Sampling of Alternatives in Multivariate Extreme Value (MEV) Models

by

CRISTIAN ANGELO GUEVARA
Universidad de los Andes, Chile
Facultad de Ingeniería y Ciencias Aplicadas
San Carlos de Apoquindo, 2200, Las Condes, Santiago, Chile
Tel:56-2-412-9621
Fax: 56-2-412-9642
aguevara@uandes.cl

and

MOSHE BEN-AKIVA
Massachusetts Institute of Technology
Department of Civil & Environmental Engineering

SUMMARY

When the number of alternatives in a choice-set is huge, sampling is unavoidable. In 1978 Daniel McFadden showed that consistent estimation under sampling of alternatives is possible if the true model is Logit; that is, if the errors of the random utilities are independent and identically distributed (*iid*) Extreme Value. However, the *iid* assumption might be easily broken in models with large choice-sets. For example, in residential location, dwelling-units are expected to be correlated depending on proximity. This paper extends McFadden's result to MEV models, a class of closed-form discrete choice models that allows for different degrees of correlation between alternatives. A methodology to achieve consistency, asymptotic normality and relative efficiency is proposed and deployed for all MEV models and then illustrated using Monte Carlo experimentation and real data for the Nested Logit model, an important member of the MEV class. Experiments show that the proposed methodology is practical, that it is substantially better than an uncorrected model, and that it yields acceptable results, even for small sample sizes. The paper finishes with a synthesis and an analysis of the impact, limitations and potential extensions of this research.

Key Words: Sampling of Alternatives, MEV, GEV models.

1 Introduction

The computational burden and the impossibility of identifying or measuring the attributes of a huge number of alternatives in spatial choice models, makes it necessary to only consider a subset of the choice-set in practical applications. McFadden (1978) demonstrated that if the model underlying the choice process is Logit, the problem of sampling of alternatives and estimation can be addressed by adding a corrective constant to the systematic utility of each alternative.

The Logit model requires the assumption that the error terms of the random utilities are uncorrelated among alternatives. This assumption may be invalid for some spatial choice models. In residential location, the error terms may be correlated according to proximity or nested according to different decision levels. Equivalently, in route choice modeling, routes that share sets of common links may be perceived as more similar than other routes that are complete substitutes, breaking from the Logit assumption.

Ignoring a non-Logit structure in spatial choice modeling may significantly impact the quality of such models. For example, if the underlying model is a Nested Logit with nests defined by geographical areas, a location subsidy will trigger more intra-area than inter-area household relocation. This effect would be impossible to capture with a Logit model, resulting in misleading guidance for urban policy analysis. This suggests extending McFadden's result on sampling of alternatives in Logit models to a more general class of models that allows correlation among the error terms of the utilities.

Building on an idea originated by Ben-Akiva[1] (2009), in this paper, we extend McFadden's results to the Multivariate Extreme Value (MEV) models, a class of closed-form discrete choice models that allows for certain degrees of correlation among alternatives. We also use Monte Carlo experimentation and real data to show the impact of the application of this novel method in the estimation of Nested Logit models while sampling alternatives.

The paper is structured as follows. The next section describes McFadden's results on sampling of alternatives in Logit models. Next, the proposed extension to MEV models is presented. The following sections describe the formulation of the proposed methodology for the Nested and the Cross-Nested Logit models, the main members of the MEV family. Then, the effects of the proposed methodology are analyzed using a Monte Carlo experiment and real data on residential location choice from Lisbon, Portugal. The final section summarizes the main conclusions, implications, and potential extensions of this research.

2 Estimation and Sampling of Alternatives in Logit Models

Consider that the random utility U_{in}, which a household n retrieves from alternative i, can be written as the sum of a systematic part V and a random error term ε, as shown in Eq. (1)

$$U_{in} = V_{in} + \varepsilon_{in} = V(x_{in}, \beta^*) + \varepsilon_{in}, \quad (1)$$

where the systematic utility depends on variables x and parameters β^*.

[1] Unpublished Manuscript.

Then, if ε is independent and identically distributed (*iid*) Extreme Value (0,μ), the probability that *n* will choose alternative *i* will correspond to the Logit model shown in Eq. (2)

$$P_n(i) = \frac{e^{\mu V_{in}}}{\sum_{j \in C_n} e^{\mu V_{jn}}}, \tag{2}$$

where C_n is the choice-set of J_n elements from which household *n* chooses an alternative. The scale μ in Eq. (2) is not identifiable and usually normalized to equal 1.

Consider that, of the true choice-set C_n, only a subset D_n with $\tilde{J}_n$ elements is sampled by the researcher. For estimation purposes, D_n must include (and therefore depends on) the chosen alternative *i* because, otherwise, the quasi-log-likelihood of the model may become unbounded, making the estimation of the model parameters impossible.

Term $\pi(i, D_n)$ the joint probability that household *n* would chose alternative *i* and that the researcher would construct the set D_n. Using the Bayes theorem, this joint probability can be rewritten as shown in Eq. (3)

$$\pi(i, D_n) = \pi(D_n \mid i) P_n(i) = \pi(i \mid D_n) \pi(D_n), \tag{3}$$

where $\pi(i \mid D_n)$ is the conditional probability of choosing alternative *i*, given that the set D_n was constructed, and $\pi(D_n \mid i)$ is the conditional probability of constructing the set D_n, given that alternative *i* was chosen.

Since the events of choosing each one of the alternatives in C_n are mutually exclusive and totally exhaustive, we can use the Total Probability theorem (see, e.g., Bertsekas and Tsitsiklis, 2002) to write the probability $\pi(D_n)$ of constructing the set D_n as shown in Eq. (4)

$$\pi(D_n) = \sum_{j \in C_n} \pi(D_n \mid j) P_n(j) = \sum_{j \in D_n} \pi(D_n \mid j) P_n(j), \tag{4}$$

where the second equality holds because $\pi(D_n \mid j) = 0 \; \forall j \notin D_n$.

Substituting Eq. (4) and the Logit choice probability $P_n(i)$ shown in Eq. (2) into Eq. (3), Eq. (5) is obtained by canceling and re-arranging terms.

$$\pi(i \mid D_n) = \frac{e^{V_{in} + \ln \pi(D_n \mid i)}}{\sum_{j \in D_n} e^{V_{jn} + \ln \pi(D_n \mid j)}} \tag{5}$$

The expression $\ln \pi(D_n \mid j)$ is termed the sampling correction.

Eq. (5) indicates that the conditional probability of choosing alternative *i*, given that a particular choice-set D_n was constructed, depends only on the alternatives in D_n. This results from the cancellation of the denominators when dividing the probabilities of two alternatives in the Logit model, which is known as the Independence of Irrelevant Alternatives (IIA) property. Note that although IIA is a convenient mathematical property, it results from the assumption that the error structure is *iid*, a statement that may be unrealistic in spatial choice models.

McFadden (1978) demonstrated that if $\pi(D_n \mid j) > 0$ and known for all j in D_n, and if the true model is Logit with choice-set C_n, it is possible to obtain consistent estimators of the model parameters β^* by maximizing the following quasi-log-likelihood function:

$$QL_{Logit,D} = \sum_{n=1}^{N} \ln \frac{e^{V(x_{in},\beta)+\ln \pi(D_n|i,x_n)}}{\sum_{j \in D_n} e^{V(x_{jn},\beta)+\ln \pi(D_n|j,x_n)}}. \quad (6)$$

McFadden´s procedure falls into the type of estimators identified by White (1982), which achieve the consistent estimation of the model parameters in spite of being misspecified.

Eq. (6) can be simplified if the sampling correction $\ln \pi(D_n \mid i)$ is the same for all alternatives. In that case, the sampling correction will cancel out in Eq. (6) and can be ignored. The effects of using other sampling protocols are studied by Manski and McFadden (1981), Ben-Akiva and Lerman (1985), Watanatada and Ben-Akiva (1979) and Frejinger *et al.* (2009).

Diverse applications of McFadden's results on sampling of alternatives for Logit models can be found in the literature. Some examples are Parsons and Kealy (1992) and Sermons and Koppelman (2001). In turn, the extension of McFadden's results to non-Logit models is a problem for which few little progress have been made in the last 30 years. Some advances have been done for choice-based samples; cases where the full choice-set is available to the researcher, but the observations are instead sampled depending on the choices. First, Manski and Lerman (1977) proposed a consistent but inefficient estimator for non-Logit models. This estimator was also used by Cosslett (1981) and by Imbens and Lancaster (1994). Later, Garrow *et al.* (2005) proposed an efficient estimator for a particular case of the Nested Logit model. Lastly, Bierlaire *et al.* (2008) proposed an alternative estimator that is applicable to MEV models with choice based samples and does not require knowledge of the sampling protocol.

Additionally, some analyses have been done regarding the impact of sampling of alternatives in Logit Mixture models. For example, McConnel and Tseng (2000), and Nerella and Bhat (2004), used Monte Carlo experimentation to study the problem of sampling of alternatives in random coefficients Logit models and found that sampling causes only small changes to parameter estimates. In turn, Chen *et al.* (2005) used Monte Carlo experimentation to show that, for Logit Mixture models that capture correlation among alternatives, the effects of sampling might be severe. Finally, Domanski (2009), citing an unpublished paper attributed to Haefen and Jacobsen, claims that the use of the expectation-maximization algorithm (see, e.g., Train, 2009) might result in the consistent estimation of model parameters while sampling of alternatives in random coefficients Logit Mixture model.

Regarding the problem of sampling of alternatives for the Nested Logit, several authors have directly applied McFadden's results for Logit without any modification. Examples of these type of applications include Berkovec and Rust (1985), Train *et al.* (1987), Hansen (1987), and Rivera and Tiglao (2005). As it will be shown later, this approach may significantly impact the estimators of the model parameters. Finally, to the best of my knowledge, the only attempt to deal with the problem of sampling of alternatives in the Nested Logit model corresponds to the work of Lee and Wadell

(2010). These authors use a method based on an idea originally suggested by Ben-Akiva (2009), which we further develop in the next section.

3 A Novel Method for MEV Models

In this section, we present a novel methodology to address the problem of sampling of alternatives and estimation for Multivariate Extreme Value (MEV) models, based on an idea originated by Ben-Akiva (2009).

The genesis of MEV models goes back to 1973, when Ben-Akiva proposed the Nested Logit model. Afterwards, McFadden (1978) showed that the Logit, the Nested Logit and other models belonged to a more general class of closed-form choice models that can handle diverse correlation structures among alternatives in the choice-set. McFadden originally denominated this class of models as Generalized Extreme Value (GEV) models. Since the error terms for this class of models follow a MEV distribution, the models themselves are termed here as MEV.

The joint distribution of the error terms of the utilities in MEV models has the following cumulative density function

$$F\left(\varepsilon_{1n},\ldots,\varepsilon_{Jn}\right)=e^{-G\left(e^{-\varepsilon_{1n}},\ldots,e^{-\varepsilon_{Jn}};\gamma\right)}, \tag{7}$$

where G is a generating function that is specific to each member of the MEV family, and γ is a set of distribution parameters. If G complies with certain requirements (McFadden, 1978) the choice model implied by Eq. (7) will be consistent with the random utility maximization theory. Later, Ben-Akiva and Lerman (1985) show that the MEV choice probability can be written in a Logit form as shown in Eq. (8)

$$P_n\left(i\right)=\frac{e^{V\left(x_{in},\beta\right)+\ln G_i\left(\left\langle e^{V_{ln}}\right\rangle_{l\in C_n};\gamma\right)}}{\sum_{j\in C_n}e^{V\left(x_{jn},\beta\right)+\ln G_j\left(\left\langle e^{V_{ln}}\right\rangle_{l\in C_n};\gamma\right)}}, \tag{8}$$

where $G_i\left(\left\langle e^{V_{ln}}\right\rangle_{l\in C_n};\gamma\right)=\frac{\partial G\left(e^{V_{1n}},\ldots,e^{V_{Jn}};\gamma\right)}{\partial e^{V_{in}}}\equiv G_{in}$.

Given the Logit form of the MEV model, it might look as if the problem of sampling of alternatives can be easily extended to MEV by following the same process of analysis deployed before for Logit, as shown in Eq. (3)-(5). That procedure results in the following expression for the conditional probability of choosing alternative i, given that set D_n was constructed:

$$\pi\left(i\mid D_n\right)=\frac{e^{V\left(x_{in},\beta\right)+\ln G_i\left(\left\langle e^{V_{ln}}\right\rangle_{l\in C_n};\gamma\right)+\ln\pi\left(D_n|i\right)}}{\sum_{j\in D_n}e^{V\left(x_{jn},\beta\right)+\ln G_j\left(\left\langle e^{V_{ln}}\right\rangle_{l\in C_n};\gamma\right)+\ln\pi\left(D_n|j\right)}}.$$

Then, the same demonstration developed by McFadden (1978) for Logit, can be claimed to show that the maximization of the following quasi-log-likelihood function

$$QL_{MEV,D,C} = \sum_{n=1}^{N} \ln \pi(i \mid D_n) = \sum_{n=1}^{N} \ln \frac{e^{V(x_{in},\beta) + \ln G_i\left(\left\langle e^{V_{ln}} \right\rangle_{l \in C_n}; \gamma\right) + \ln \pi(D_n|i)}}{\sum_{j \in D_n} e^{V(x_{jn},\beta) + \ln G_j\left(\left\langle e^{V_{ln}} \right\rangle_{l \in C_n}; \gamma\right) + \ln \pi(D_n|j)}} \tag{9}$$

leads to consistent estimators of the model parameters.

