Cross-nested logit models

Estimating complex network correlations

Thor Donsby Noe * Kristian Urup Larsen†

May 17, 2018

This paper investigates the usefulness of generalized extreme value models, and in particular the cross-nested logit model, in economics. We present the theoretical background required to understand the inner workings of the models, and derive marginal effects and a number of other useful measures. Like regular logit models the GEV models have closed form likelihood functions, and as such lend themselves to regular maximum likelihood estimation. As model complexity grows they will however require increasingly optimized code to feasibly evaluate and optimize. Likewise the number of constaints and identification issues rise rapidly with complexity, meaning specialized tools are the only time efficient way to estimate parameters in GEV models. Additionally the models complexity severely limits interpretability. 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 it's usefulness in applied econometrics and compare various specifications of GEV models.

To the discussant: Many of the details regarding our code are currently not correct as we're still trying to get our estimator working, overall things should however be about correct. If you have any insight into identification of the model, we'd really like to hear it. Also we've left our notes in the paper, so you know what we already know to improve.

*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 model with cross nesting, also simply known as the cross-nested logit model. The GEV framework is a flexible framework encompassing a broad range of discrete choice models which are all consistent with random utility theory and have closed form likelihood functions. So far GEV models, including the CNL model have received little attention by economists, perhaps primarily 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 of 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. An example is in the study of route choice - here roads, train lines etc. are naturally limiting the choice possibilities, but exactly how they affect decisionmakers is preferable left for the estimation to decide.

The CNL model is in short a structural model of sequential or nested choices allowing choice nodes partial membership of multiple nests. This allows the model large flexibility, but implies a large number of parameters restricted by both linear and nonlinear constraints. Thus estimating CNL models require imposing parameter restrictions, often ad hoc as knowledge of proper identification is lacking. We simulate GEV data and compare data generated by the nested- and cross-nested logit models. We find that the GEV models allow extreme heterogeneity in choice probabilities and have complicated marginal effects, but at the same time is conceptually very close to the regular multinomial logit model.

Lastly we estimate three choice models of varying complexity on data from the Danish DREAM database, and compare results from the different models.

2. Literature

The literature on cross nested logit models 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 identifiability are still to be addressed for the cross nested logit. 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 experimental data.

A large part of the existing literature is due to Michel Bierlaire (Bierlaire, 2006; Bierlaire, 2001; Bierlaire, 2003; Bierlaire, 2008; Bierlaire et al., 2009) who have contributed both by unifying the various formulations of the CNL that have been proposed (e.g. Bierlaire (2006)) and by developing the software package *Biogeme* (Bierlaire (2003)) for estimation of discrete choice models. Other work on the cross nested logit specifically include Papola (2004) but in general the literature due to others than Bierlaire has been superceded by Bierlaires attempts to unify the work of many others.

There is some literature on the estimation of complex network models, e.g. by Newman (2018) and Mai et al. (2017) however we don't 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 of nested models. The oldest of the free software packages is Biogeme which has now been extended to include the more user friendly pythonBiogeme for python 3. (Michel Bierlaire, 2016). A newer alternative, also made for python 3.6 is Larch, presented in Newman (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 and has a significant speed advantage over Biogeme.

The two options differ slighly in the ways data should be formatted, and in their optimization routines, but should in almost all cases be interchangable.

3. Theory

3.1. Random Utility Models (RUM's)

Common for all choice models related to the cross nested logit is that they're random utility models, that is models which forecast discrete choices based on the assumption that all individuals seek to maximize utility from the choices they make. 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 suitable for analyzing almost any human-choice-problem and advanced RUM's are heavily used in i.e. traffic research and route choice problems. Add: On "randomness", disadvantages, and the broad meaning of "utility" (McFadden, 2005), (Richter, 1966).

Because of the similarities between the econometric models and economic theory the random utility models are also used extensively in microeconomics.

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

$$V_{iq} = \beta_{i0} + \sum_{k=1}^{K} \beta_{ik} x_{iqk} \tag{3.2}$$

Note though, that whether $\beta_0 \neq 0$ and for that matter if the constant is allowed to vary between alternatives (subscript i) is dependent on the model specification. Determining the minimal necessary restrictions on parameters required for identification is not trivial in the nested and cross-nested models.

The deterministic model (3.2) is not able 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 (1977) shows that assuming a joint error distribution

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

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

3.2. Discrete choice structures

The discrete choice models taken into account in this paper predict 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. In econometric terms this is decided by the assumptions about the distribution of random error terms across the different alternatives and thus the cross-elasticities (Wen and Koppelman, 2001).

In this section we first take a look at the simple model and its assumption about equal competition among all pairs of alternatives, after which we relax this assumption and look at two different models that allows for cases where it is assumed that there is not equal competition among all pairs of alternatives.

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 multinomial logit model. Letting $\mathcal{J}$ be a stochastic variable with support on the choice set $\mathcal{C}$ of a 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 relative odds ratio 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).

Add a consideration about the Conditional Logit Model CL and the Mixed Logit model. The following general formulations of the Nested Logit model, Cross-nested Logit model and the the GEV-model class all contain alternative-specific regressors, though, the data we apply it to only contain individual-specific variables.

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 popuralized by R. Duncan Luce (1959) 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 odds ratio between two 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 (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 very strict IIA assumption is not fulfilled, i.e. when the inclusion of other alternatives is indeed not irrelevant.

One of the most classic examples of violation of the IIA assumption is classic "red bus/blue bus paradox" Koppelman and Bhat, 2006. The starting point is the commuters choice between car and a blue bus line as the means of transport for everyday commuting to work. Say that the commuter would take the car with probability $\frac{2}{3}$ and the blue bus with probability $\frac{1}{3}$, thus, the relative probability ratio is

$$\frac{\Pr(\mathcal{J} = \text{car}|\mathcal{C})}{\Pr(\mathcal{J} = \text{blue bus}|\mathcal{C})} = \frac{2/3}{1/3} = 2 \tag{3.6}$$

A competing bus company introduces a bus line on the same route with the only difference being that the bus is painted red instead of blue. Besides that the red bus line has the same characteristics as the blue bus line.

