frac{1}{3}$, thus, the ratio of probabilities is

$$\frac{\Pr(\mathcal{J} = \text{car}|\mathcal{C}^0)}{\Pr(\mathcal{J} = \text{bus}|\mathcal{C}^0)} = \frac{2/3}{1/3} = 2 \tag{B.1}$$

Now let us say a railway is being build and a local *train* line is introduced with several stations not far from the bus line used by the commuter. Furthermore the same ticket system applies for both of the public transport alternatives.

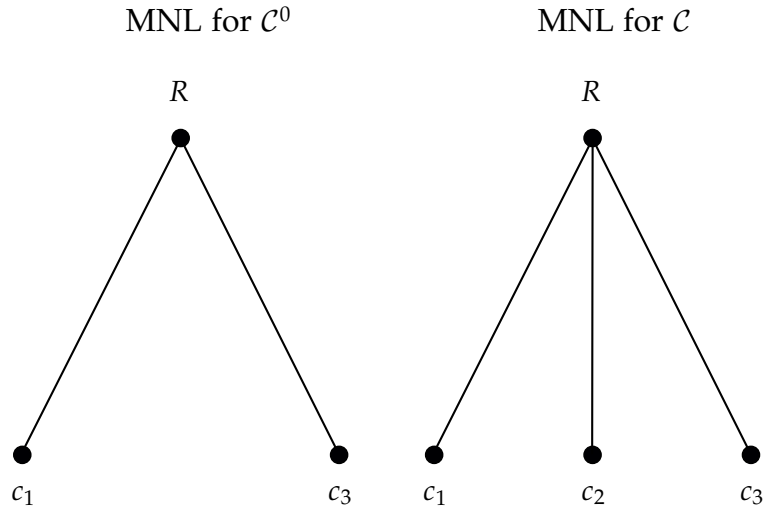


Figure B.3: Examples of Multinomial Logit models for two choice sets. For our analogy c_1 is bus, c_2 is train, and c_3 is car.

Looking at this new choice set $\mathcal{C}$ in figure B.3 let us first assume for simplicity that the commuter with a RUM-type utility function (3.1) would choose *train* and *bus* with equal probability as the two alternatives share somewhat similar characteristics (note though, that the probability of taking the train could be set to any number $\in]0, 1[$ for this example to carry through).

$$\Pr(\mathcal{J} = \text{train}|\mathcal{C}) = \Pr(\mathcal{J} = \text{bus}|\mathcal{C}) \tag{B.2}$$

In line with the IIA assumption the ratio of probabilities for (B.1) is kept constant, such that $\frac{2}{3}$ of the new train passengers took the car before and $\frac{1}{3}$ of the train passengers took the bus before as this was the distributions before the introduction of the train and IIA assumes equal competition. Take into account the probability of taking the *train* (B.2) and that the probabilities of the choices have to sum to one. The solution under these

three conditions is that the probabilities of the extended choice set $\mathcal{C}$ should be

$$\Pr(\mathcal{J} = \text{car}|\mathcal{C}) = \frac{1}{2}, \quad \Pr(\mathcal{J} = \text{bus}|\mathcal{C}) = \frac{1}{4}, \quad \Pr(\mathcal{J} = \text{train}|\mathcal{C}) = \frac{1}{4} \tag{B.3}$$

Thus, the MNL model predicts that the effect of adding another public transport option is that the probability of taking the *car* drops from $\frac{2}{3}$ to $\frac{1}{2}$ while the joint probability of taking *public transport* goes up from $\frac{1}{3}$ to $\frac{1}{2}$.

Implying that twice as many *train* passengers took the *car* before as opposed to the *bus* makes it obvious that IIA is a wrong assumption in this case. That is, the MNL model overestimates the joint probability of taking *public transport* due so keeping the prior ratio of probabilities between taking the car and taking the bus (B.1) constant despite the introduction of a train-alternative.

Deviating from the IIA-axiom, a more realistic assumption (though extreme too) could be that *bus* and *train* are such similar alternatives that the introduction of a train line does not affect the probability of taking the *car* but only takes over a share of the prior probability of taking the bus. The ratio of probabilities of the pair *(car, bus)* (B.1) would not be IIA in this case and would instead be expected to change.

C. Substitution patterns in nesting structures

A virtue of finding a nesting structure that resembles reality well is, that it can be used for policy advice as the choice structure implies to which alternatives the probabilities would shift for each possible alteration of the availability or attributes of each alternative in the choice set. In this section we propose three general properties for the ratio of probabilities in all possible nesting and cross-nesting structures.