However, it can be immediately noted that Eq. (9) is not practical. Even though the denominator of the choice probability depends only on D_n, the argument of the term $\ln G_{in}$ still depends on the full choice-set C_n. Ben-Akiva (2009) suggests that this problem might be solved if G_{in} is replaced by an estimator that depends only on the subset D_n.

In this paper, we formalize the idea proposed by Ben-Akiva (2009), analyze the conditions required for its success, study the asymptotic properties of the estimators resulting from it, determine the correct expansion factors required in some relevant examples, and study the properties of the estimators using Monte Carlo experimentation and real data.

The results on consistency, asymptotic normality and efficiency can be summarized in the following theorem:

Theorem: Given N observations, a choice-set C_n of cardinality J_n, and a subset D_n of cardinality $\tilde{J}_n$. If

a) $\pi(D_n \mid j) > 0 \quad \forall j \in D_n$ and $\pi(D_n \mid j) = 0 \quad \forall j \notin D_n$,

b) the choice model is MEV and $G_{in} = \dfrac{\partial G\left(e^{V_{1n}}, ..., e^{V_{Jn}}; \gamma\right)}{\partial e^{V_{in}}}$,

c) $G_{in} = f\left(B_i(C_n)\right)$ where f is continuous and twice-differentiable,

d) $\hat{B}_i(D_n)$ is a consistent (in $\tilde{J}_n$) and unbiased estimator of $B_i(C_n)$, and

e) $Var\left(\hat{B}_{in}\right) = K_n / \tilde{J}_n$ with K_n scalar;

then, the maximization of the quasi-log-likelihood function

$$QL_{MEV,D} = \sum_{n=1}^{N} \ln \hat{\pi}(i \mid D_n) = \sum_{n=1}^{N} \ln \frac{e^{V(x_{in},\beta) + \ln f\left(\hat{B}_i(D_n)\right) + \ln \pi(D_n|i)}}{\sum_{j \in D_n} e^{V(x_{jn},\beta) + \ln f\left(\hat{B}_j(D_n)\right) + \ln \pi_n(D_n|j)}} \tag{10}$$

yields, under general regularity conditions, consistent estimators (in N) of the model parameters β^*, as $\tilde{J}_n$ increases with N at any rate. If $\tilde{J}_n$ increases faster than $\sqrt{N}$, the estimators of the model parameters will be consistent, asymptotically normal, and as efficient as the estimators obtained from the maximization of a quasi-log-likelihood shown in Eq. (9). Finally, if J_n is finite and the protocol is sampling without replacement, $\tilde{J}_n$ needs to increase only up to $\tilde{J}_n = J_n$ in order to achieve consistency and relative efficiency.

Proof. Given that $\hat{B}_{in}$ is a consistent estimator of B_{in}, as $\tilde{J}_n$ increases, the Slutsky theorem guarantees that $\ln f\left(\hat{B}_i(D_n)\right)$ will also be a consistent estimator of $\ln G_{in}$, because

the log and *f* are continuous. Equivalently, since $\pi(i \mid D_n)$ is continuous in $\ln G_{in}$, the Slutsky theorem guarantees that $\hat{\pi}(i \mid D_n)$ will be a consistent estimator of $\pi(i \mid D_n)$. Finally, McFadden's consistency results for Logit, shown in Eq. (6), guarantees that the maximization of the quasi-log-likelihood shown in Eq. (10) will result in the consistent estimation of the model parameters as *N* increases.

Note that the claim of McFadden's consistency result is established as *N* increases, but the consistency of $\hat{B}_{in}$, $\ln f\left(\hat{B}_i(D_n)\right)$ and $\hat{\pi}(i \mid D_n)$ is established as $\tilde{J}_n$ increases. To rely legitimately on the Slutsky theorem, it is indispensable to determine a concordance between $\tilde{J}_n$ and *N*. This concordance can be established by analyzing the asymptotic properties of the estimators.

The asymptotic distribution of the estimators of the model parameters that result from the maximization of the quasi-log-likelihood shown in Eq (10) can be derived using the two-stage approach employed by Train (2009, pp. 247-257) to analyze the asymptotic properties of simulation-based estimators. In a first stage, we will analyze the asymptotic distribution of the sample average of the score, which is defined as the gradient of the quasi-log-likelihood shown in Eq. (10). In a second stage we will use those results to derive the asymptotic distribution of the estimators of the model parameters.

Consider that the choice-sets *C* and *D*, of cardinalities *J* and $\tilde{J}$ respectively, do not vary across observations, and that there is a single term $\ln G_n$ that needs to be approximated for each observation *n*. Then, instead of B_{in}, the term considered in this case should be B_n. These assumptions are not essential, and can be easily generalized, but help in substantially to reduce notation burden. With the same purpose, we will refer to the whole set of model parameters β and μ, just as β.

Under this setting, the sample average of the score evaluated using the estimator $\hat{B}_n$ will correspond to:

$$\hat{g}(\beta) = \frac{1}{N}\sum_{n=1}^{N}\hat{g}_n(\beta) = \frac{1}{N}\sum_{n=1}^{N}\frac{\partial \ln \hat{\pi}_n}{\partial \beta} = \frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial \beta}\ln\frac{e^{V(x_{in},\beta)+\ln f\left(\hat{B}_n\right)+\ln \pi_n(D|i,\beta)}}{\sum_{j\in D} e^{V(x_{jn},\beta)+\ln f\left(\hat{B}_n\right)+\ln \pi_n(D|j,\beta)}}.$$

To study the asymptotic distribution of $\hat{g}(\beta)$ in the vicinity of the true values β^*, consider the following re-arrangement of terms

$$\hat{g}(\beta^*) = \underbrace{g(\beta^*)}_{A_1} + \underbrace{\left[E(\hat{g}(\beta^*)) - g(\beta^*)\right]}_{A_2} + \underbrace{\left[\hat{g}(\beta^*) - E(\hat{g}(\beta^*))\right]}_{A_3}.$$

The first term $A_1 = g(\beta^*)$ is the statistic that is being approximated by $\hat{g}(\beta^*)$, where

$$g(\beta) = \frac{1}{N}\sum_{n=1}^{N}\frac{\partial \ln \pi_n(\beta)}{\partial \beta} = \frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial \beta}\ln\frac{e^{V(x_{in},\beta)+\ln G_n(C)+\ln \pi_n(D|i,\beta)}}{\sum_{j\in D} e^{V(x_{jn},\beta)+\ln G_n(C)+\ln \pi_n(D|j,\beta)}}.$$

The second term $A_2 = E(\hat{g}(\beta^*)) - g(\beta^*)$ corresponds to the bias of the estimator of $g(\beta^*)$. The third term A_3 corresponds to the noise of the approximation, which is the difference between a particular realization of $\hat{g}(\beta^*)$, and its expected value.

Consider the noise term A_3, which can be rewritten as follows:

$$A_3 = \hat{g}(\beta^*) - E(\hat{g}(\beta^*))$$

$$A_3 = \frac{1}{N}\sum_n [\hat{g}_n(\beta^*) - E(\hat{g}_n(\beta^*))]$$

$$A_3 = \frac{1}{N}\sum_n d_n,$$

where each d_n is the deviation of $\hat{g}(\beta^*)$ from its expectation for observation *n*. Note that each d_n depends on a particular draw of alternatives to construct the set *D*. This means that there is a distribution of values of d_n depending on all possible draws of alternatives in *D*. The distribution of d_n has zero mean because the expectation is subtracted in the creation of d_n. Also, note that the variance of d_n should decrease with the cardinality of *D* because $\hat{g}(\beta^*)$ should become closer to its expected value as $\tilde{J}$ increases. To account for this effect, the variance of d_n can be expressed as $S_n/\tilde{J}$, where S_n is the variance when $\tilde{J} = 1$. Then, relying on the generalized version of the central limit theorem (see, e.g., Train, 2009, pp.246), the noise A_3 will have the following limiting distribution:

$$\sqrt{N}A_3 \xrightarrow{d} \text{Normal}\left(0, \mathbf{S}/\tilde{J}\right),$$

where **S** is the population mean of S_n. Consequently, the asymptotic distribution of the noise A_3 will be

$$A_3 \overset{a}{\sim} \text{Normal}\left(0, \mathbf{S}/\tilde{J}N\right).$$

It is interesting to note what occurs with the noise A_3 when *N* increases but $\tilde{J}$ is fixed. In this case, $\sqrt{N}A_3$ will have a limiting distribution, but will not vanish as *N* increases. In turn, the asymptotic variance of the noise A_3 will decrease as *N* increases, even if $\tilde{J}$ is fixed. Note also that when the protocol is sampling without replacement and *J* is finite, $\tilde{J}$ needs to increase only up to *J*, since from that point $\hat{g}(\beta) = E(\hat{g}(\beta)) = g(\beta)$ because any resorting of the alternatives in the choice-set *C* will have no impact on the choice probabilities.

Consider the bias term A_2. This bias exists because the method described in Eq. (10) considers an unbiased estimator $\hat{B}_n$ of B_n, but the calculation of $\hat{g}(\beta)$ involves a series of nonlinear transformations of $\hat{B}_n$. The bias can be studied by taking a second order Taylor's approximation of $\hat{g}(\beta)$ around $\hat{B}_n = B_n$. Noting that $\hat{g}_n(\beta, B_n) = g_n(\beta)$, it follows that

$$\hat{g}_n(\beta) = g_n(\beta) + \frac{\partial \hat{g}_n}{\partial \hat{B}_n}\left[\hat{B}_n(\beta) - B_n(\beta)\right] + \frac{1}{2}\frac{\partial^2 \hat{g}_n}{\partial \hat{B}_n^2}\left[\hat{B}_n(\beta) - B_n(\beta)\right]^2 + o_n.$$

Then, taking expectations (over possible realizations of the set *D*), recalling that $\hat{B}_n$ is an unbiased estimator of B_n, and considering that the discrepancy o_n has zero mean, this Taylor's approximation can be rewritten as

$$E(\hat{g}_n(\beta)) - g_n(\beta) = \frac{1}{2}\frac{\partial^2 \hat{g}_n(\beta)}{\partial \hat{B}_n^2} Var\left(\hat{B}_n(\beta)\right).$$

Note that the $Var\left(\hat{B}_n(\beta)\right)$ should decrease as $\tilde{J}$ increases because then $\hat{B}_n$ will become progressively closer to B_n. Assuming that this relationship can be captured by the expression $Var\left(\hat{B}_n(\beta)\right) = K_n / \tilde{J}$, where K_n is a scalar, the bias A_2 can be rewritten as

$$A_2 = E(\hat{g}(\beta)) - g(\beta) = \frac{1}{N}\sum_n [E(\hat{g}_n(\beta)) - g_n(\beta)]$$

$$A_2 = \frac{1}{N}\sum_n \frac{1}{2}\frac{\partial^2 \hat{g}_n(\beta)}{\partial \hat{B}_n^2}\frac{K_n}{\tilde{J}}$$

$$A_2 = \frac{Z}{\tilde{J}}$$

where Z is the sample average of $\frac{K_n}{2}\frac{\partial^2 \hat{g}_n}{\partial \hat{B}_n^2}$.

The bias A_2 will vanish as N increases, if and only if $\tilde{J}$ increases also with N. Otherwise, $\hat{g}(\beta)$ will be an inconsistent estimator of $g(\beta)$. Instead, an even stronger assumption is required to achieve asymptotic normality. To understand why, consider the bias A_2 normalized for sample size N

$$\sqrt{N} A_2 = \frac{\sqrt{N}}{\tilde{J}} Z .$$

This term will vanish as N increases, if and only if $\tilde{J}$ increases faster than $\sqrt{N}$. Otherwise, the estimator $\hat{g}(\beta)$ will have neither a limiting nor an asymptotic distribution.

Equivalent to what occurred with the noise A_3, note that when the protocol is sampling without replacement and J is finite, $\tilde{J}$ needs to increase only up to J, since from that point $E(\hat{g}(\beta)) = g(\beta)$ because any resorting of the alternatives in the choice-set C will have no impact on the choice probabilities.