Looking at this new choice set $\mathcal{C}'$ it is first of all clear that the utility optimizing commuter with a RUM-type utility function (3.1) should choose blue bus and red bus with equal probability as they share the same characteristics

$$\Pr(\mathcal{J} = \text{blue bus}|\mathcal{C}') = \Pr(\mathcal{J} = \text{red bus}|\mathcal{C}') \tag{3.7}$$

In line with the IIA assumption the relative odds ratio (3.6) is kept constant. Take also into account the condition (3.7) and that the probabilities of the choices have to sum to one. The solution under these three conditions is that the the probabilities of the extended choice set C should be

$$\Pr(\mathcal{J} = \text{car}|\mathcal{C}') = \frac{1}{2},\ \Pr(\mathcal{J} = \text{blue bus}|\mathcal{C}') = \frac{1}{4},\ \Pr(\mathcal{J} = \text{red bus}|\mathcal{C}') = \frac{1}{4} \tag{3.8}$$

Thus, the MNL model predicts that the effect of adding another bus on the same route is that the probability of taking the car drops from $\frac{2}{3}$ to $\frac{1}{2}$ while the joint probability of taking any bus goes up from $\frac{1}{3}$ to $\frac{1}{2}$.

It is obvious that IIA is a wrong assumption in this case. The relative probability ratio between the probability of taking the car and taking the blue bus should not be kept constant when adding an alternative such as a red bus that is likely to be irrelevant to the probability of taking the car but certainly not irrelevant to the probability of taking the blue bus, thus the relative probability ratio (3.6) should not be expected to be constant.

3.2.3. The Nested Logit model (NL)

The more commonly used relaxation of the MNL model, and thus the IIA assumption, is the Nested Logit Model (NL) that was introduced by H.C.W.L. Williams (1977).

Alternatives are allocated to different nests where the IIA-axiom is assumed to be violated for the pair, while being nested together in cases where it is assumed that the IIA-axiom holds pairwise, e.g. for c_1, c_2 that both belong to a nest m_1 the relative probability of choosing c_1 over c_2 is assumed to be independent from whether other alternatives exist or possible attributes of these (Train, 2009). What can be a bit unclear in the litteraure is that the root R in itself is regarded as a nest, such that two alternatives, say c_3, c_4 that are both kept "unnested", similiarly to the alternatives in the MNL model, are actually both regarded as a direct child of the R-nest and thus the pair is assumed to be IIA. See the NL-tree in **figure 1** below for illustration.

Technically alternatives should be nested together when they share similar attributes which will lead to correlation of the error terms (Koppelman and Bhat, 2006). Thus the utilities of the nested alternatives c_1, c_2 in the nest m_1 are

$$\begin{aligned} U_{c_1} &= W_{m_1} + V_{c_1} + \varepsilon_{m_1} + \varepsilon_{c_1} \\ U_{c_2} &= W_{m_1} + V_{c_2} + \varepsilon_{m_1} + \varepsilon_{c_2} \end{aligned} \tag{3.9}$$

While the utility of the "unnested" alternatives c_3 are equal to that of the MNL model (3.1):

$$U_{c_3} = V_{c_3} + \varepsilon_{c_3} \tag{3.10}$$

As Train (2009) emphasizes $W_{m_1} = \beta_{m_1} x_{m_1}$ only depends on potential nest-specific regressors, i.e. x_{m_1} values that vary over nests but not over the alternatives in the nest m_1. Thus, while we include the deterministic nest utility W_{m_1} here for the general representation it will be set to zero in cases without observed information on the attributes of a nest like for the data we use for this paper. Likewise a W_R component could be added to all of the utility equations as they are all part of the root nest, but it would be 0 anyway.

Though the NL model makes no assumption about the actual order of the choice process, decomposing the NL model into two levels of equations lets us write it in a quite intuitive way (Jong and Kroes, 2014) where the probability of making a choice within a nest m is

$$\Pr(\mathcal{J} = m|\mathcal{C}) = \frac{\exp\{W_m + \frac{1}{\mu_m}\Gamma_m\}}{\sum_{n\in\mathcal{C}} \exp\{W_n + \frac{1}{\mu_n}\Gamma_n\}} \tag{3.11}$$

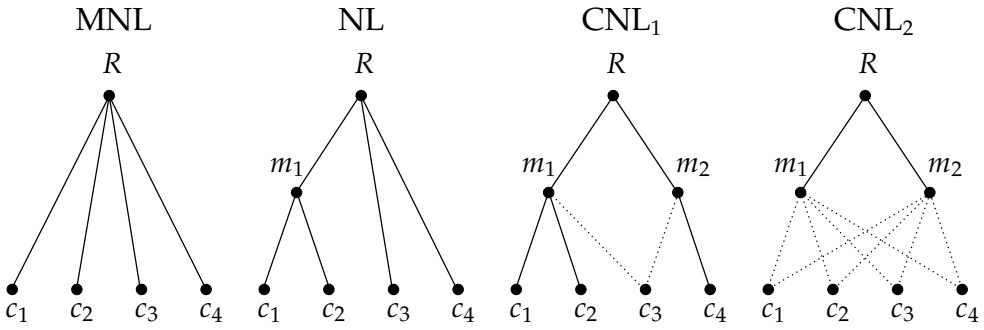

Figure 1: Examples of different choice models and structures for four choices.

3.2.4. The Cross-nested Logit model (CNL)

Preferable different assumptions about choice structures should be empirically tested and compared. The Cross-nested Logit model (CNL) allows for actual estimation of the cross-elasticities between each pair of alternatives and thus leaves less to be decided by the modeler.

3.3. The GEV model class

A generalized framework for models similar to the logit is the 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, 1977). The core of the class is a generating function G:

$$G(y_1, y_2, ..., y_J) \ \in \ \mathbb{R}^J_+ \tag{3.12}$$

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

1. $\mathcal{G}$ is nonnegative, differentiable and homogeneous of degree $\mu > 0$.
2. $\forall j \in \mathcal{J} : \ \lim_{y_j\to\infty} G(\cdots) = \infty, \mathcal{C} = 1, ..., J$
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.

Daniel 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

$$\Pr(\mathcal{J} = i|\mathcal{C}) = \frac{e^{V_i}\frac{\partial G}{\partial y_i}(e^{V_1}, e^{V_2}, ..., e^{V_J})}{\mu G(e^{V_1}, e^{V_2}, ..., e^{V_J})}, \quad i \in \mathcal{C} = 1, ..., J \tag{3.13}$$

where V_j is the deterministic element of utility, which is observed by the researchers, and/or parametrized by some known function. μ is the degree of homogeneity of $G(\cdot)$. How $G(\cdots)$ is then defined gives rise to a variety of models, including when $G(y_1, ..., y_J) = \sum_{j=1}^{J} y_j$ the regular multinomial logit model. However all functions $G \in \mathcal{G}$ are valid, and generalized specifications lead to the *nested-* and *cross-nested* logit models.