Here we show how the ratio of probabilities can be used to examine the substitution patterns for a pair of alternatives in a specific choice structure. Furthermore we thoroughly use this methods to prove the properties **a.-e.** that are condensed into property $1. - 3.$ for Nested Logit and Cross-nested Logit models in section 3.3.1.

Inspired by Bierlaire (1997) and by using examples from the right side of figure C.4 we use the following terminology for the different nodes (alternatives) in the choice structure. The *parent node of* B_l *and* B_r is the nest B directly above the alternatives, while reversely the *children of* B are the alternatives B_l, B_r immediately belonging in the nest B. The complete choice structure consists of the root R, the *structural nodes* which is the set of the alternatives that are nests in themselves (B, B_l), and the *elemental nodes* which is the set of alternatives without children themselves (B_r, B_{ll}, B_{lr}). The *branch* following a structural node B would contain all the elemental and structural nodes that could be reached further down the tree after choosing B (i.e. B_l, B_r, B_{ll}, B_{lr}).

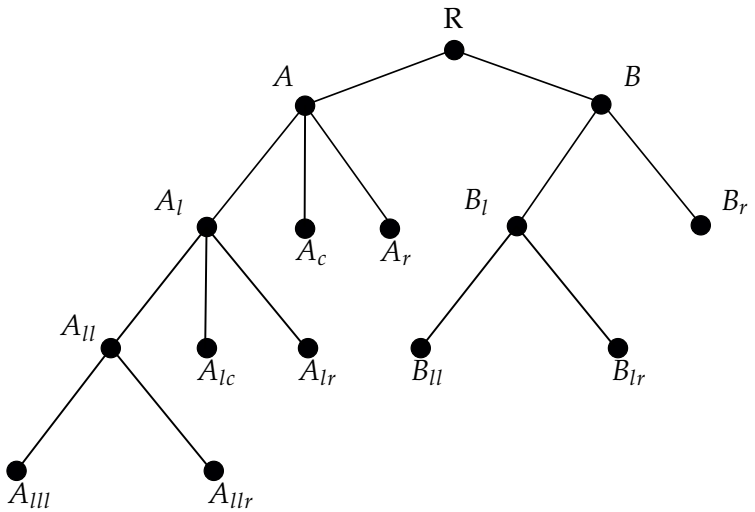


Figure C.4: Example of nesting in the actual choice structure.

As shown by Kenneth Train (2009) it holds for the Nested Logit (NL) model that

1. A pair of elemental nodes (i, l) within the same nest is IIA as their ratio of probabilities will be independent from any existence or modification of other alternatives.

2. A pair of alternatives (i, l) belonging to *different* nests is not IIA in general as the ratio of probabilities can depend on the alternatives in their respective nests.

To find out for which cases the relative probability for any pair of alternatives (e.g. A_{lc}, A_{lr} in figure C.4) is independent to different changes in the choice set, the simplest approach is to write up and analyze the ratio of probabilities for the pair by using the following slight rewriting of the probability function (3.23) for the NL in GEV-form. That is, it has the feature that one will never need to look at alternatives above the given nest A_l (or substitute in the whole path of conditional probabilities up till the root) as the probability of choosing A_{lc} in a given nest A_l over all other alternatives in the choice set $\mathcal{C}$ only depends on the attributes V_{A_lc} and V_j for the alternatives $j\mathcal{C}_{\Updownarrow} = A_{ll}, A_{lc}, A_{lr}$ that all belong to the same nest $A_l l$ and not on alternatives (Train, 2009). This works as utility maximizing under perfect information ensures that the utility of any possible alternative in the choice set is taken into account by the rational agent already at the root (McFadden, 1977a).

$$\Pr(\mathcal{J} = i|\mathcal{C}) \frac{e^{\mu_m V_i} \left(\sum_{j \in \mathcal{C}_m} e^{\mu_m V_j} \right)^{\frac{\mu}{\mu_m} - 1}}{\sum_{n \in \mathcal{M}} \left(\sum_{j \in \mathcal{C}_n} e^{\mu_n V_j} \right)^{\frac{\mu}{\mu_m}}} \tag{C.1}$$