In summary, it was shown that if $\tilde{J}$ increases with N at any rate, $\hat{g}(\beta) \xrightarrow{p} g(\beta)$ and when $\tilde{J}$ increases faster than $\sqrt{N}$, $\hat{g}(\beta)$ will be asymptotically Normal. Given that $\hat{g}(\beta) \xrightarrow{p} g(\beta)$, the limiting and asymptotic distributions of $\hat{g}(\beta)$ will be the same as those of $g(\beta)$.

To study the asymptotic properties of $g(\beta)$, label **W** the population variance of $g_n(\beta^*)$. Then, assuming that $g(\beta)$ equals zero in the population, by the central limit theorem, the limiting distribution of $g(\beta)$ corresponds to

$$\sqrt{N}(g(\beta^*) - 0) \xrightarrow{d} \text{Normal}(0, \mathbf{W}),$$

and the asymptotic distribution corresponds to

$$g(\beta^*) \overset{a}{\sim} \text{Normal}(0, \mathbf{W}/N).$$

It is then possible to combine the results for the components of $\hat{g}(\beta)$ in order to study the asymptotic distribution of the estimators $\hat{\beta}$ of the model parameters β. This can be achieved by taking a first-order Taylor's expansion of $\hat{g}\left(\hat{\beta}\right)$ around the true values β^*

$$\hat{g}(\hat{\beta}) = \hat{g}(\beta^*) + \hat{R}[\hat{\beta} - \beta^*] + o_n,$$

where $\hat{R} = \partial\hat{g}/\partial\beta$. Then, note that the estimators $\hat{\beta}$ of the model parameters β are defined by the condition $\hat{g}(\hat{\beta}) = 0$, because dividing Eq. (10) by N does not impact the solution of the problem. Assuming that and the discrepancy o_n disappears asymptotically, it follows that the limiting distribution of the estimators is

$$\sqrt{N}(\hat{\beta} - \beta^*) = \sqrt{N}(-\hat{R}^{-1})\hat{g}(\beta^*) = \sqrt{N}(-\hat{R}^{-1})(A_1 + A_2 + A_3). \quad (11)$$

As established before, if $\tilde{J}$ increases faster than $\sqrt{N}$ the terms A_2 and A_3 will vanish. Under this condition, the term A_1 in Eq. (11) becomes asymptotically equal to $g(\beta)$, which has a limiting distribution of $\sqrt{N}(g(\beta^*) - 0) \xrightarrow{d} \text{Normal}(0, \mathbf{W})$. Note that $\hat{R} \xrightarrow{p} \mathbf{R}$, where $\mathbf{R} = E(\hat{R})$. This implies that the limiting distribution of the estimators of the model parameters becomes

$$\sqrt{N}(\hat{\beta} - \beta^*) \xrightarrow{d} \text{Normal}(0, \mathbf{R}^{-1}\mathbf{W}\mathbf{R}^{-1}), \quad (12)$$

and their asymptotic distribution will be

$$\hat{\beta} \overset{a}{\sim} \text{Normal}(\beta^*, \mathbf{R}^{-1}\mathbf{W}\mathbf{R}^{-1}/N) = \text{Normal}(\beta^*, \boldsymbol{\Omega}/N), \quad (13)$$

where $\boldsymbol{\Omega} = \mathbf{R}^{-1}\mathbf{W}\mathbf{R}^{-1}$, $\mathbf{W} = Var\left(\frac{\partial \ln \pi_n(\beta^* \mid D)}{\partial \beta}\right)$ and $\mathbf{R} = E\left(\frac{\partial^2 \ln \pi_n(\beta^* \mid D)}{\partial \beta \partial \beta'}\right)$.

$\boldsymbol{\Omega}$ is usually defined as the "robust" or "sandwich" variance-covariance matrix of the estimators of the model parameters (see, e.g., Train, 2009, pp. 201). Berndt *et al.* (1974) proposed an estimator of $\boldsymbol{\Omega}$ that is known as the BHHH matrix and is used, for example, by the discrete-choice estimation software Biogeme (Bierlaire, 2003). To deploy the BHHH matrix for this case, note that $\mathbf{R}$ is the Hessian of the model shown in Eq. (9). A consistent estimator of $\mathbf{R}$ is its sample analog, which can be constructed from the Hessian of the quasi-log-likelihood shown in Eq. (10). Equivalently, the variance-covariance matrix of the score of the model shown in Eq. (10), evaluated at the estimated values $\hat{W}(\hat{\beta})$, is a consistent estimator of $\mathbf{W}$. Given that $\hat{g}(\hat{\beta}) = 0$, $\hat{W}(\hat{\beta})$ can be calculated as the outer product of the scores of the model shown in Eq. (10). In summary, the BHHH estimator for the variance-covariance matrix of the estimators of the model parameters resulting from the maximization of the quasi-log-likelihood function shown in Eq. (10), corresponds to the following expression:

$$\hat{\Omega} = \left[\frac{\partial^2 \ln \hat{\pi}(\hat{\beta} \mid D)}{\partial \beta \, \partial \beta'}\right]^{-1} \left[\sum_{n=1}^{N} \frac{\partial \ln \hat{\pi}_n(\hat{\beta} \mid D)}{\partial \beta} \frac{\partial \ln \hat{\pi}_n(\hat{\beta} \mid D)}{\partial \beta'}\right] \left[\frac{\partial^2 \ln \hat{\pi}(\hat{\beta} \mid D)}{\partial \beta \, \partial \beta'}\right]^{-1}.$$

These results imply that the estimators obtained by the maximization of Eq. (10) will have the same asymptotic variance-covariance matrix as the estimators that would be obtained by using Eq. (9); that is, if the full choice-set C is available for the calculation of the expansion of the term $\ln G_n$. Then, it can be affirmed that estimators obtained by maximizing Eq. (10) are efficient among all possible approximations of the model described in Eq. (9). **Q.E.D.**

It is interesting to note that the estimators obtained by maximizing Eq. (9) are not globally efficient because Eq. (9) is not the true log-likelihood and therefore the

Crammer-Rao lower bound is not attained. This also implies that the estimators obtained by using McFadden's (1978) method for Logit are also inefficient. McFadden (1978) did not study the asymptotic distribution of his estimators. However, following the same line of analysis deployed in this section, it can be shown that the asymptotic distribution of McFadden's (1978) estimators will be equal to Eq. (13), using instead Eq. (6) to calculate the terms **R** and **W**.

Additionally, the fact that the estimators obtained with the method deployed in Eq. (10) will not be consistent unless $\widetilde{J}$ increases with *N*, implies that, in practice, we should test the stability of the estimators of the model parameters as a function of $\widetilde{J}$. If the estimators for different values of $\widetilde{J}$ are statistically equal, we can be sure that the finite sample (of alternatives) bias is negligible. Otherwise, $\widetilde{J}$ should be increased until attaining stability. This is equivalent to the need for testing for the stability of Logit Mixture's estimators as a function of the number of draws, in the simulated maximum-likelihood framework (Walker, 2001).

The practical implementation of the method to achieve consistency and asymptotic normality under sampling of alternatives in MEV models depends on the specific MEV model and the sampling protocol being considered. In the next two sections, we analyze this implementation in detail for the Nested and the Cross-Nested Logit models, respectively. Then, for illustrative purposes, in Section 6, we develop a Monte Carlo experiment where the performance of the method is analyzed under different circumstances. Finally, in Section 7, the methodology is applied to a Nested Logit of residential location choice that was estimated using real data from Lisbon, Portugal.

4 Formulation of the Method for Nested Logit

The Nested Logit model is a closed-form discrete choice model that allows for the correlation among random components of the utilities of alternatives that belong to mutually exclusive and totally exhaustive subsets (or nests) of the full choice-set. In this model, the marginal choice probabilities are written as the product of the conditional probability of choosing each alternative (given that the respective nest is chosen) and the marginal probability of choosing the respective nest. The utility of a nest is defined as the expected maximum utility of choosing the alternatives that belong to that nest, what is known as the inclusive value (Ben-Akiva and Lerman, 1985).

McFadden (1978) showed that the Nested Logit model can be alternatively formulated as a member of the MEV family. The generating function *G* for a Nested Logit model with *M* nests is

$$G\left(\left\langle e^{V_{ln}}\right\rangle_{l\in C_n};\gamma\right)=\sum_{m=1}^{M}\left(\sum_{i\in C_{m(i)n}}e^{\mu_m V_{in}}\right)^{\frac{\mu}{\mu_m}}, \qquad (14)$$

where *m(i)* is the nest to which *i* belongs, γ is the set of scales μ_m of the nests, and $C_{m(i)n}$ is the set of alternatives that belong to the nest $m(i)$. In this case, $\ln G_{in}$ corresponds to the expression shown in Eq. (15).

$$\ln G_{in} = \left(\frac{\mu}{\mu_{m(i)}} - 1 \right) \left(\ln \sum_{j \in C_{m(i)n}} e^{\mu_{m(i)} V_{jn}} \right) + \ln \mu + \left(\mu_{m(i)} - 1 \right) V_{in} \qquad (15)$$

Then, if a sample $D_{m(i)n}$, is drawn from the true choice-set $C_{m(i)n}$, the only term that would be affected (and therefore needs to be approximated) is the sum of the exponentials of the systematic utilities, the argument of the *logsum*. The sum of the exponentials will be denoted as

$$B_{in} = \sum_{j \in C_{m(i)n}} e^{\mu_{m(i)} V_{jn}} .$$

One way of approximating B_{in} is by constructing an expanded sum of the exponentials of the utilities of the alternatives in $D_{m(i)n}$. Then, the challenge would be to determine the expansion factors w_{jn} required to obtain an unbiased and consistent estimator of the sum of the exponentials.

To obtain an unbiased estimator, the expansion factors have to comply with the conditions shown in Eq. (16), where the first expectation is taken over all values of x, and the second expectation is taken over x and all potential sets $D_{m(i)n}$.

$$E\left(B_{in}\right) - E\left(\hat{B}_{in}\right) = 0 = E_x \left(\sum_{j \in C_{m(i)n}} e^{\mu_{m(i)} V_{jn}} \right) - E_{x,D} \left(\sum_{j \in D_{m(i)n}} w_{jn} e^{\mu_{m(i)} V_{jn}} \right) \qquad (16)$$

Note that each $e^{\mu_{m(i)} V_{jn}}$ can be seen as a random variable with mean $\eta_{m(i)n}$, the mean of the empirical distribution of $e^{\mu_{m(i)} V_n}$. In this case the first component of Eq. (16) becomes

$$E\left(B_{in}\right) = E\left(\sum_{j \in C_{m(i)n}} e^{\mu_{m(i)} V_{jn}} \right) = J_{m(i)n} \eta_{m(i)n} .$$

The expansion factors w_{jn} required to obtain an unbiased estimator of B_{in} shall depend on the sampling protocol. For analytical purposes we will consider first that the protocol is sampling without replacement and then that it is sampling with replacement. Finally, we will show that the expansion factors w_{jn} required in both cases can be summarized in a single expression.

Consider first that the protocol is sampling **without** replacement by nest. Then, using the following indicator function

$$1_{j \in D_{m(i)n}} = \begin{cases} 1 & \text{if } j \in D_{m(i)n} \\ 0 & o/w \end{cases}$$

it is possible to rewrite $E\left(\hat{B}_{in}\right)$ in Eq. (16) as follows:

$$E\left(\hat{B}_{in}\right) = E\left(\sum_{j \in D_{m(i)n}} w_{jn} e^{\mu_{m(i)} V_{jn}} \right) = E\left(\sum_{j \in C_{m(i)n}} 1_{j \in D_{m(i)n}} w_{jn} e^{\mu_{m(i)} V_{jn}} \right).$$

Then, by the Law of Total Expectations (also known as the Law of Iterated Expectations), which is equivalent to the total probability theorem used in Eq. (4),

$$E\left(\hat{B}_{in}\right) = E\left(E\left(\sum_{j \in C_{m(i)n}} 1_{j \in D_{m(i)n}} w_{jn} e^{\mu_{m(i)} V_{jn}} \mid 1_{j \in D_{m(i)n}} \right) \right)$$

$$E\left(\hat{B}_{in}\right)=E\left(\sum_{j\in C_{m(i)n}}1_{j\in D_{m(i)n}}w_{jn}E\left(e^{\mu_{m(i)}V_{jn}}\mid 1_{j\in D_{m(i)n}}\right)\right)=E\left(\sum_{j\in C_{m(i)n}}1_{j\in D_{m(i)n}}w_{jn}\eta_{m(i)n}\right)$$

$$E\left(\hat{B}_{in}\right)=\sum_{j\in C_{m(i)n}}E\left(1_{j\in D_{m(i)n}}\right)w_{jn}\eta_{m(i)n}\ ,$$

where $E\left(e^{\mu_{m(i)}V_{jn}}\mid 1_{j\in D_{m(i)n}}\right)=\eta_{m(i)n}$ results from the fact that the distribution of $e^{\mu_{m(i)}V_n}$ determines the sampling of $D_{m(i)n}$, but the causality does not go in the other direction.