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

$$G(e^{V_1}, e^{V_2}, ..., e^{V_J}) = \sum_{J} e^{V_j}\frac{\partial G}{\partial e^{V_j}}(e^{V_1}, e^{V_2}, ..., e^{V_J}) \tag{3.14}$$

By redefining $z_i = e^{V_i}$ the probability in (3.13) can be reexpressed as

$$\Pr(\mathcal{J} = j|\mathcal{C}) = \frac{z_j\frac{\partial G}{\partial z_j}}{\sum_{i\in\mathcal{J}} z_i\frac{\partial G}{\partial z_i}} \tag{3.15}$$

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 multinomial logit 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 (3.15) as

$$\Pr(\mathcal{J} = j) = \frac{e^{V_j + \ln\frac{\partial G}{\partial z_j}}}{\sum_{i\in\mathcal{J}} e^{V_i + \ln\frac{\partial G}{\partial z_i}}} \tag{3.16}$$

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

3.4. Logit and nested logit in the GEV framework

The multinomial logit is perhaps the simplest of the GEV class models, and ss mentioned above setting the derivative of G equal to one collapses the expresion of $Pr(\mathcal{J} = j)$ to that from the multinomial logit. The reason for this is that setting $G_{\text{mult}}(\cdots) = \sum_{j\in\mathcal{J}} e^{V_j}$ gives exactly the multinomail logit since $\forall i : \ \frac{\partial}{\partial e^{V_i}} G_{\text{mult}} = 1$, whereby the expression in (3.16) collapses to multinomial logit model. Assuming this structure on G very clearly implements a fixed and equal relationship between all alternatives, and as a consequence yields *independence of irrelevant alternatives* across all $j \in \mathcal{J}$.

[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 $\sum_i z_i f'_{z_i}(\boldsymbol{z}) = qf(\boldsymbol{z})$

3.4.1. A first generalization - the nested logit model

Before jumping to the cross nested logit it is worth spending some time studying the slightly simpler nested logit model.

3.5. Cross nested logit

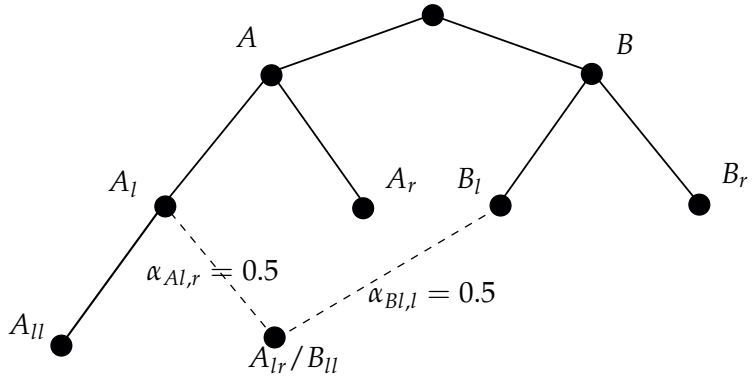

Figure 2: Example of cross nesting in the actual choice structure

In the cross nested logit, only one additional complication is added, namely the possibility of each choice being in multiple nests, such that they belong to each nest with some weight α_{jm}. 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

$$G(z_1, ..., z_J) = \sum_m \left(\sum_{j \in \mathcal{J}} \alpha_{jm} z_j^{\mu_m} \right)^{\frac{\mu}{\mu_m}} \tag{3.17}$$

where m is a nest-index of the set $\mathcal{M}$ of nodes in the graph, $\mathcal{J}$ 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 $\mathcal{J}$ 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 $j \in \mathcal{J}$, the summation is restricted to a set $\mathcal{J}_k \in \mathcal{J}$ of nests where $\alpha_{jm} \neq 0$.

One thing to note is that in the above specification each choice in $\mathcal{J}$ can belong to all nests at once. Other specification only sum over those choices that are nested deeper

in a decisiontree-like structure. While we keep the general 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 the cross nested structures in question, 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}.

In order for this function to satisfy the three criteria mentioned in section **??** a number of a priori restrictions have to be made on the parameters. 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 $G(tz_1, tz_2, ...tz_J) = \sum_m \left(\sum_j \alpha_{jm} z_j^{\mu_m} t^{\mu_m}\right)^{\frac{\mu}{\mu_m}} = t^{\mu} G(z_1, z_2, ..., z_J)$
- $\lim_{z_j \to \infty} G(\cdot) = \infty \ \forall j$ 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

$$\frac{\partial^k G(z)}{\partial z_{i_1}, ..., \partial z_{i_k}} = \sum_m \left(\mu_m^k \prod_{n \in \{i_1, ..., i_k\}} (\alpha_{nm} z_n^{\mu_m - 1}) \prod_{n=0}^{k-1} \left(\frac{\mu}{\mu_m} - n \right) y_m^{\frac{\mu - k\mu_m}{\mu_m}} \right) \tag{3.18}$$

  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 (3.20)). 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

$$\frac{\partial^k G(z)}{\partial z_{i_1}, ..., \partial z_{i_k}} \begin{cases} \geq 0, \text{ if } k \text{ is odd} \\ \leq 0, \text{ if } k \text{ is even} \end{cases} \tag{3.19}$$

This argument also makes it clear that while $\mu > \mu_m$ might produce a valid GEV generating function, it depends crucially on the relative size of the parameters.

3.5.1. Derivatives of G_i and marginal effects

In calculating the probability in (3.16) one need the derivative of (3.17). Fortunately we can easily derive this as

$$\begin{aligned}\frac{\partial G}{\partial z_i} &= \sum_m \frac{\partial}{\partial z_i}\left(\sum_{j\in\mathcal{J}} \alpha_{jm} z_j^{\mu_m}\right)^{\frac{\mu}{\mu_m}} \\ &= \sum_m \left[\frac{\mu}{\mu_m}\left(\sum_{j\in\mathcal{J}} \alpha_{jm} z_j^{\mu_m}\right)^{\frac{\mu}{\mu_m}-1} \cdot \alpha_{im}\mu_m z_i^{\mu_m-1}\right]\end{aligned} \tag{3.20}$$