As the probability function (C.1) for any two alternatives $i, l \in \mathcal{C}$ will have the same denominator then we only need the numerators for analyzing the ratio of probability for the pair. Letting i be a child of m_i and l of the nest m_l we have

$$\frac{\Pr(\mathcal{J} = i|\mathcal{C})}{\Pr(\mathcal{J} = l|\mathcal{C})} = \frac{e^{\mu_{m_i} V_i} \left(\sum_{j \in \mathcal{C}_{m_i}} e^{\mu_{m_i} V_j} \right)^{\frac{\mu}{\mu_{m_i}} - 1}}{e^{\mu_{m_l} V_l} \left(\sum_{j \in \mathcal{C}_{m_l}} e^{\mu_{m_l} V_j} \right)^{\frac{\mu}{\mu_{m_l}} - 1}} \tag{C.2}$$

If i and l are elemental nodes of the same parent nest (e.g. A_{lc}, A_{lr} in figure C.4) then $m_i = m_l$ and

$$\frac{\Pr(\mathcal{J} = i|\mathcal{C})}{\Pr(\mathcal{J} = l|\mathcal{C})} = \frac{e^{\mu_{m_i} V_i}}{e^{\mu_{m_l} V_l}} = \frac{e^{V_i}}{e^{V_l}} \tag{C.3}$$

Which is the result shown by Train (2009). We have written up his 1st property in a slightly elaborated version to underline that it holds for *elemental nodes* within the same nest while he seems to not regard nests as alternatives in themselves.

To disprove that IIA should holds in general for all pairs of alternatives within a nest including structural nodes, we again let i, l be children of the same nest but for i being a structural node and l an elemental node (e.g. A_{ll}, A_{lr} in figure C.4). While our result at first sight is identical to (C.3) we have to keep in mind that from (3.15) we have that a structural node l has the utility $V_l = W_m + \frac{1}{\mu_m}\Gamma_m$ where we from (3.16) have the logsum

utility for all alternatives in the nest below. If any of these are structural nodes the logsum utilities for all lower subnests should likewise be iteratively substituted into the joint utility of a structural node l. That we do not have IIA in general is to be expected from property 2. Though, it is clear from (C.2) and (C.3) respectively that a 3rd and 4th property can be added:

3. For any pair of alternatives i, l their ratio of probabilities is independent from all nodes n that are *at, next to, or prior to* the lowest structural node from which both i and l can be reached as well as from all nodes in branches of nodes n that do not reach i or l.

4. For a pair of alternatives i, l within the same nest m where at least one of them is a *structural node* their ratio of probabilities is independent from other alternatives within that nest, though, is not independent to any alternatives belonging to any branch following i or l.

E.g. in figure C.4 the ratio of probabilities for (A_{ll}, A_{lr}) is independent of all other elemental and structural nodes but for changes in A_{lll} or A_{llr}.

Next we analyze if under any circumstances the ratio of probabilities of i, l could be independent for i in nest m_i that is different from the nest m_l containing l. Applying these conditions to (C.2) we see that no terms go out. Even the share of an elemental node i within a structural node $m_i = l$ is only independent of alternatives at higher levels in the nesting structures, i.e. for i belonging to the nest m_i and $m_i = l$ belonging to the nest m_l. We see that property 2 actually holds such that i, l is not independent from any choices but the ones mentioned in property 3. For poperty 2 we are thus remminded that depending on alternatives in their respective nests also implies dependence on every alternative in a branch that starts in their nests.

Substitution patterns under cross-nesting

From equation (4.2) below we get that the equivalent to (C.2) for the *Cross-nested Logit* model where nodes are allowed to be cross-nested

$$\frac{\Pr(\mathcal{J}=i|\mathcal{C})}{\Pr(\mathcal{J}=l|\mathcal{C})} = \frac{\sum_m \alpha_{im_i} e^{\mu_{m_i} V_i} \left(\sum_{j \in \mathcal{C}_{m_i}} \alpha_{jm_j} e^{\mu_{m_i} V_j} \right)^{\frac{\mu}{\mu_{m_i}} - 1}}{\sum_m \alpha_{lm_l} e^{\mu_{m_l} V_l} \left(\sum_{j \in \mathcal{C}_{m_l}} \alpha_{jm_j} e^{\mu_{m_l} V_j} \right)^{\frac{\mu}{\mu_{m_l}} - 1}} \tag{C.4}$$

We find that we can add a 5th property such that for nesting structures where some nodes are allowed to be a part of several nests