Given this result, one way for Eq. (16) to equal zero is by having

$$w_{jn}=1/E\left(1_{j\in D_{m(i)n}}\right),$$

where $E\left(1_{j\in D_{m(i)n}}\right)$ is the probability of drawing alternative *j*, because the protocol in this case is sampling without replacement.

Consider now that the protocol is sampling **with** replacement by nest. Then it is necessary to define the set $\tilde{D}_{m(i)n}$ and the indicator function $\tilde{n}_{jn}$. The former is a set that includes all the repetitions of the alternatives sampled, and the latter corresponds to the number of times alternative *j* is repeated in the set $\tilde{D}_{m(i)n}$. Then $\hat{B}_{in}$ can be rewritten as follows

$$\hat{B}_{in}=\sum_{j\in\tilde{D}_{m(i)n}}\tilde{w}_{jn}e^{\mu_{m(i)}V_{jn}}=\sum_{j\in C_{m(i)n}}\tilde{n}_{jn}\tilde{w}_{jn}e^{\mu_{m(i)}V_{jn}}\text{, and therefore}$$

$$E\left(\hat{B}_{in}\right)=\sum_{j\in C_{m(i)n}}\tilde{w}_{jn}\eta_{m(i)n}E\left(\tilde{n}_{jn}\right)\ \text{ and then }\ \tilde{w}_{jn}=1/E\left(\tilde{n}_{jn}\right).$$

Finally, since

$$\hat{B}_{in}=\sum_{j\in\tilde{D}_{m(i)n}}\tilde{w}_{jn}e^{\mu_{m(i)}V_{jn}}=\sum_{j\in D_{m(i)n}}\tilde{n}_{jn}\tilde{w}_{jn}e^{\mu_{m(i)}V_{jn}}=\sum_{j\in D_{m(i)n}}w_{jn}e^{\mu_{m(i)}V_{jn}},$$

the expansion factors required to obtain an unbiased estimation of the sum of the exponentials, for the case of sampling with replacement, are equal to

$$w_{jn}=\tilde{n}_{jn}/E\left(\tilde{n}_{jn}\right).$$

The expansion factors required when the protocol is with or without replacement can be summarized in a single expression by noting that, when the protocol is sampling without replacement, $\tilde{n}_{jn}$=1 if *j* is in $D_{m(i)n}$, and $E\left(1_{j\in\tilde{D}_{m(i)n}}\right)$ is also the expected number of times alternative *j* would be drawn to construct the set $D_{m(i)n}$. Then, the general expression for the expansion factors required to obtain an unbiased estimator of B_{in} can be denoted as shown in Eq. (17).

$$w_{jn}=\frac{\tilde{n}_{jn}}{E\left(\tilde{n}_{jn}\right)} \tag{17}$$

The next step is to prove that the expansion factors shown in Eq. (17) will lead to consistent estimators of B_{in} as $\tilde{J}_{m(i)n}$ increases. This results directly from any weak Law of Large Numbers. Actually, consistency would be granted even if no expansion factors were considered at all. As $\tilde{J}_{m(i)n}$ increases, even an estimator of B_{in} that only considers the simple sum of the exponentials of the alternatives in $D_{m(i)n}$ will eventually be as near

to B_{in} as desired, as $\tilde{J}_{m(i)n}$ increases. The difference is that the expansion factors shown in Eq. (17) will converge faster, leading to better finite sample properties. In addition, using Eq. (17) is what allows obtaining an unbiased estimator, condition required in the derivation of the results on efficiency and asymptotic normality.

5 Formulation of the Method for Cross-Nested Logit

The Cross-Nested Logit model is a closed-form discrete choice model that allows for correlation among the random components of the utilities of all alternatives in the choice-set. Similar to the Nested Logit, the Cross-Nested Logit considers a set of nests *m*. However, in the Cross-Nested Logit model the nests are totally exhaustive but not mutually exclusive in the coverage of the alternatives in the choice-set. The correlation structure is defined by a non-negative weight α_{jm} representing the degree of belonging of alternative j to the nest *m*. Examples of applications of the Cross-Nested Logit model and variations of it are the works of Small (1987), Vovsha (1997), Vovsha and Bekhor (1998), Bierlaire (2001), and Papola (2004).

The Cross-Nested Logit model can be formulated as a member of the MEV family. In general, with *M* nests, the generating function *G* that results in the Cross-Nested Logit model is

$$G\left(\left\langle e^{V_{ln}}\right\rangle_{l\in C_n};\gamma\right)=\sum_{m=1}^{M}\left(\sum_{i\in C_n}\alpha_{jm}e^{\mu_m V_{in}}\right)^{\frac{\mu}{\mu_m}},$$

where *m* are the nests, γ corresponds to the set of scales μ_m of the nests, and α_{jm} are the weights. Then, the term $\ln G_{in}$ corresponds to the following expression:

$$\ln G_{in}=\ln\sum_{m=1}^{M}\left(\mu\alpha_{im}e^{V_i(\mu_m-1)}\left(\sum_{j\in C_n}\alpha_{jm}e^{\mu_m V_j}\right)^{\frac{\mu-\mu_m}{\mu_m}}\right).$$

Just as it occurred with the Nested Logit, if a sample D_n is drawn from the true choice-set C_n, the only term affected will be the sum of the exponentials, which is now weighed by the terms α_{jm}. Then, consistency, relative efficiency, and asymptotic normality can be achieved for the Cross-Nested Logit while sampling of alternatives, using the following estimator:

$$\hat{B}_{in}=\sum_{j\in D_n}w_{jn}\alpha_{jm}e^{\mu_m V_j}\approx B_{in}=\sum_{j\in C_n}\alpha_{jm}e^{\mu_m V_j}\ .$$

The same derivation used in Eq. (16)-(17) can be used to show that the expansion factors w_{jn} required in this case are also those shown in Eq. (17).

6 Monte Carlo Experiment

6.1 Model Setting

A Monte Carlo experiment was performed to analyze and illustrate the properties of the proposed method in achieving consistency in the case of sampling of alternatives in MEV models. The setting of this experiment is summarized by Figure 1. The true or underlying

model is a Nested Logit with 1,005 alternatives, among which the first 5 belong to one nest ($J_1 = 5$) and the other 1,000 to a second nest ($J_2 = 1{,}000$). The systematic utilities V_{in} depend upon two variables, x_1 and x_2, which were constructed *iid* Uniform (-1,1) for the N=2,000 observations. The true parameters of the model are $\mu = 1$, $\mu_1 = 2, \mu_2 = 3, \beta_{x_1} = \beta_{x_2} = 1$.

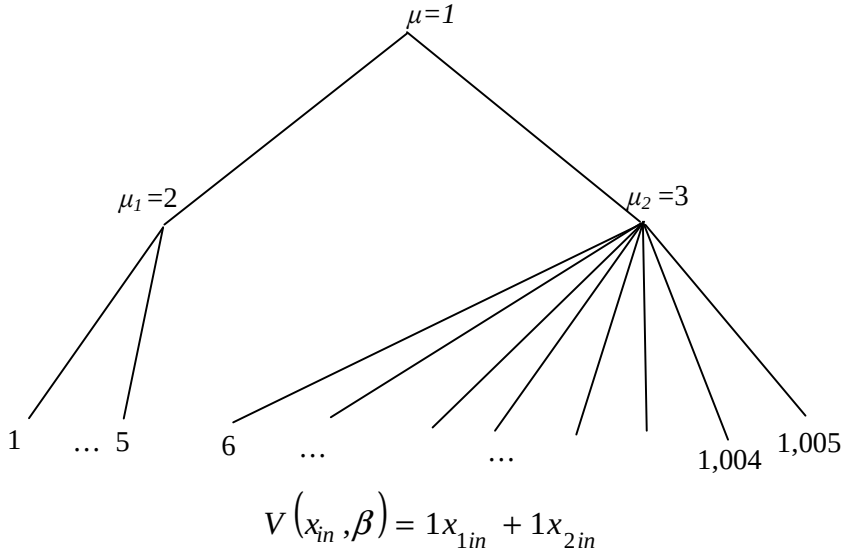

Figure 1 Monte Carlo Experiment: Nesting Structure. 1,005 Alternatives

N=2,000 $J_1 = 5$ $\tilde{J}_1 = 5$; $J_2 = 1{,}000$ $\tilde{J}_2 = 5$ and 500

The methodology used to implement the Nested Logit model shown in Figure 1 for Monte Carlo experimentation was performed in several steps. First, the choice probability was calculated replacing the true values of the parameters in Eq. (8). Then, these choice probabilities were used to build a discrete cumulative density function by alternative. Afterwards, a random number Uniform (0,1) was generated for each observation. Finally, the chosen alternative was determined as the inverse of the cumulative density function, evaluated for each random number.

The sampling protocol used to draw alternatives from the choice-set in this experiment was stratified importance sampling without replacement by nest. First, the chosen alternative for each observation was included. Then non-chosen alternatives were randomly sampled, without replacement by nest, to make a total of $\tilde{J}_1 = 5$ for the first nest, and $\tilde{J}_2 = 5$ and $\tilde{J}_2 = 500$ for the second nest.

Given this sampling protocol, the conditional probability of constructing a particular set D_n for observation n, given that alternative i was chosen, corresponds to

$$\pi_n(D \mid i) = \begin{pmatrix} J_{m(i)} - 1 \\ \tilde{J}_{m(i)} - 1 \end{pmatrix}^{-1} \begin{pmatrix} J_{m' \neq m(i)} \\ \tilde{J}_{m' \neq m(i)} \end{pmatrix}^{-1},$$

where *m'≠m(i)* is the nest to which *i* does not belong and the expression on parenthesis corresponds to the binomial coefficient.

It can be shown that

$$\binom{J_{m(i)}-1}{\tilde{J}_{m(i)}-1}=\frac{(J_{m(i)}-1)!}{(\tilde{J}_{m(i)}-1)!(J_{m(i)}-1-(\tilde{J}_{m(i)}-1))!}=\frac{\tilde{J}_{m(i)}}{J_{m(i)}}\binom{J_{m(i)}}{\tilde{J}_{m(i)}},$$

and therefore, the conditional probability of constructing the set D_n, given that alternative *i* was chosen, corresponds to

$$\pi_n(D\mid i)=\frac{J_{m(i)}}{\tilde{J}_{m(i)}}\left[\binom{J_1}{\tilde{J}_1}^{-1}\binom{J_2}{\tilde{J}_2}^{-1}\right]. \tag{18}$$

Given that the second term in Eq. (18) does not vary across alternatives, it will cancel out when taking the log to calculate the sampling correction $\ln\pi(D_n\mid i)$. Then, the estimator of the conditional probability of choosing alternative *i*, given that the set D_n was constructed, will correspond to Eq. (19)

$$\hat{\pi}(i\mid D_n)=\frac{e^{V(x_{in},\beta)+\ln f\left(\hat{B}_{in}(D_n)\right)+\ln\frac{J_{m(i)}}{\tilde{J}_{m(i)}}}}{\sum_{j\in D_n}e^{V(x_{jn},\beta)+\ln f\left(\hat{B}_{jn}(D_n)\right)+\ln\frac{J_{m(j)}}{\tilde{J}_{m(j)}}}}, \tag{19}$$

where $\ln f\left(\hat{B}_{in}(D_n)\right)=\left(\frac{\mu}{\mu_{m(i)}}-1\right)\left(\ln\sum_{j\in D_{m(i)n}}w_{jn}e^{\mu_{m(i)}V_{jn}}\right)+\ln\mu+(\mu_{m(i)}-1)V_{in}$.

The final step corresponds to the specification of the expansion factors w_{jn}. This task is substantially different depending on whether the set D_n (used for the sampling correction) is or is not also used for the expansion of the sum of the exponentials.

Consider first that the set D_n is also used for the expansion of the sum of the exponentials. Then, given that the sampling protocol is without replacement, the numerator in Eq. (17) will equal 1. $E(\tilde{n}_{jn})$, the expected number of times alternative *j* might be sampled to construct the set D_n, remains to be calculated. Given that the protocol is without replacement, $E(\tilde{n}_{jn})$ corresponds to the probability of sampling alterative *j*.

$E(\tilde{n}_{jn})$ can be calculated using the Law of Total Expectations. The idea is to divide the space into mutually exclusive and totally exhaustive events with known probabilities of occurrence, and for which the conditional expectation of $\tilde{n}_{jn}$ is also known. Consider the following events:

A_1: The chosen alternative is *j*

A_2: The chosen alternative is not *j*, but it is within those in the nest *m*(*j*)

A_3: The chosen alternative does not belong to the nest *m*(*j*).

The events A_1, A_2 and A_3 are totally exhaustive and mutually exclusive because only one alternative is chosen and the nests in the Nested Logit model are mutually exclusive and totally exhaustive. The probabilities of these three events depend on the choice probabilities:

$P(A_1) = P_n(j)$: The probability of choosing alternative *j*.