This gives us an analytical gradient of G. To calculate marginal effects with respect to data x the second derivative $\frac{\partial G_i}{\partial x}$ 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

$$\begin{aligned}\frac{\partial^2 G}{\partial z_i \partial x} = \sum_m \frac{\mu}{\mu_m}\alpha_{im}\mu_m \Bigg[&\left(\frac{\mu}{\mu_m}-1\right)\left(\sum_j \alpha_{jm} z_j^{\mu_m}\right)^{\frac{\mu}{\mu_m}-2} \cdot \left(\sum_j \alpha_{jm} z_j^{\mu_m-1}\mu_m\beta_j\right)\cdot z_i^{\mu_m-1} \\ &+ \left(\sum_j \alpha_{jm} z_j^{\mu}\right)^{\frac{\mu}{\mu_m}-1} \cdot (\mu_m-1) z_i^{\mu_m-2}\beta_i\Bigg]\end{aligned} \tag{3.21}$$

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

$$\begin{aligned}\frac{\partial \Pr(i|\mathcal{C})}{\partial x} &= \frac{\frac{\partial}{\partial x}\left[e^{x\beta_i}\frac{\partial G}{\partial z_i}\right]\cdot\left(\sum_j e^{x\beta_j}\frac{\partial G}{\partial z_j}\right) - e^{x\beta_i}\frac{\partial G}{\partial z_i}\frac{\partial}{\partial x}\left(\sum_j e^{x\beta_j}\frac{\partial G}{\partial z_j}\right)}{\left(\sum_j e^{x\beta_j}\frac{\partial G}{\partial z_j}\right)^2} \\ &= \frac{e^{x\beta_i}\left(\beta_i\frac{\partial G}{\partial z_i} + \frac{\partial^2 G}{\partial z_i \partial x}\right)\cdot\left(\sum_j e^{x\beta_j}\frac{\partial G}{\partial z_j}\right) - e^{x\beta_i}\frac{\partial G}{\partial z_i}\sum_j e^{x\beta_j}\left(\beta_j\frac{\partial G}{\partial z_j} + \frac{\partial^2 G}{\partial z_j \partial x}\right)}{\left(\sum_j e^{x\beta_j}\frac{\partial G}{\partial z_j}\right)^2} \\ &= \beta_i \Pr(i|\mathcal{C}) + \frac{e^{x\beta_i}\frac{\partial^2 G}{\partial z_i \partial x}}{\sum_j e^{x\beta_j}\frac{\partial G}{\partial z_j}} - \Pr(i|\mathcal{C})\cdot\sum_j \beta_j \Pr(j|\mathcal{C}) + \frac{e^{x\beta_j}\frac{\partial^2 G}{\partial z_j \partial x}}{\sum_{j'} e^{x\beta_{j'}}\frac{\partial G}{\partial z_{j'}}}\end{aligned} \tag{3.22}$$

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

$$\frac{\partial \Pr(i|\mathcal{C})}{\partial x} = \Pr(i|\mathcal{C})\left(\beta_i - \sum_j\left[\beta_j \Pr(j|\mathcal{C}) + \frac{e^{x\beta_j}\frac{\partial^2 G}{\partial z_j \partial x}}{\sum_{j'} e^{x\beta_{j'}}\frac{\partial G}{\partial z_{j'}}}\right]\right) + \frac{e^{x\beta_i}\frac{\partial^2 G}{\partial z_i \partial x}}{\sum_j e^{x\beta_j}\frac{\partial G}{\partial z_j}} \tag{3.23}$$

3.6. The CNL model as a route choice model

Papola (2004) suggests a use for the CNL model that has applications in a wide range of problems beyond tree-like decision processes. Specifically he suggests a way to convert route-choice problems into CNL models. Consider the two networks shown in figure 3 and let the network (1) represent a road network connecting the origin O with the destination D. Furthermore let each of the arrows represent a road with nodes 2,3,4 being intersections. There are then four paths through the road-network, i.e. one could choose the highlighted path $O \to 2 \to D$. This network can then easily be reshaped into a cross nested structure by considering the four full paths from O to D instead of the individual road segments.

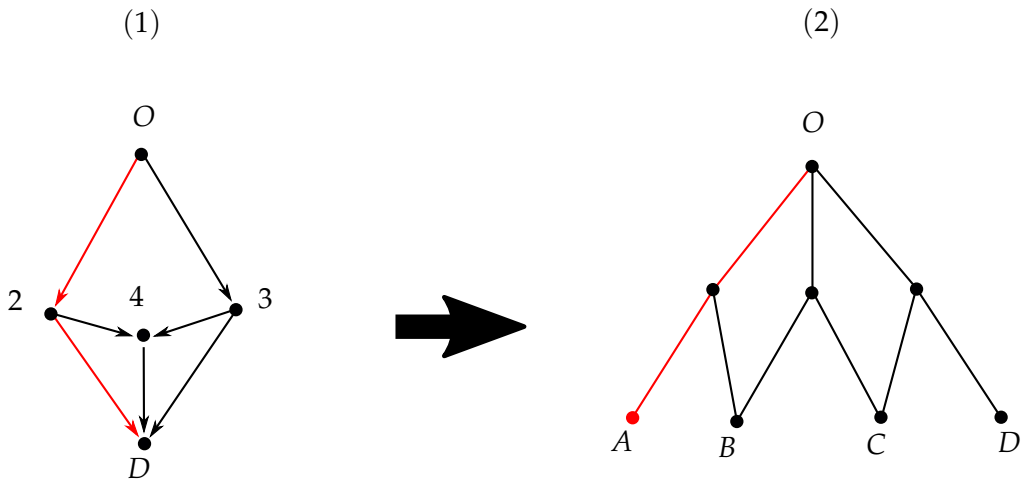

Figure 3: CNL for route choice modelling

This reformulation in terms of full paths is shown in (2) with choices A, B, C, D being a route from O to D. The underlying assumption is simply that choices only correlate if they share some part of their route. For example A and B share the road-segment $O \to 2$ and are therefore assumed to correlate. In this paper we do not particularly work with this type of network CNL's, but note that this interpretation is essentially valid for any CNL model, significantly broadening the interpretability and depth of the models.

3.7. 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 the utility is additively composted of known values V and noise ϵ. In the case of the logit model, the specification of this utility is straight forward as there is no sequentiality in choices. In the nested and cross-nested models, things are different however. Here individuals might derive utility from their choices along their 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 at structural

nodes. It can be shown that a consistent representation of the within-nest utilities of choices is

$$\log \sum_j \exp\{x\beta_j \cdot \mu_m\} \tag{3.24}$$

and the direct utility of a structural node can be parameterized simply as $W_m = x\beta_m$. Thus the total utility from choosing a structural node is

$$V_m = x\beta_m + \log \sum_j \exp\{x\beta_j \cdot \mu_m\} \tag{3.25}$$

In the following we don't explicitly distinct between strucutral and non-structural nodes, but always denote the derived utility V.