5. In general property 1. and 4. holds only for alternatives i, l that are *not* cross-nested, i.e. $\forall n \in \mathcal{M} : \ \alpha_{n,i}, \alpha_{n,l} \in 0, 1$. Property 3. is violated if the branches

following structural nodes i or l contains a node that is crossed to a nest not in the branches following i or l-

Where in figure 2 the pair (A_{lc}, A_{lr}) nested in A_l with $\alpha_{A_l,A_{lc}} = \alpha_{A_l,A_{lr}} = 1$ is still IIA despite (partly) sharing nest with the cross-nested node c as (C.6) collapses to

$$\frac{\Pr(\mathcal{J} = A_{lc}|\mathcal{C})}{\Pr(\mathcal{J} = A_{lr}|\mathcal{C})} = \frac{\alpha_{A_l,A_{lc}} e^{\mu_{m_{A_l}} V_{A_{lc}}}}{\alpha_{A_l,A_{lr}} e^{\mu_{m_{A_l}} V_{A_{lr}}}} = \frac{e^{V_{A_{lc}}}}{e^{V_{A_{lr}}}} \tag{C.5}$$

The proof of property 4. is parallel to that of property 1. above, both conditioned on 5. It is trivial that property 2. and 3. conditional on property 5. holds for the CNL as for a cross-nested node like c its probability will be a sum with a term for every $\alpha > 0$ which in general hinders further reduction in the ratio of probabilities wrt. other alternatives.

We are able to come up with just one very specific special case where a pair of alternatives with at least one of them being cross-nested would be IIA. Both would need to be cross-nested to the exact same nests and with equivalent α-values. E.g. if A_{lr} like c was cross-nested to the nest B_l with $\alpha_{B_l,A_{lr}}$ then (C.6) would collapse to

$$\frac{\Pr(\mathcal{J} = c|\mathcal{C})}{\Pr(\mathcal{J} = A_{lr}|\mathcal{C})} = \frac{e^{V_c}}{e^{V_{A_{lr}}}} \tag{C.6}$$

D. Identification w.r.t "utility shifting"

In the nested logit the equations in (4.11) are independent and can therefore be solved one by one. We can show that there is a utility shifting potential, where utility is subtracted in the nest, and added in the nest-children in the ACS'es ψ by

$$\begin{aligned} V_m^{\psi} &= W_m - \psi + \frac{1}{\mu_m} \ln \sum_{j \in \mathcal{C}_m} e^{\mu_m V_j + \psi} \\ &= W_m - \psi + \frac{1}{\mu_m} \ln \sum_{j \in \mathcal{C}_m} e^{\mu_m (V_j + \psi)} \\ &= W_m - \psi + \frac{1}{\mu_m} \ln e^{\mu_m \psi} \sum_{j \in \mathcal{C}_m} e^{\mu_m V_j} \\ &= W_m - \psi + \psi + \frac{1}{\mu_m} \ln \sum_{j \in \mathcal{C}_m} e^{\mu_m V_j} \\ &= V_m \end{aligned} \tag{D.1}$$

In a similar fashion we can show that this problem also exists when considering scaling the β's in V_i. This is omitted for brevity but follows the exact same lines as the above.

In order to solve this issue in the NL model, all nests should therefore have their ACS restricted to a fixed value, such as 0.

E. On vectorization of cross nested models

There are a few different ways of implementing computationally heavy functions in high level languages such as R, Python etc. The go-to method is vectorization, but other alternatives, such as JIT compilation, libraries like Cython and code parallelization. In the creation of this paper we optet for the vectorization approach, as the Numpy library offers very simple and very powerful vectorization. However as cross nested logit allows cross nesting, this turn out to be problematic.

First consider that our data is generated as:

`for each individual` k`, for each nest` m `create a row,`
`indexed by` $c \in \mathcal{C}$`, for each option in` m

This entails that all cross nested choices $c_\times$ will occur once for each nest they're accessible from. Now note that calculating probabilities for simulation purposes involve the β matrix, which has one row per choice in the choice set $\mathcal{C}$. This is exactly the challenge: in the data we have more rows than there are choices, because of duplicates arising from cross nesting. In the parameter matrix we do not, yet these two must be multiplied together.

We handle this by also duplicating the relevant rows in β, and consequently "masking" them from the results by multiplying with a boolean vector indicating the individuals actual choice. This method, while valid, produces quite some overhead and is therefore to be avoided if possible. For future work we therefore suggest looking at Cython as a starting point.