$P(A_2) = \sum_{\substack{l \in C_{m(j)} \\ l \neq j}} P_n(l)$: The probability of choosing other alternatives in *m*(*j*), which is equal to the sum of their choice probabilities.

$P(A_3) = 1 - \sum_{l \in C_{m(j)}} P_n(l)$: The probability of choosing an alternative outside *m*(*j*), which is equal to 1 minus the probability of the nest *m*(*j*).

The conditional expectations of $\tilde{n}_{jn}$ given the events A_1, A_2 and A_3 are also known:

$E(\tilde{n}_{jn} \mid A_1) = 1$: Because the chosen alternative is always sampled.

$E(\tilde{n}_{jn} \mid A_2) = \frac{\tilde{J}_{m(j)} - 1}{J_{m(j)} - 1}$: Because if *j* is not chosen, but the chosen alternative is in *m*(*j*), only $\tilde{J}_{m(j)} - 1$ out of $J_{m(j)} - 1$ alternatives remain to be sampled from the nest *m*(*j*).

$E(\tilde{n}_{jn} \mid A_3) = \frac{\tilde{J}_{m(j)}}{J_{m(j)}}$: Because if the chosen alternative is in not in *m*(*j*), $\tilde{J}_{m(j)}$ out of $J_{m(j)}$ alternatives remain to be sampled from the nest *m*(*j*).

Then, by the Law of Total Expectations, the expected number of times alternative *j* might be drawn will correspond to

$$E(\tilde{n}_{jn}) = E(\tilde{n}_{jn} \mid A_1)P(A_1) + E(\tilde{n}_{jn} \mid A_2)P(A_2) + E(\tilde{n}_{jn} \mid A_3)P(A_3).$$

By replacing terms, Eq. (20) is finally obtained.

$$E(\tilde{n}_{jn}) = P_n(j) + \frac{\tilde{J}_{m(j)} - 1}{J_{m(j)} - 1} \sum_{\substack{l \in C_{m(j)} \\ l \neq j}} P_n(l) + \frac{\tilde{J}_{m(j)}}{J_{m(j)}} \left(1 - \sum_{l \in C_{m(j)}} P_n(l) \right) \tag{20}$$

The expression shown in Eq. (20) for the denominators of the expansion factors depends on the choice probabilities, which are unknown beforehand in an application with real data. In section 6.3, we analyze alternatives to achieve this goal in practice.

Consider now the case when a set D_n is used for the sampling correction $\ln \pi(D_n \mid i)$, and a different set $\tilde{D}_n$ is drawn to construct the expansion of the sum of the exponentials. We term this alternative procedure re-sampling. In this case, the conditional probability of choosing alternative *i*, given that the sets D_n and $\tilde{D}_n$ were drawn, will correspond to Eq. (21).

$$\hat{\pi}(i \mid D_n, \tilde{D}_n) = \frac{e^{V(x_{in},\beta) + \ln f\left(\hat{B}_{in}(\tilde{D}_n)\right) + \ln \frac{J_{m(i)}}{\tilde{J}_{m(i)}}}}{\sum_{j \in D_n} e^{V(x_{jn},\beta) + \ln f\left(\hat{B}_{jn}(\tilde{D}_n)\right) + \ln \frac{J_{m(j)}}{\tilde{J}_{m(j)}}}} \tag{21}$$

As stated before, the set D_n must include the chosen alternative. Otherwise, the quasi-log-likelihood of the model may become unbounded, making impossible the estimation of the model parameters. In turn, the set $\tilde{D}_n$ used for the expansion of the sum of the exponentials in Eq. (21) does not need to include the chosen alternative, as long as D_n does it. This small difference is relevant because, if the sampling protocol used to build the set $\tilde{D}_n$ does not require drawing the chosen alternative forcedly, there is no need for knowing the choice probabilities beforehand to calculate the expansion factors w_{jn}.

The implementation of the expansion method in practice when $\tilde{D}_n$ does not require drawing the chosen alternative forcedly becomes then considerably simpler. Consider for example that the sampling protocol used to build the set $\tilde{D}_n$ was importance sampling without replacement by nest. Under this setting, the denominators of the expansion factors, the equivalent to Eq. (20), would simply be the ratio shown in Eq. (22), where $\tilde{\tilde{J}}_{m(j)}$ corresponds to the cardinality of $\tilde{D}_n$.

$$E\left(\tilde{n}_{jn}\right) = \frac{\tilde{\tilde{J}}_{m(j)}}{J_{m(j)}} \tag{22}$$

6.2 Assessment of the Methods with and without Re-sampling

Given this Monte Carlo experiment, the sampling protocol described and the expansion proposed, five models were estimated and the results are shown in Table 1. The first (*No Sampling* in Table 1) corresponds to the true model, where no sampling was applied. This model is estimated as a benchmark for the best possible estimators that could be expected for this particular experiment.

Table 1 Monte Carlo Experiment: Sampling in MEV with and without Re-Sampling

Experiments		No Sampling		Full $\ln G_{in}$		Unexpanded		Expanded True Prob.		Expanded Re-Sampling	
		est.	s.e	est.	s.e	est.	s.e	est.	s.e	est.	s.e
$\tilde{J}_1 = 5$ $\tilde{J}_2 = 5$	β_{x_1}	1.009	0.04681	0.9906	0.06112	2.570	0.1612	0.9102	0.06020	0.9301	0.06705
	β_{x_2}	1.062	0.04933	1.027	0.06253	2.630	0.1649	0.9276	0.06124	0.9558	0.06818
	μ_1	2.055	0.2076	2.111	0.2289	0.2655	0.006477	2.211	0.2688	1.976	0.2913
	μ_2	2.824	0.1125	2.881	0.1291	1.130	0.07562	3.313	0.1786	2.853	0.1567
	$L(\hat{\beta})$	-10,312.09		-1,942.70		-2,036.24		-1,968.59		-2,030.30	
	$L(0)$	-13,825.49		-4,605.17		-4,605.17		-4,605.17		-4,605.17	
	$\bar{\rho}^2$	0.2544		0.5790		0.5587		0.5734		0.5583	
$\tilde{J}_1 = 5$ $\tilde{J}_2 = 500$	β_{x_1}	1.009	0.04681	1.005	0.04678	0.7534	0.04708	1.005	0.04679	1.004	0.04679
	β_{x_2}	1.062	0.04933	1.055	0.04915	0.7913	0.04950	1.056	0.04918	1.055	0.04917
	μ_1	2.055	0.2076	2.065	0.2088	2.730	0.3086	2.063	0.2088	2.065	0.2091
	μ_2	2.824	0.1125	2.832	0.1130	3.785	0.2186	2.831	0.1131	2.834	0.1133
	$L(\hat{\beta})$	-10,312.09		-9,115.24		-9,117.40		-9,115.91		-9,115.37	
	$L(0)$	-13,825.49		-12,449.12		-12,449.12		-12,449.12		-12,449.12	
	$\bar{\rho}^2$	0.2544		0.2681		0.2679		0.2681		0.2675	

N=2,000. $J_1 = 5$, $\tilde{J}_1 = \tilde{\tilde{J}}_1 = 5$; $J_2 = 1{,}000$ $\tilde{J}_2 = \tilde{\tilde{J}}_2 = 5$ and 500

The second model (*Full* $\ln G_{in}$ in Table 1) corresponds to the application of sampling of alternatives and the corresponding sampling correction, but using the full choice-set to

evaluate the term $\ln G_{in}$, as shown in Eq. (9). Even though this model is impractical because it requires knowledge of the full choice-set, it was estimated to show that Eq. (9) is correct, and to quantify and to differentiate the effects of sampling of alternatives, when having a reduced choice-set, from its effects in the approximation of $\ln G_{in}$.

The third model estimated (*Unexpanded* in Table 1) considers that a set D_n was sampled from the full choice-set C_n, that the corresponding sampling correction was applied, and that the same set D_n was used to construct the term $\ln G_{in}$, without any expansion term. This model acts as a benchmark because it corresponds to what has been used to date by the researchers to estimate Nested Logit models under sampling of alternatives (see, e.g., Berkovec and Rust, 1985; Train *et al.*,1987; Hansen, 1987; and Rivera and Tiglao, 2005).

The fourth model estimated (*Expanded True Prob.* in Table 1) corresponds to the method proposed for cases where the same set D_n is used for the sampling correction and for the expansion of $\ln G_{in}$ using Eq. (20). The calculation of Eq. (20) involves knowledge of the choice probabilities, which are unknown beforehand in a real application. However, in this Monte Carlo experiment the true choice probabilities are available beforehand and are therefore used to show the performance of the method proposed for the expansion of the sum of the exponentials.

The last model estimated (*Expanded Re-sampling* in Table 1) corresponds to the method proposed for cases where a set D_n is used for the sampling correction, and a different set $\tilde{\tilde{D}}_n$ (constructed independently from the chosen alternative) is used for the expansion of $\ln G_{in}$ using Eq. (22). For fair comparison with other models, the number of alternatives considered in the set $\tilde{\tilde{D}}_n$ is the same as that used for the set D_n; that is, $\tilde{\tilde{J}}_n = \tilde{J}_n$.

The first result that should be noted in Table 1 is that, as expected, all estimated parameters for the *No Sampling* and *Full* $\ln G_{in}$ models are statistically equal (with 95% confidence) to the true values. Regarding *Full* $\ln G_{in}$, note that, as the sample size increases, the standard error of the estimators is reduced as a result of the increment in the number of cases $N\left(\tilde{J}-1\right)$. In other words, efficiency increased as more information became available.

Regarding the model *Unexpanded*, note that for $\tilde{J}_2 = 5$, the model estimates are very far from the true values. Remarkably, one of the scale parameters is even below one, which makes this result inconsistent with utility maximization (Ben-Akiva and Lerman, 1985). The bias in this model is reduced substantially for $\tilde{J}_2 = 500$. This occurs because the *Unexpanded* formulation collapses to the true model as the sample size increases. However, even for $\tilde{J}_2 = 500$, the estimators are still statistically different (with 95% confidence) from the true values.

In the case of the *Expanded True Prob.* method, all estimates in Table 1 are remarkably better than those of the *Unexpanded* model and statistically equal (with 95%

confidence) to the true ones with 95% confidence, even for $\tilde{J}_2 = 5$. For the bias, note that it is not negligible for $\tilde{J}_2 = 5$, but for $\tilde{J}_2 = 500$, it is significantly reduced.

Figure 2 shows the evolution of the estimators as $\tilde{J}_2$ is increased for the model *Expanded True Prob*. As $\tilde{J}_2$ approaches J_2, the estimators of the model collapse to those of the *No Sampling* model. Remarkably, the estimators quickly stabilize for $\tilde{J}_2$ below 100 and are never far from the true values. As shown in Table 1, even for a sample size as small as $\tilde{J}_2 = 5$, all the estimators are statistically equal (with 95% confidence) to the true values.

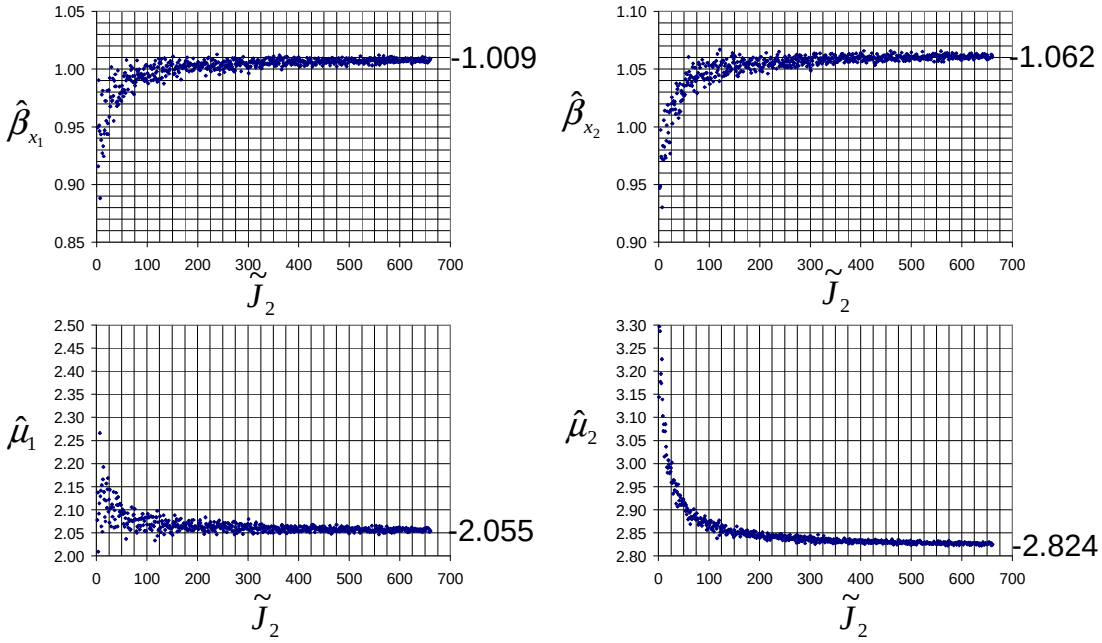

Figure 2 Monte Carlo Experiment: Estimators as $\tilde{J}_2$ Increases. Expanded True Prob.