4. Estimation

4.1. Likelihood function

In order to derive the proper likelihood of observing data generated by a cross nested DGP, begin from the GEV definition of $\Pr(\mathcal{J} = i|\mathcal{C})$ given in equation (3.13) and G given in (3.17), to write

$$\begin{aligned}\Pr(\mathcal{J} = i|\mathcal{C}) &= \frac{z_i \left(\mu \sum_m \alpha_{im} z_i^{\mu_m - 1} \left(\sum_j \alpha_{jm} z_j^{\mu_m}\right)^{\frac{\mu}{\mu_m} - 1}\right)}{\mu \sum_n \left(\sum_j \alpha_{jn} z_j^{\mu_n}\right)^{\frac{\mu}{\mu_n}}} \\ &= \frac{\sum_m \alpha_{im} z_i^{\mu_m} \left(\sum_j \alpha_{jm} z_j^{\mu_m}\right)^{\frac{\mu}{\mu_m} - 1}}{\sum_n \left(\sum_j \alpha_{jn} z_j^{\mu_n}\right)^{\frac{\mu}{\mu_n}}}\end{aligned} \tag{4.1}$$

where n is a secondary nest index, as both $G(\cdot)$ and $G_i(\cdot)$ sums over all nests. Now taking one count of $\sum_j \alpha_{jm} z_j^{\mu_m}$ out of 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

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

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

$$\Pr(\mathcal{J} = i|\mathcal{C}) = \sum_m \Pr(m|\mathcal{C}) \Pr(i|m) \tag{4.3}$$

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.

Using the expression of probability given in (4.2) it is relatively simple to derive the likelihood. Letting (4.2) be the probability of observing an observation z_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

$$\mathcal{L}(\beta, \alpha, \mu_m, \mu | z) = \prod_{k=1}^{K} \Pr(\mathcal{J} = i | \mathcal{C})^{d_{ki}} \tag{4.4}$$

and as a direct extention thereof

$$\begin{aligned} \ln \mathcal{L}(\beta, \alpha, \mu_m, \mu | z) &= \sum_{k=1}^{K} d_{ki} \ln \Pr(\mathcal{J} = i | \mathcal{C}) \\ &= \sum_{k=1}^{K} d_{ki} \ln \left(\sum_m \frac{\left(\sum_j \alpha_{jm} z_j^{\mu_m} \right)^{\frac{\mu}{\mu_m}}}{\sum_n \left(\sum_j \alpha_{jn} z_j^{\mu_n} \right)^{\frac{\mu}{\mu_n}}} \times \frac{\alpha_{im} z_i^{\mu_m}}{\sum_j \alpha_{jm} z_j^{\mu_m}} \right) \end{aligned} \tag{4.5}$$

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 (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 doesn't 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. Conditions for a GEV generating function

The generating function of the CNL needs to fulfil several criteria in order to be a GEV generating function (equation 3.12). See Theorem 2, (Bierlaire, 2006)

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.

4.3. Identification

Due to the complexity of the generating function $G(z_1, ..., z_J)$ (eq.3.17) the exact set of necessary constraints for model identification is not known. Though, in the litterature there are proposed several different sufficient restructions for model estimation that are either complementary or substitutes.

Inserting the definition of $z_i = e^{V_i}$ into the probability function (4.2) we get:

$$\Pr(\mathcal{J} = i|\mathcal{C}) = \sum_m \frac{\left(\sum_j \alpha_{jm} e^{\mu_m V_j}\right)^{\frac{\mu}{\mu_m}}}{\sum_n \left(\sum_j \alpha_{jn} e^{\mu_n V_j}\right)^{\frac{\mu}{\mu_n}}} \times \frac{\alpha_{im} e^{\mu_m V_i}}{\sum_j \alpha_{jm} e^{\mu_m V_j}} \tag{4.6}$$

By imposing $V_j = (\beta_j^0 + \psi) + x\beta_j$ it can be shown that alternative specific constant (ASC) β_j^0 needs to be constrained in order to be identified as it otherwise can be arbitrarily extended by adding the same constant ψ to the utility of all alternatives without affecting the joint likelihood:

$$\Pr(\mathcal{J} = i|\mathcal{C}) = \frac{e^{\mu\psi}}{e^{\mu\psi}} \sum_m \frac{e^{\mu_m\psi}}{e^{\mu_m\psi}} \frac{\left(\sum_j \alpha_{jm} e^{\mu_m(\beta_j^0 + x\beta_j)}\right)^{\frac{\mu}{\mu_m}}}{\sum_n \left(\sum_j \alpha_{jn} e^{\mu_n(\beta_j^0 + x\beta_j)}\right)^{\frac{\mu}{\mu_n}}} \times \frac{\alpha_{im} e^{\mu_m(\beta_i^0 + x\beta_i)}}{\sum_j \alpha_{jm} e^{\mu_m(\beta_j^0 + x\beta_j)}} \tag{4.7}$$

Regarding the MNL model there exists a common strategy for solving this by arbitrarily fixing $\beta_j^0 = 0$ for one of the alternatives j (Ben-Akiva et al., 1985).

For the NL model there exist two traditional approaches for solving the ASC-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 (junctions of the three-structure besides the root). 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 three. Bierlaire (1997) introduces the "ortogonal strategy" where the "constraint subspace" contains all the necessary constraints and is choosen in order to be orthogonal to the "invariant subspace" that contains all the overspecification due to ASCs. The main virtues of the ortogonal strategy is the nice geometrical properties as well as removing the arbitrary selection, thus, both reducing 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). Bierlaire (1997) furthermore mentions that the overspecification for both the elemental and structural strategy is motivated by both ensuring a straightforward interpretation of the parameters as well as simplifying the estimation and thus ensuring convergence. Nonetheless, using more robust methods it is possible to ensure convergence even without restricting any ASCs, though, convergence can be very slow (Bierlaire, 1997).