Figure 2 is also useful for analyzing the small sample bias. First, note that the parameter that has the poorest convergence behavior (larger variance and slope) is $\hat{\mu}_2$, the scale of the second nest. It can be hypothesized that this occurs because sampling is performed only from the second nest in this experiment. Figure 2 also shows that both scales $\hat{\mu}_1$ and $\hat{\mu}_2$ are biased upward and that the model parameters $\hat{\beta}$ are biased downward. The experiments analyzed did not allow proposing hypotheses to explain this result. Further analysis of the finite sample properties of this estimator, and potential ways to improve them, are left for future research.

Finally, the last column in Table 1 shows that the results for the *Expanded Re-Sampling* method are statistically equal (with 95% confidence) to those obtained by using the *Expanded True Prob.* method, and also statistically equal (with 95% confidence) to the true values. This implies that if re-sampling to perform the expansion of the sum of the exponentials is possible, it should be preferred because it avoids approximating the

choice probabilities. In the next section, we analyze the performance of different procedures that can be used in practice when re-sampling is not possible.

6.3 Expansion in Practice when Re-sampling is not Possible

When re-sampling is not possible the results of the method for sampling of alternatives in MEV shown in Table 1, require knowledge of the choice probabilities, which are not available in an application with real data. To avoid this problem, three methods used to approximate the choice probabilities are examined and the results are summarized in Table 2.

Table 2 Monte Carlo Experiment: Different Estimators of Choice Probabilities

Experiments		Expanded True Prob.		Expanded All or Nothing		Expanded Population Shares		Expanded Iterative Prob.	
		est.	s.e	est.	s.e	est.	s.e	est.	s.e
$\tilde{J}_1 = 5$ $\tilde{J}_2 = 5$	β_{x_1}	0.9102	0.06020	0.7440	0.05335	1.133	0.06906	0.9444	0.06528
	β_{x_2}	0.9276	0.06124	0.7565	0.05417	1.158	0.07020	0.9630	0.06641
	μ_1	2.211	0.2688	2.787	0.3327	1.685	0.2151	2.031	0.2734
	μ_2	3.313	0.1786	4.328	0.2817	2.714	0.1251	3.210	0.1808
	$L(\hat{\beta})$	-1,968.59		-1,864.44		-1,982.65		-1,991.85	
	$L(0)$	-4,605.17		-4,605.17		-4,605.17		-4,605.17	
	$\bar{\rho}^2$	0.5734		0.5960		0.5703		0.568	
$\tilde{J}_1 = 5$ $\tilde{J}_2 = 500$	β_{x_1}	1.005	0.04679	1.005	0.04673	1.007	0.04681	1.005	0.04679
	β_{x_2}	1.056	0.04918	1.055	0.04912	1.058	0.04920	1.056	0.04918
	μ_1	2.063	0.2088	2.066	0.2088	2.059	0.2083	2.063	0.2088
	μ_2	2.831	0.1131	2.833	0.1131	2.825	0.1125	2.831	0.1130
	$L(\hat{\beta})$	-9,115.91		-9,114.88		-9,115.92		-9,115.92	
	$L(0)$	-12,449.12		-12,449.12		-12,449.12		-12,449.12	
	$\bar{\rho}^2$	0.2681		0.2682		0.2681		0.2681	

N=2,000. $J_1 = 5, \tilde{J}_1 = 5; J_2 = 1{,}000$ $\tilde{J}_2 = 5$ and 500

One alternative is to approximate the probability of the chosen alternative to equal 1, and the probability of the non-chosen alternatives to equal zero. This model is termed *Expanded All or Nothing* in Table 2. Replacing these assumptions in Eq. (20), the expansion factors used in this case will correspond to the following:

$w_{jn} = 1$ if *j* is the chosen alternative

$w_{jn} = \dfrac{J_{m(j)} - 1}{\tilde{J}_{m(j)} - 1}$ if *j* is not chosen, but another alternative in *m(j)* is chosen

$w_{jn} = \frac{J_{m(j)}}{\widetilde{J}_{m(j)}}$ if j is not chosen, and no other alternative in *m(j)* is chosen.

The expansion factors that result in this case are equivalent to those used by Frejinger *et al.* (2009) to approximate the denominator of a Logit model with sampling of alternatives, and to those used by Lee and Waddell (2010) to expand a Nested Logit model under sampling of alternatives. That is, although it is not mentioned by those authors, they implicitly approximated the probability of the chosen alternative to 1, and the probability of the non-chosen alternatives to 0.

A second possibility to approximate the choice probabilities needed for the calculation of the expansion factors is to use the population shares of each alternative. Although the true population shares are not available in a real application, good approximations of them are clearly plausible from different sources (Census data for spatial choice models or flow counts in route choice modeling). This method is termed *Expanded Population Shares* in Table 2. Replacing the population shares in Eq. (20), the expansion factors implied by this procedure are the following:

$W_j = \text{population share of alternative } j$

$$w_{jn} = \frac{1}{W_j + \frac{\widetilde{J}_{m(j)} - 1}{J_{m(j)} - 1} \sum_{\substack{l \in C_{m(j)n} \\ l \neq j}} W_l + \frac{\widetilde{J}_{m(j)}}{J_{m(j)}} \left(1 - \sum_{l \in C_{m(j)n}} W_l \right)} \qquad \forall n = 1, \cdots, N; \forall j \in C_n .$$

Finally, an iterative method can be proposed. This method starts with an estimation of the population shares of each alternative, and then estimates the choice probabilities for each observation, iteratively, until convergence. This method is termed *Expanded Iterative Prob.* in Table 2 and can be summarized as follows.

Step 0:

k=0

$W_j = \text{population share of alternative } j$

$$w_{jn}^{k} = \frac{1}{W_j + \frac{\widetilde{J}_{m(j)} - 1}{J_{m(j)} - 1} \sum_{\substack{l \in C_{m(j)n} \\ l \neq j}} W_l + \left(1 - \sum_{l \in C_{m(j)n}} W_l \right) \frac{\widetilde{J}_{m(j)}}{J_{m(j)}}} \quad \forall n = 1, \cdots, N; j \in C_n$$

Step 1:

Estimate the model using w_{jn}^{k} to obtain $\hat{\beta}$ and $\hat{P}_n^k(j) = \frac{e^{\hat{V}(x_{jn}, \hat{\beta}) + \ln f\left(\hat{B}_{in}(w^k)\right)}}{\sum_{l \in D_n} w_{ln}^{k} e^{\hat{V}(x_{ln}, \hat{\beta}) + \ln f\left(\hat{B}_{jn}(w^k)\right)}}$

Step 2:

$$w_{jn}^{k+1} = \frac{1}{\hat{P}_n^k(j) + \frac{\widetilde{J}_{m(j)} - 1}{J_{m(j)} - 1} \sum_{\substack{l \in D_{m(j)n} \\ l \neq j}} w_{ln}^{k} \hat{P}_n^k(l) + \frac{\widetilde{J}_{m(j)}}{J_{m(j)}} \left(1 - \sum_{l \in D_{m(j)n}} w_{ln}^{k} \hat{P}_n^k(l) \right)}$$

Step 3:

k=k+1

Go to Step 1 until convergence.

Convergence can be stated in terms of the estimated parameters of the model, the expansion factors, or the choice probabilities. For the applications of the iterative procedure in this research, the following stopping criterion was used:

$\max_{n,j}\left|\hat{P}_n^k(j)-\hat{P}_n^{k+1}(j)\right| \le 1/(10J)$.

The three methods proposed to approximate the choice probability when re-sampling is not possible were used in the estimation of the problem of sampling of alternatives for the Nested Logit model described in Figure 1. Table 2 shows the results of the three methodologies, compared to the results obtained with the *Expanded True Prob.* method.

Consider the case of the *Expanded All or Nothing* and the *Expanded Population Shares* procedures. Table 2 shows that for $\tilde{J}_2 = 5$, the estimators of both methods are statistically different (with 95% confidence) to the true ones. Although, comparing these results with those of the *Unexpanded* method reported in Table 1, it should be noted that the new estimators have a smaller bias. For $\tilde{J}_2 = 500$, the *Expanded All or Nothing* and the *Expanded Population Shares* estimators are statistically equal (with 95% confidence) to those obtained by using the *Expanded True Prob.* method, and also statistically equal (with 95% confidence) to the true values.

Finally, for the *Expanded Iterative Prob.* method, Table 2 shows that for $\tilde{J}_2 = 5$ and $\tilde{J}_2 = 500$ the estimates are statistically equal (with 95% confidence) to those obtained using the *Expanded True Prob.* method, and also statistically equal (with 95% confidence) to the true values. This implies that, when re-sampling is not possible, the iterative procedure proposed to approximate the choice probabilities should be preferred to expand the sum of the exponentials.

6.4 Additional Experiments

In this section, we present four additional experiments to illustrate the performance of the proposed method for addressing sampling of alternatives in MEV models, under different circumstances.

The first three experiments explore the effect of the distribution of the data. These experiments consider the same structure described in Figure 1. The only difference is that the distributions of attributes x_1 and x_2 vary across observations. Under this setting, the estimators of the model parameters were obtained for 100 repetitions using the *Expanded True Prob.* method and for different values of $\tilde{J}_2$. Table 3 reports the bias, mean squared error (MSE) and t-test against the true value of the scale of the second nest $\hat{\mu}_2$ for each experiment.

The first experiment is termed *Uniform Mixture*. For the first 1,000 observations, x_1 was drawn from an *iid* Uniform (-1,1) distribution and x_2 from an *iid* Uniform (-1.5,1.5) distribution. For the second half of the observations, x_1 was drawn from an *iid* Uniform (0,2) distribution and x_2 from an *iid* Uniform (-3,1) distribution. The results of this model are shown in the first column (after the labels) of Table 3. It can be noted that the sample size required to obtain an estimator of $\hat{\mu}_2$ statistically equal (with 95% confidence) to its true value is larger than 25 alternatives in this case. This value is larger than that obtained

for the experiment reported in Table 1 and confirms that the threshold required for attaining valid estimates of the model parameters depends on the data.

The second experiment is termed *Varying* $\tilde{J}_2$. This experiment considers the same structure and distribution of the data used in the *Uniform Mixture* experiment. The only difference is that the number of drawn alternatives varies across individuals following a Discrete Uniform distribution with limits

$$\left[\left\lfloor \tilde{J}_2/2 \right\rfloor, \left\lceil 2\tilde{J}_2 \right\rceil\right].$$

Then, for example, for $\tilde{J}_2$=10 in Table 3, the number of alternatives considered for each of the 2,000 observations can be any integer between 5 and 20, with equal probability. The results of this experiment are shown in the second column of Table 3. Although this experiment is not directly comparable with the *Uniform Mixture* setting, it can be affirmed that the fact that, in both cases, sample sizes around 25 were large enough to obtain an estimator of the scale of the second nest that was statistically equal (with 95% confidence) to its true value, is evidence that varying the sample size across observations causes only minor impacts in the estimation procedure.

Table 3 Monte Carlo Experiment: Additional Experiments on Sampling in MEV

$\hat{\mu}_2$	**Uniform Mixture**			**Varying** $\tilde{J}_2$			**Normal Uniform**		
$\tilde{J}_2$	**Bias**	**MSE**	**t-test true**	**Bias**	**MSE**	**t-test true**	**Bias**	**MSE**	**t-test true**
10	0.4878	0.2624	3.125	0.4502	0.2317	2.641	1.359	1.989	3.616
25	0.2760	0.09404	2.065	0.2623	0.08881	1.854	1.063	1.223	3.493
50	0.1512	0.03729	1.259	0.1465	0.03799	1.139	0.7938	0.6914	3.206
100	0.07260	0.01733	0.6612	0.07413	0.02034	0.6085	0.5505	0.3516	2.498
250	0.01492	0.01156	0.1402	0.01941	0.01369	0.1682	0.2968	0.1228	1.592
500	-0.006329	0.01105	-0.06033	0.0002020	0.01305	0.001768	0.1473	0.04756	0.9155

N=2,000. $J_1 = 5$, $J_2 = 1{,}000$ $\tilde{J}_1 = 5$; Average and variance from 100 repetitions. *Expanded True Prob.*

The third experiment is termed *Normal Uniform*. In this case x_1 is *iid* Normal (0,1) for the first 1,000 observations and Normal (1,2) for the rest. In turn x_2 *iid* Uniform (1,3) for the first 1,000 observations and Uniform (0,4) for the rest. The results of this experiment are shown in the third column of Table 3. This experiment shows that the sample size required to attain estimators that are statistically equal (with 95% confidence) to the true values is now between 100 and 250. This result is further evidence that the performance of the method can be significantly affected by the distribution of the data.