Similarly to equation (4.7) when imposing $V_j = (\beta_j^0 + x(\beta_j + \delta)$ it can be shown

that there is a scalation problem and at least one of the parameters in the β_j-matrix needs to be constrained for the model to be identified as:

$$\Pr(\mathcal{J} = i|\mathcal{C}) = \frac{e^{\mu x\delta}}{e^{\mu x\delta}} \sum_m \frac{e^{\mu_m x\delta}}{e^{\mu_m x\delta}} \frac{\left(\sum_j \alpha_{jm} e^{\mu_m(\beta_j^0 + x\beta_j)}\right)^{\frac{\mu}{\mu_m}}}{\sum_n \left(\sum_j \alpha_{jn} e^{\mu_n(\beta_j^0 + x\beta_j)}\right)^{\frac{\mu}{\mu_n}}} \times \frac{\alpha_{im} e^{\mu_m(\beta_i^0 + x\beta_i)}}{\sum_j \alpha_{jm} e^{\mu_m(\beta_j^0 + x\beta_j)}} \tag{4.8}$$

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 (4.2) 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 when assuming away the added complexity 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, (4.2) 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 dataset of this kind.

This section makes no sense. Fix when the code is closer to completion

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. This yields full paths through the graph. In essence our approach is the same as in the nested case, that is we simulate each nest as a separate event, but in the cross-nested case we 'bias' the multinomial probabilities with α's. This approach will however produce impossible paths as each individuals choice is simulated in every nest. For that is we will 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.

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.

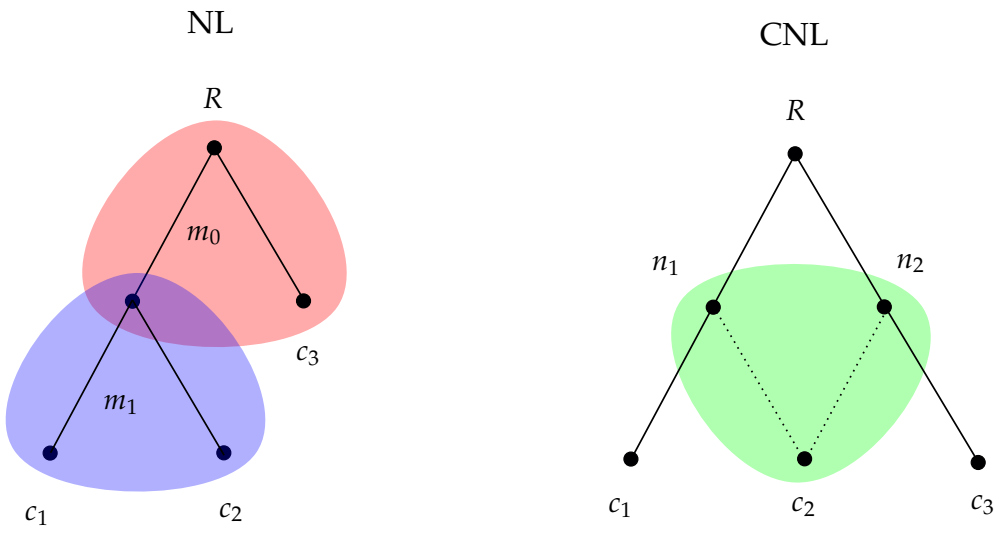

Figure 4: The nesting structures employed as simulated data

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, which we handle by simulating five independent vectors $x_1, ..., x_5$ and (somewhat arbitrarily) combining them in pairs of three, so that there is a total of 5 equations (4 in the NL case) of the form

$$V_j = x\beta_j \tag{5.1}$$

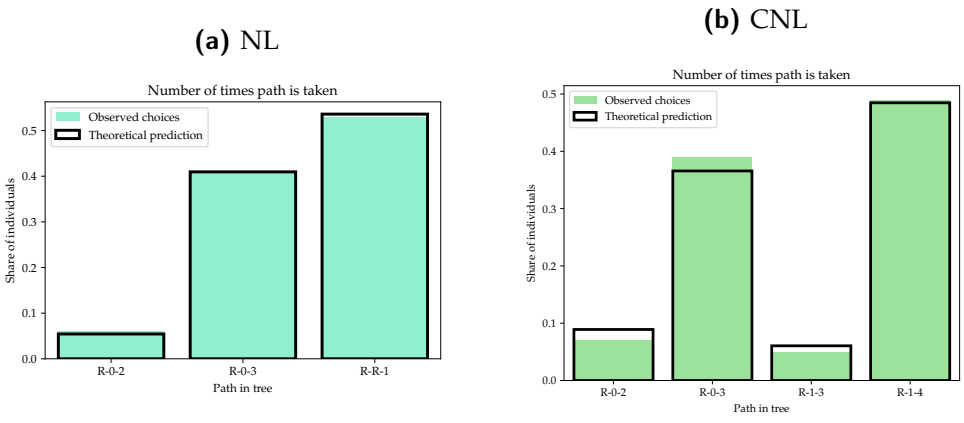

Figure 5: Actual and theoretical choice probabilities

By simulating and selecting data as described above we construct a dataset of 1000 "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

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

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 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.

Is this where we include a reference to github where our code can be downloaded ?

5.3. 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 6 show for each of 100 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 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

(a) Individual probabilities as function of β's

(b) Population effects of altering x_1 by a constant

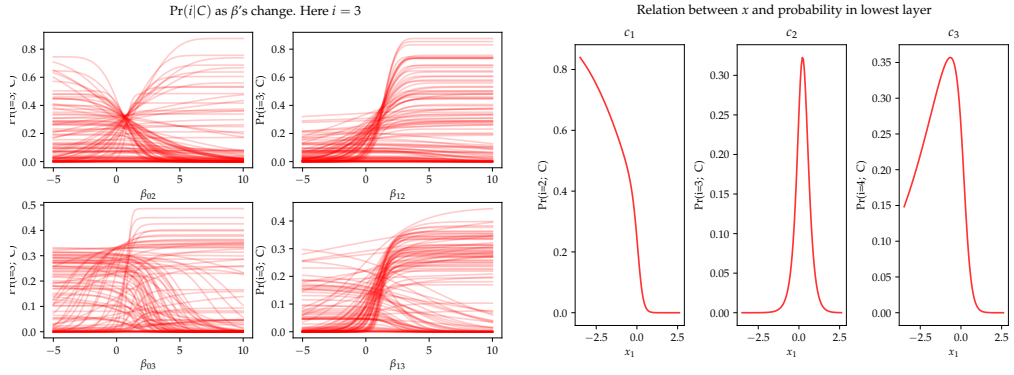

Figure 6: Variations in $P(i|C)$ when altering β's and x's

curves. Figure A.2 show a plot similar to the one in figure 6 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 Github repository. 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.

5.4. 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*, DREAM. More specifically the december 2017-version of the database that lays the foundation for their socioeconomic projection model where all of the population week-by-week is subdivided labour market categories.

For simplicity we model the one-year course for the group of 84.123 persons on sick leave that received unemployment benefits in week 48 of 2016 (last week of november). Thus, the selection is already 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 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 that in week 48 of 2017 (last week of november) are on either paternity leave, ordinary early retirement ("efterløn"), ...

We end up with 67.516 observations.

The choices are

6. Code

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.

mention that this all assumes knowledge of the structural nodes, which is not normally assumed.

We implement reasonably fast versions of the probability equations in 4.1, 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 1. 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 $J \times N_x$ parameters jointly fails completely when using the standard SciPy minimization routine with BFGS. This should be expected as BFGS assumes the optimization problem is unconstrained, which is not the case when solving GEV models. Instead we attempt SLSQP[2] which allows for both linear and non-linear constraints. But we have to understand how this works before moving forward

Normalize these times to make it less obviously system dependent

Table 1: Speed comparison single evaluation of $\log \mathcal{L}_K$ at different data sizes

	K=10	K=50	K=100	K=1000	K=5000
0	0.855978	0.992007	0.989961	1.009362	0.961808

Numbers are means 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 are 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.

[2] *Sequential (Least Squares) Quadratic Programming* - this is also the optimizer used by Larch.

6.1. Iterative estimation

The perhaps simplest way of estimating the model parameters is following 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 each choice (with some set to 0), so we can arrange β as a matrix of size $J \times N_x$, (Note that the fact that there are as many x'es as there are choices is coincidental.)

Update all of this when you know the final dimensions of stuff

$$\beta = \begin{array}{c} \\ c_0 \\ c_1 \\ c_2 \\ c_3 \\ c_4 \end{array} \begin{array}{c} \begin{array}{ccccc} b_0 & b_1 & b_2 & b_3 & b_4 \end{array} \\ \begin{pmatrix} \beta_{00} & \beta_{01} & \beta_{02} & \beta_{03} & \beta_{04} \\ \beta_{10} & \beta_{11} & \beta_{12} & \beta_{13} & \beta_{14} \\ \beta_{20} & \beta_{21} & \beta_{22} & \beta_{23} & \beta_{24} \\ \beta_{30} & \beta_{31} & \beta_{32} & \beta_{33} & \beta_{34} \\ \beta_{40} & \beta_{41} & \beta_{42} & \beta_{43} & \beta_{44} \end{pmatrix} \end{array} \tag{6.1}$$