The fourth experiment sheds light on whether or not the sample size required to attain a desirable bias can be stated as a percentage of the cardinality of the true choice-set. The experiment described in Figure 1 was modified only regarding the number of alternatives in the second nest, which is 1,000,000 in this case. The distribution of x_1 and x_2 are again *iid* Uniform (-1,1) for the *N*=2,000 observations. The model is described in Figure 3 and the results are reported in Table 4.

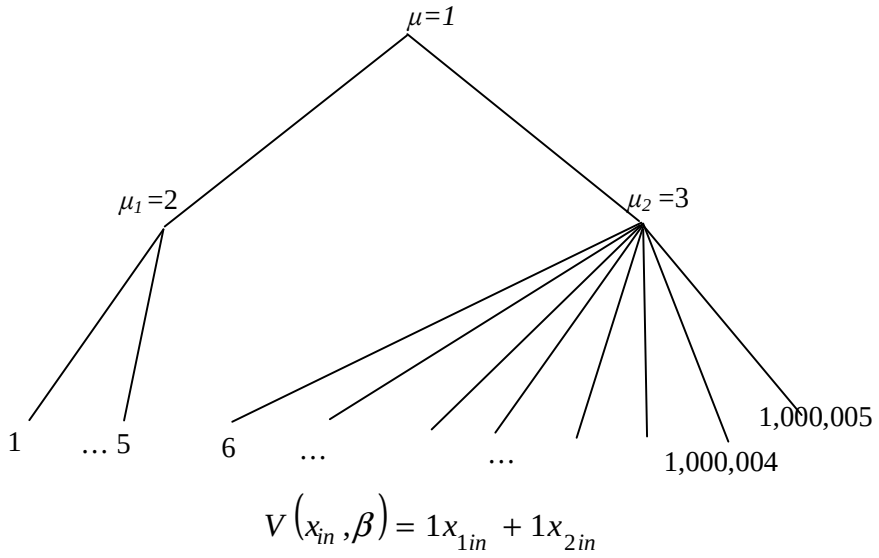

Figure 3 Monte Carlo Experiment: Nesting Structure. 1,00,005 Alternatives

Table 4 Monte Carlo Experiment: Sampling in MEV. 1,000,005 Alternatives

Experiments		True Values	Unexpanded		Expanded True Prob.	
			Est.	s.e	est.	s.e
$\tilde{J}_1=5$ $\tilde{J}_2=5$	β_{x_1}	1	2.947	0.3594	0.9403	0.07193
	β_{x_2}	1	2.820	0.3412	0.9118	0.06894
	μ_1	2	0.1427	0.003651	1.877	0.5237
	μ_2	3	1.073	0.1322	3.372	0.2203
	$L(\hat{\beta})$		-1,348.82		-1,341.87	
	$L(0)$		-3,670.51		-3,670.51	
	$\bar{\rho}^2$		0.6336		0.6355	
$\tilde{J}_1=5$ $\tilde{J}_2=500$	β_{x_1}	1	1.887	0.4404	1.014	0.05930
	β_{x_2}	1	1.784	0.4162	0.9629	0.05586
	μ_1	2	0.1896	0.02426	1.836	0.455
	μ_2	3	1.645	0.3837	3.054	0.162
	$L(\hat{\beta})$		-9,253.96		-9,241.78	
	$L(0)$		-12,710.46		-12,710.46	
	$\bar{\rho}^2$		0.2723		0.2732	

N=2,000. $J_1=5, \tilde{J}_1=\tilde{\tilde{J}}_1=5; J_2=1{,}000{,}000\ \tilde{J}_2=\tilde{\tilde{J}}_2=5$ and 500

In this case the true model is not estimatable with commercial software because the computational costs are unbearable. In turn, it is possible to simulate independently the choices for each observation, then to sample a small number of alternatives from the true choice-set, and store that subset of information, for subsequent estimation. Using this procedure, samples of 5 and 500 alternatives were drawn from the second nest.

Table 4 contrasts the estimators that are obtained using the *Unexpanded* and *Expanded True Prob.* methods. Similar to what occurred in the experiments reported in Table 1, the estimators of the *Expanded True Prob.* method are also statistically equal (with 95% confidence) to their true values, even for a sample size as small as 5. However, comparing Table 1 with Table 4, it can be noted that the confidence is smaller in the case where the true choice-set has 1,000,005 alternatives.

Given that the quality of the estimators obtained with samples of 5 and 500 are qualitatively equal when the cardinality of the true choice-set is 1,005 or 1,000,005, it can be affirmed that there is evidence that the sample size required to obtain acceptable estimators is independent of the true cardinality of the choice-set.

7 Application to Real Data

The final step corresponds to the demonstration of the method proposed for sampling of alternatives and estimation in MEV models using real data. The case study corresponds to a residential location choice model situated in the Portuguese municipalities of Lisbon, Odivelas and Amadora, which are located at the center of the Lisbon Metropolitan Area (LMA).

The data to estimate the model was constructed using the combination of two sources. The first source was a small convenience online survey (SOTUR) conducted in 2009 by Martinez *et al.* (2010) in the LMA. The second source corresponds to a snapshot of the dwellings that were advertised for sale in February 2007 within the municipalities of Lisbon, Odivelas and Amadora (Martinez and Viegas, 2009). The details on the construction of the database by matching both sources can be found in Guevara (2010).

The database is compounded of 11,501 alternatives, from which only 63 correspond to chosen dwellings. The main descriptive statistics of the database are shown in Table 5. Regarding dwelling attributes, Table 5 shows that dwellings from the Lisbon municipality tend to be more expensive and older than those from Odivelas and Amadora, although the differences are not statistically significant (with 95% confidence). Also, the dwellings from both regions have approximately equal area. Finally, dwellings from Lisbon are significantly closer, in average, to the workplace of the head-of-the-households of the sample. Table 5 also shows the distribution of household location, classified by income. It should be noted that 51 out of 63 households reside in Lisbon municipality and that the larger share of households in the sample have an income that is between 2,000 and 5,000 Euros per month (€/M).

Using the database described in Table 5 we considered a Nested Logit model allowing for correlation between alternatives on a geographic base. The structure used is shown in Figure 4. We considered one nest for the 3,483 alternatives that belong to the Municipalities of Odivelas and Amadora, and the other 8,018 alternatives from the Municipality of Lisbon, were considered to belong to the root of this Nested Logit model.

The nesting structure used is simple principally because of the small number of observations available. More interesting structures, such as multilevel nests by area, were impossible to estimate. However, the nesting structure considered does serve well its main purpose of demonstrating the methodology for sampling of alternatives and estimation developed in this research. Furthermore, despite its simplicity, the nesting structure is concordant with what is observed in the city. The municipalities of Odivelas and Amadora are approximately what Rayle (2008)[2] defined (using a factor-analysis approach) as the "Inner Periphery" of the central LMA, a sector that has marked differences with the Lisbon's Municipality.

Table 5 Summary of Lisbon's Residential Location Choice Database for Estimation

Municipality	Average Dwelling Attributes (Standard Deviation)				Total Dwellings Available	Household Location			
	Price 100,000 [€]	Distance to Workplace [Km]	Area [m^2]	Age [Years]		Income <2,000 [€/M]	Income 2,000-5,000 [€/M]	Income >5,000 [€/M]	Tot.
Lisbon	2.356 (1.354)	4.508 (2.389)	99.30 (41.77)	39.93 (36.21)	8,018	16	28	7	51
Odivelas and Amadora	1.680 (0.8365)	10.581 (1.253)	98.44 (32.59)	32.17 (31.68)	3,483	5	7	0	12
Total	2.151 (1.260)	6.347 (3.499)	99.01 (39.22)	37.58 (35.08)	11,501	21	35	7	63

€/M: Euros per month. Standard errors in parenthesis.

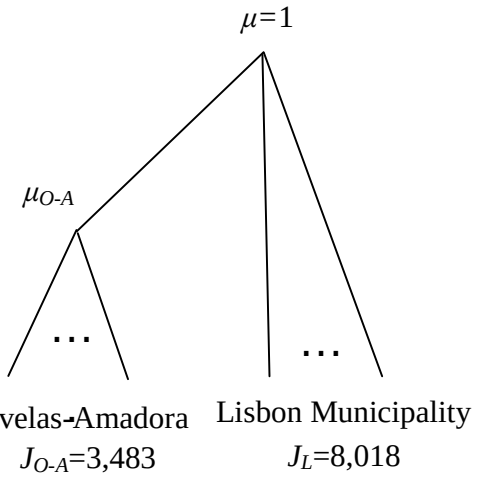

Figure 4 Nesting Structure of Lisbon's Nested Logit Residential Location Choice Model

[2] Unpublished Manuscript.

Under this setting, a Nested Logit model was estimated with the assumption that the 11,501 alternatives corresponded to the true choice-set. This model considered the correction for endogeneity caused by the omission of attributes using the Two-stage-control-function (2SCF) method (Hausman, 1978; Heckman, 1978), as it is described in detail by Guevara (2010).

The results of this model are reported in the second column of Table 6 and are repeated in Table 7. It should be noted that the signs of the coefficients of the models are as expected. The coefficient of dwelling area ($\hat{\beta}_5$) is positive, meaning that households prefer larger dwellings. The contrary occurs with dwelling price ($\hat{\beta}_1$), age ($\hat{\beta}_6$), and distance to workplace of the head-of-the-household ($\hat{\beta}_4$), which are perceived negatively. Also, the impact of dwelling price decreases with household income since $\hat{\beta}_2, \hat{\beta}_3 > 0$, but is negative for all stratum since $\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3$ <0 for all cases. Finally, it should be noted that the scale of the nest is statistically different (with 95% confidence) from 1, what implies that the null hypothesis that the choice model is a Logit, is rejected.

Table 6 Corrected and Uncorrected Estimators for Lisbon's Nested Logit Model

Sampling 5 + 5 Alternatives

Variables	No Sampling		Unexpanded		Expanded Iterative Prob.	
	$\hat{\beta}$	s.e	$\hat{\beta}$	s.e	$\hat{\beta}$	s.e
1. Dwelling price (in 100,000 €)	-4.393	0.7058	-3.095	0.6498	-5.374	0.8947
2. Dwelling price * 1[Income > 2,000 €/M]	1.213	0.5769	1.291	0.4756	1.834	0.7048
3. Dwelling price * 1[Income > 5,000 €/M]	0.9463	0.5284	0.5298	0.5364	0.7604	0.6779
4. Distance to Workplace (in Km)	-0.1774	0.0538	-0.1617	0.0528	-0.1732	0.0639
5. Log [Dwelling Area (in m^2)]	4.217	0.7854	2.220	0.5530	4.454	1.0324
6. Log [Dwelling Age (in years) +1]	-0.6381	0.1158	-0.4850	0.1180	-0.7252	0.1604
7. $\hat{\delta}$ Control-function Aux. Var.	1.987	0.4711	0.6193	0.3864	2.145	0.5763
8. $\mu_{O\text{-}A}$ Odivela-Amadora Nest	1.329	0.09414	5.480	3.053	1.392	0.1266
Log likelihood at Convergence $L(\hat{\beta}, \hat{\mu})$	-547.89		-94.96		-93.53	
Log likelihood at Zero $L(\hat{\beta}=0, \hat{\mu}=1)$	-589.06		-134.13		-134.13	
Adjusted ρ^2	0.08518		0.3666		0.3623	
Sample Size N	63		63		63	
Choice-set Size J	11,501		10		10	
Estimation Time (in seconds]	363.0		1.080		10.65	

Nest Amadora and Odivelas. Root Lisbon municipality. Models include sampling correction. Models corrected for endogeneity with 2SCF. Sample 5 alts. from Odivelas-Amadora nest and 5 from Lisbon municipality. €/M: Euros per month.

To demonstrate the method proposed in this paper to achieve sampling of alternatives and estimation in MEV models, we performed two experiments where we sampled a set of alternatives in the choice-set and then re-estimated the model with and without the

expansion of the sum of the exponentials proposed in this paper. The sampling protocol used in the first experiment was the following. First, the chosen alternative was included. Then, alternatives were randomly drawn from the Odivelas-Amadora nest and from the root (Lisbon) up to make a total of 5 alternatives for each case.

Table 7 Corrected and Uncorrected Estimators for Lisbon's Nested Logit Model

Sampling 500 + 500 Alternatives

Variables	No Sampling		Unexpanded		Expanded Iterative Prob.	
	$\hat{\beta}$	s.e	$\hat{\beta}$	s.e	$\hat{\beta}$	s.e
1. Dwelling price (in 100,000 €)	-4.393	0.7058	-4.349	0.6780	-4.347	0.7054
2. Dwelling price * 1[Income > 2,000 €/M]	1.213	0.5769	1.242	0.5649	1.184	0.5776
3. Dwelling price * 1[Income > 5,000 €/M]	0.9463	0.5284	0.9566	0.5290	0.9923	0.5333
4. Distance to Workplace (in Km)	-0.1774	0.0538	-0.1766	0.05288	-0.1811	0.05380
5. Log [Dwelling Area (in m^2)]	4.217	0.7854	4.177	0.7450	4.223	0.7902
6. Log [Dwelling Age (in years) +1]	-0.6381	0.1158	-0.6362	0.1123	-0.6321	0.1161
7. $\hat{\delta}$ Control-function Aux. Var.	1.987	0.4711	1.908	0.4460	1.937	0.4683
8. $\mu_{O\text{-}A}$ Odivela-Amadora Nest	1.329	0.09414	1.510	0.1618	1.326	0.09340
Log likelihood at Convergence $L(\hat{\beta},\hat{\mu})$	-547.89		-382.38		382.95	
Log likelihood at Zero $L(\hat{\beta}=0,\hat{\mu}=1)$	-589.06		-424.25		424.25	
Adjusted ρ^2	0.08518		0.1223		0.1162	
Sample Size N	63		63		63	
Choice-set Size J	11,501		1,000		1,000	
Estimation Time (in seconds)	363.0		55.27		220.8	

Nest Amadora and Odivelas. Root Lisbon municipality. Models include sampling correction. Models corrected for endogeneity with 2SCF. Sample 500 alts. from Odivelas-Amadora nest and 500 from Lisbon municipality. €/M: Euros per month.

The results of the model estimated using this sampling protocol, are shown in Table 6. In the third column are reported the estimators of the *Unexpanded* model where the sampling correction was applied but the sum of the exponentials of the Odivelas-Amadora nest was calculated using only the 5 alternatives sampled from the nest. Note that several estimators are statistically different (with 95% confidence) from those of the original model. Remarkably, the estimator of the scale of the Odivelas-Amadora nest is highly positively biased. This means that the use of the *Unexpanded* model for simulation would cause an important overestimation of the substitution among dwellings in the Odivelas-Amadora nest.

The fourth column of Table 5 reports the estimators of the model estimated using the *Expanded Iterative Prob.* method, where the sampling correction is applied and the sum of the exponentials is expanded using the iterative procedure described in Section 6.3. Equivalent to what occurred in the Monte Carlo experiments, the estimators are remarkably similar to those of the model without sampling and statistically equal (with 95% confidence) to them.

The second experiment corresponded to the application of the same sampling protocol as before, but with alternatives that were sampled up to make a total of 500 for the Odivelas-Amadora nest and 500 for the root (Lisbon). The results of the models estimated using this sampling protocol are shown in Table 7. Equivalent to what occurred with the Monte Carlo experiments, the estimators of the *Unexpanded* and of the *Expanded Iterative Prob.* models are similar to those of the model without sampling when $\tilde{J}$ is large. All estimators are statistically equal (with 95% confidence) in both cases. The most significant difference is that the bias of the estimator of the scale of the Odivelas-Amadora's nest is smaller for the *Expanded Iterative Prob.* model.

Finally, Table 6 and Table 7 report also the computational time used in the estimation of the different models. In the case where only 10 alternatives were sampled, the differences in computational costs were huge. The true model that considers the full choice-set of 11,501 alternatives took approximately 350 times more to be estimated than the *Unexpanded* model, and approximately 35 times more than the *Expanded Iterative Prob.* method. The differences are reduced to 7 and 1.7 times respectively, when 1,000 alternatives are sampled. These differences in estimation time, together with the evidence gathered from the Monte Carlo experiment with one million alternatives, reflect the significant gains that can be obtained with sampling. The methodological developments of this paper will allow taking benefit of these gains in the implementation of spatial choice models with more realistic error structures rendering the development of better tools for policy analysis.

8 Conclusion

Sampling of alternatives for non-Logit models is a problem that has been open for over 30 years, and that have hindered the development of suitable spatial choice models. This paper proposes a novel method to address this issue for MEV models and illustrates its properties by means of a Monte Carlo experiment applied to the Nested Logit model, and a case study based on real data on residential location choice from Lisbon, Portugal.

Monte Carlo experiments showed that the sampling of alternatives causes a significant bias in the estimators of the model parameters when the choice model is Nested Logit. In addition, the proposed method for expanding the sum of the exponentials performed well, even for small sample sizes. In cases where it is possible to obtain an additional sample to expand the sum of the exponentials, the method proposed is easily applicable. When it is not possible to re-sample, the method requires knowledge of the choice probabilities in order to build the expansion factors. In this final case, an iterative procedure showed satisfactory results.

Monte Carlo experiments additionally offered evidence that the sample size required to obtain good estimators while sampling alternatives in MEV models depends on the distribution of the data available and cannot be expressed as a percentage of the cardinality of the true choice-set. In general, an appropriate strategy to determine if the size of the sample of alternatives is large enough might be to test the stability of the estimators with different number of alternatives sampled.

The application with real data demonstrated that the proposed method to achieved consistency while sampling of alternatives in MEV is practical and may have a

significant impact in the design of more sophisticated modeling tools for policy analysis in urban systems.

Different lines for future research may be proposed to address the limitations of this study. First, it would be interesting to apply the method developed in this paper into larger real databases and other spatial choice models, such as job and firm location, route choice or activity scheduling. Finally, it would be interesting to assess the full impact of the methodological advances of this research in policy analysis by applying them in the framework of an operational microscopic integrated urban model such as UrbanSim (Waddell *et al.*, 2008).

Acknowledgments

This publication was made possible in part by the generous support of the Portuguese Government through the Portuguese Foundation for International Cooperation in Science, Technology and Higher Education, undertaken by the MIT-Portugal Program. Additional funding for this research came from the Martin Family Society of Fellows for Sustainability. We would also like to thank Luis Martinez for the provision, processing and analysis of the real data. All Monte Carlo and real data experiments were generated and estimated using the open-source software R (R Development Core Team, 2008).

References

Ben-Akiva M. 1973. Structure of Passenger Travel Demand Models. Ph.D. Thesis, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA.

Ben-Akiva M, Lerman S. 1985. *Discrete Choice Analysis, Theory and Application to Travel Demand.* MIT Press: Cambridge, MA.

Berkovec J, Rust J. 1985. A Nested Logit Model of Automobile Holdings for One Vehicle Households. *Transportation Research B* **19**: 275-285.

Berndt E, Hall H, Hall R, Hausman J. 1974. Estimation and Inference in Nonlinear Structural Models. *Annals of Economic and Social Measurement* **3/4**: 653-665.

Bertsekas D, Tsitsiklis, J. 2002. *Introduction to Probability.* Athena Scientific Press: Belmont, MA.

Bierlaire M. 2003. BIOGEME: A Free Package for the Estimation of Discrete Choice Models. *Proceedings of the 3rd Swiss Transportation Research Conference.* Ascona, Switzerland.

Bierlaire M. 2001. A Theoretical Analysis of the Cross-Nested Logit Model. *Annals of Operations Research* **144**(1): 287-300.

Bierlaire M, Bolduc D, McFadden, D. 2008. The Estimation of Generalized Extreme Value Models from Choice-Based Samples. *Transportation Research Part B: Methodological* **42**(4):381-394.

Chen Y, Duann L, Hu W. 2005. The Estimation of Discrete Choice Models with Large Choice-Set. *Journal of the Eastern Asia Society for Transportation Studies* **6**: 1724-1739.

Domanski A. 2009. Estimating Mixed Logit Recreation Demand Models with Large Choice Sets. Presented at the Agricultural and Applied Economics Association Annual Meeting, Milwaukee, WI.

Frejinger E, Bierlaire, M, Ben-Akiva, M. 2009. Sampling of Alternatives for Route Choice Modeling. *Transportation Research Part B: Methodological* **43**(10): 984-994.

Garrow L, Koppelman, F, Nelson L. 2005. Efficient Estimation of Nested Logit Models using Choice-Based Samples. In *Transportation and Traffic Theory Flow, Dynamics and Interactions: Proceedings of the 16th International Symposium on Transportation and Traffic Theory*, Mahmassani (ed) Oxford, UK: 525-544.

Guevara C. (2010). Endogeneity and Sampling of Alternatives in Spatial Choice Models. Ph.D. Thesis, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA.

Hansen E. 1987. Industrial Location Choice in Sao Pablo Brazil, A Nested Logit Model. *Regional Science and Urban Economics* **17**: 89-108.

Hausman J. 1978. Specification Tests in Econometrics. *Econometrica* **46**: 1251-1272.

Heckman J. 1978. Dummy Endogenous Variables in a Simultaneous Equation System. Econometrica **46**: 931-959.

Imbens G, Lancaster L. 1994. Combining Micro and Macro Data in Microeconometric Models. *Review of Economic Studies* **61**(4): 655–80.

Lee B, Waddell P. 2010. Residential Mobility and Location Choice: a Nested Logit Model with Sampling of Alternatives. *Transportation* **37**(4): 587-601.

Manski C, Lerman S. 1977. The Estimation of Choice Probabilities from Choice Based Samples. *Econometrica* **45**(8): 1977-1988.

Manski C, McFadden D. 1981. Alternative Estimators and Sample Designs for Discrete Choice Analysis. In *Structural Analysis of Discrete Data with Econometric Applications*, Manski and McFadden (eds). MIT Press, Cambridge, MA, 2-50.

Martinez L, Viegas, J. 2009. Effects of Transportation Accessibility on Residential Property Values: A Hedonic Price Model in the Lisbon Metropolitan Area. *Transportation Research Record* **2115**: 127-137.

Martinez L, Abreu J, Viegas J. 2010. Assessment of Residential Location Satisfaction in Lisbon Metropolitan Area. Presented at the 89th Transportation Research Board Annual Meeting, Washington, DC.

McConnel K, Tseng, W. 2000. Some Preliminary Evidence on Sampling of Alternatives with the Random Parameters Logit. *Marine Resource Economics* **14**: 317-332.

McFadden D. 1978. Modeling the Choice of Residential Location. In *Spatial Interaction Theory and Residential Location*, Karlquist, Lundqvist, Snickers and Weibull (eds). North Holland, Amsterdam, 75-96.

Nerella S, Bhat C. 2004. A Numerical Analysis of the Effect of Sampling of Alternatives in Discrete Choice Models. *Transportation Research Record* **1894**: 11-19.

Newey W, McFadden, D. 1986. Large Sample Estimation and Hypothesis Testing. In *Handbook of Econometrics* **4**(36): 2111-2245, Engle and McFadden (eds).

Parsons G, Kealy M. 1992. Randomly Drawn Opportunity Sets in a Random Utility Model of Lake Recreation. *Land Economics* **68**(1): 93-106.

Papola A. 2004. Some Developments on the Cross-nested Logit Model. Transportation Research Part B **38**: 833–851.

R Development Core Team. 2008. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Rivera M, Tiglao, N. 2005 Modeling Residential Location Choice, Workplace Location Choice and Mode Choice of Two-Worker Households in Metro Manila. *Proceedings of the Eastern Asia Society for Transportation Studies* **5**: 1167 – 1178.

Small K. 1987. A Discrete Choice Model for Ordered Alternatives. *Econometrica* **55**(2): 409-424.

Sermons M, Koppelman, F. 2001. Representing the Differences between Female and Male Commute Behavior in Residential Location Choice Models. *Geography* **9**: 101–110.

Train K. 2009. *Discrete Choice Methods with Simulation, 2nd Edition*. Cambridge University Press: New York, NY.

Train K, McFadden D, Ben-Akiva M. 1987. The Demand for Local Telephone Service: A fully Discrete Model of Residential Calling Patterns and Service Choice. *Rand Journal of Economics* **18**: 109–123.

Vovsha P. 1999. Comparative Analysis of Different Spatial Interaction Models. Presented at the 78th Annual Meeting of the Transportation Research Board, Washington, DC.

Vovsha P, Bekhor S. 1998. The Link-Nested Logit Model of Route Choice: Overcoming the Route Overlapping Problem. *Transportation Research Record*: 1645, 133–142.

Waddell P, Wang L, Liu X. 2008. UrbanSim: An Evolving Planning Support System for Evolving Communities. In *Planning Support Systems for Cities and Regions*, Brail (eds): 103-138. Lincoln Institute for Land Policy: Cambridge, MA.

Walker J. 2001. Extended Discrete Choice Models: Integrated Framework, Flexible Error Structures, and Latent Variables. Ph.D. Thesis, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA.

Watanatada T, Ben-Akiva M. 1979. Forecasting Urban Travel Demand for Quick Policy Analysis with Disaggregate Choice Models: a Monte Carlo Simulation Approach. *Transportation Research* **13**(A): 241–248.

White H. 1982. Maximum Likelihood Estimation of Misspecified Models. *Econometrica* **50**: 1–25.