Our iterative optimization scheme then initialized $\beta^{\text{init}} = \mathbf{1}$, sets β_{i0}^{init} to their true parameters for all i (this is required for identification) and then iteratively optimize each element of β assuming all other elements fixed. After each iteration β is updated to include the found optimal element. 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

```
for b in [b_0, b_1, b_2, b_3, b_4]{
    for c in [c_0, c_1, c_2, c_3, c_4]{
      β*_[b][c] = OptimizeScalar(log L_K(β, x, Θ) w.r.t β_[b][c])
      β_[b][c] = β*_[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 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) 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.

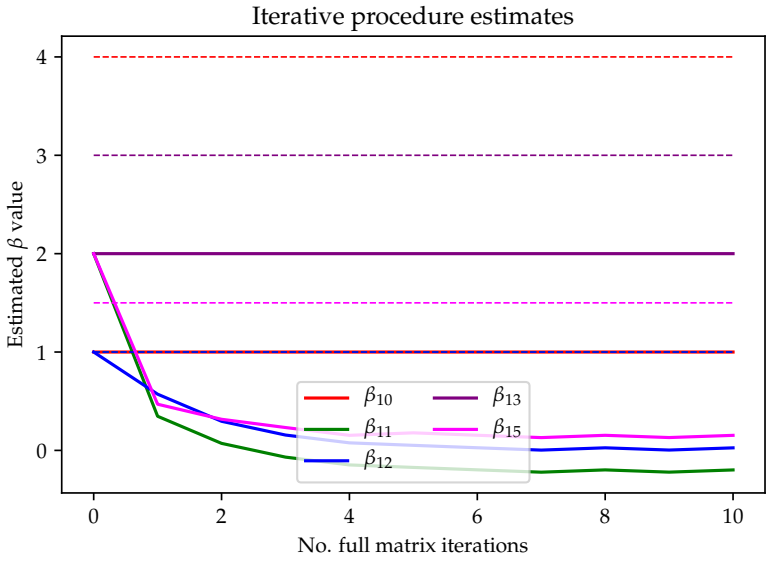

Figure 7: Parameter development using iteration. Dashed lines are true parameter values.

From figure 7 it is clear that this simple method suffers from a variety of biases and divergence issues. However this was to be expected, and the method is surprisingly good for some parameters. 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. Nest based optimization

7. Application to DREAM data

In the following we estimate MNL, NL and CNL models on the DREAM data described above. For the estimation we use Larch, which uses two optimization routines SLSQP and BFGS to estimate utility parameters as well as a number of nesting related parameters. As mentioned there is no real consensus about whether structural nodes should be assigned linear utility or simply represent components of the error terms. Larch does

Table 2: Optimization results, DREAM data

	MNL		NL		CNL	
	Final value	StdError	Final value	stdError	Final value	StdError
ASC2	-0.522772	0.215321	-0.522687	-8.77458	0.686658	-0.0903655
age2	0.0544355	0.0105717	0.0544319	-0.13285	0.000269098	-0.00380324
agesq2	-0.000389638	0.000124591	-0.000389602	-0.000824869	-2.26028e-06	-3.18926e-05
ASC3	1.10506	0.198562	1.1051	-0.582595	1.31448	-0.299703
age3	0.0192727	0.00984581	0.0192708	-0.046215	-0.0134031	-0.00316774
agesq3	-0.000217209	0.000116917	-0.000217191	5.76122e-05	1.36615e-05	5.66309e-05
a1			1	-4.92115	0.00286446	-0.0416262
a2					0.739591	-0.337225
phi					2.32581e-20	1.8819e-11
Iterations	4		33		106	
Optimizer	BHHH		SLSQP		SLSQP, BHHH, SLSQP	

not directly model structural utility, but instead lets the utility of choosing a specific nest be[3]

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

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

$$\alpha_{im} = \frac{\exp(\phi_i Z)}{\exp(\phi_i Z) + \exp(\phi_{-i} Z)} \tag{7.2}$$

to avoid dealing with optimization constraints, while still ensuring $\alpha_{im} \in [0, 1]$. naturally this defaults to 0.5 when not specified.

Table 2 show estimated parameters in MNL, NL and CNL models on the DREAM data described in section 5.4. First of it is interesting how similar parameter estimates are in the MNL and NL models, while standard errors vary significantly. why?. Of course 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.

8. Conclusion

- Computational issues require specialist knowledge
- low interpretability
- high flexibility
- etc

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

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*.

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

Cameron, Adrian Colin and P. K. Trivedi (2005). *Microeconometrics: methods and applications*. Cambridge ; New York: Cambridge University Press. ISBN: 978-0-521-84805-3.

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 (1959). *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. (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 (1973). "Conditional Logit Analysis of Qualitative Choice Behavior". In: *Frontiers in Econometrics*. New York: Academic Press, pp. 105–142.

– (1977). *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).

Michel Bierlaire (2016). "PythonBiogeme: a short introduction". In: 2016.

Newman, Jeffrey P. (2018). "Computational methods for estimating multinomial, nested, and cross-nested logit models that account for semi-aggregate data". eng. In: *Journal of choice modelling* 26, pp. 28–40. ISSN: 1755-5345. DOI: 10.1016/j.jocm.2017.11.001.

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.

Wen, Chieh-Hua and Frank S Koppelman (2001). "The generalized nested logit model". In: *Transportation Research Part B: Methodological* 35.7, pp. 627–641. ISSN: 0191-2615. DOI: 10.1016/S0191-2615(00)00045-X. URL: http://www.sciencedirect.com/science/article/pii/S019126150000045X (visited on 04/10/2018).

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. Appendix A

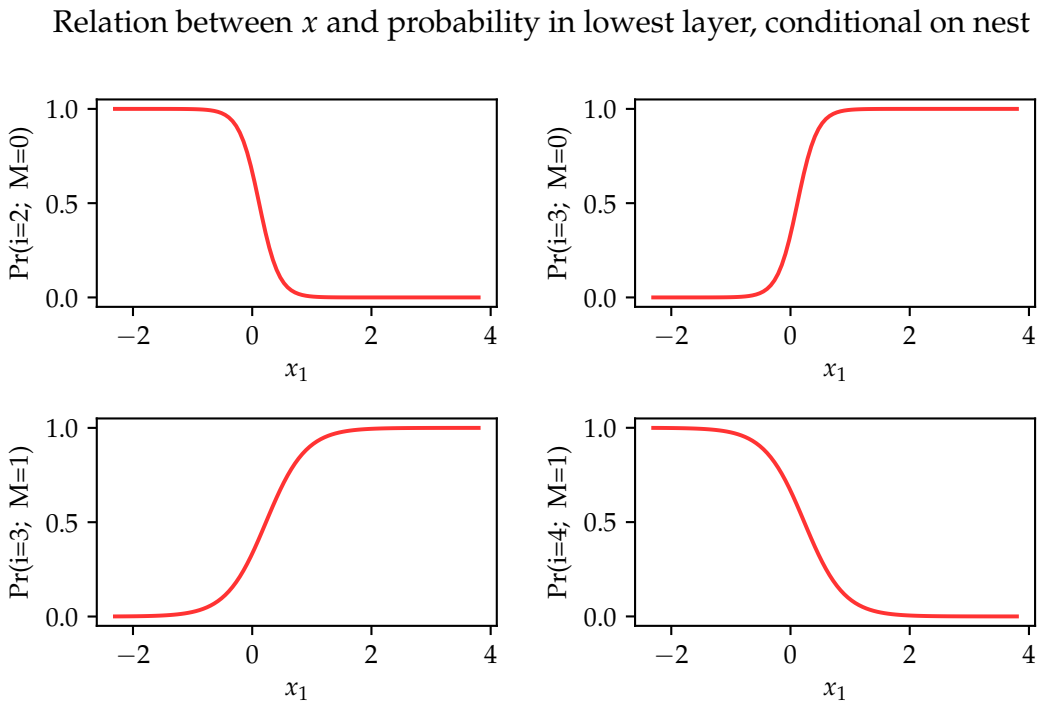

Figure A.1: Individual probabilities as function of β's, conditional on m

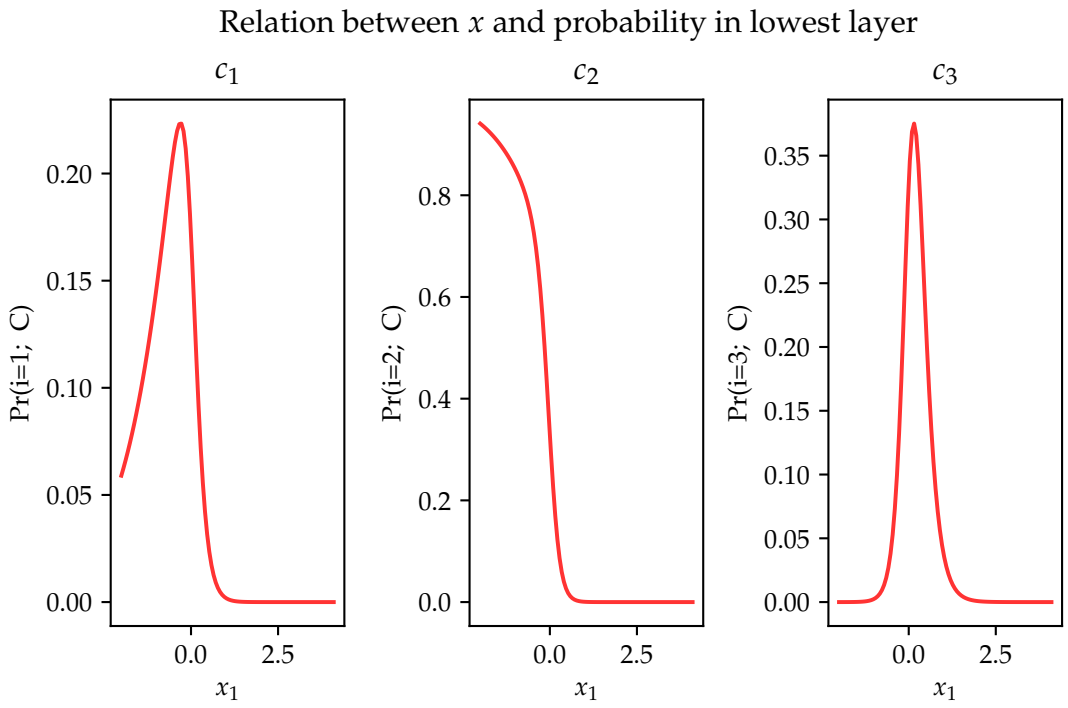

Figure A.2: Relation between x_1 and choice probabilities in the Nested Logit model