Estimation techniques for MEV models with sampling of alternatives

Ricardo Hurtubia, Gunnar Flötteröd and Michel Bierlaire
Transport and Mobility Laboratory (TRANSP-OR), EPFL

Abstract

Estimation of MEV models with large choice sets requires sampling of alternatives, which might be a difficult task due to the correlated-structure of the error terms. Standard sampling techniques like the ones traditionally used for Multinomial Logit models can not be directly applied in the estimation of more complex MEV models. State of the art estimators for MEV models with sampling of alternatives either require knowledge of the full choice set or produce biased estimates for small sample sizes. This paper proposes two estimation techniques for MEV models with sampling of alternatives. The first technique is based on bootstrapping and allows to reduce the bias for existing estimators. The second technique introduces a new estimator, based on importance sampling, which generates unbiased parameter estimates for small sample sizes.

1 Introduction

In discrete choice models, sampling of alternatives is commonly used when the choice set is large. Typical examples of this are the problems of residential location choice, destination choice or route choice, where the identification of each available alternative becomes difficult.

Sampling a subset of alternatives from the full choice set allows for a simpler estimation of the parameters in the utility function by reducing the computational complexity of the estimator. In the case of a Multinomial Logit model (MNL), where the error terms of the random utilities are independent and identically distributed (*iid*), it is possible to estimate parameters that are consistent and unbiased by adding a corrective constant to the utility of each alternative (McFadden, 1978). However, if the *iid* assumption is discarded, the sampling correction method usually utilized in MNL models will generate biased estimates. This is the case for the Nested Logit model

(NL), the Cross Nested Logit (CNL) and other members of the Multivariate Extreme Value (MEV) family of models.

The estimation of nested-structured MEV models with sampling of alternatives is difficult because of the error-correlation structures that makes the probability of choosing the sampled alternative dependent on the utilities of all the alternatives in the corresponding nest. For example, in a Nested Logit model, the inclusive value (or logsum) will include the full choice set in the nest, regardless of the selected alternatives in the sample for the choice probabilities. Bierlaire et al. (2008) propose an estimator for MEV models with sampling of alternatives. However it assumes that the probability generating function of the MEV can be computed accurately, which again requires the full choice set. They do not describe how to apply sampling to compute it.

In the context of route choice models, Frejinger et al. (2009) introduce the concept of "Extended Path Size", where the Path Size is computed based on a sample of alternatives, and corrected using an expansion factor. Guevara and Ben-Akiva (2010) generalize this approach and derive an expansion factor for a general MEV model. This method generates asymptotically unbiased estimates of the unknown parameters; however, the quality of the estimates depends on the sample size, generating biased results for relatively small samples.

This paper proposes two improved estimators for MEV models with sampling of alternatives. We take as a starting point (and benchmark) the method proposed by Guevara and Ben-Akiva and develop two estimation procedures that reduce the bias of the estimates. The first procedure utilizes bootstrapping techniques to reduce the bias of the estimates generated by the benchmark method while the second proposes a new approximation of the logsum and an importance sampling strategy to reduce the bias in the estimates of the parameters. Both procedures are tested over synthetic data using Monte Carlo experiments; results are compared with those obtained when using the method proposed by Guevara and Ben-Akiva.

The paper is organized as follows: Section 2 reviews the sampling methodology for MEV models and the method proposed by Guevara and Ben-Akiva. Section 3 introduces two techniques for bias reduction in the estimation of MEV model under sampling of alternatives: importance sampling and bootstrapping. Section 4 describes a Monte Carlo experiment using the bootstrapping approach. Section 5 shows the results of an experiment using the importance sampling approach. Finally, Section 6 concludes the paper and identifies further research.

2 Sampling of alternatives in random utility models

In the following section we analyze methods for sampling of alternatives in MEV models. We start with the simple case of the Multinomial Logit to then describe the more general method for sampling of alternatives in MEV models. The section concludes describing the state of the art for sampling of alternatives in Nested Logit models.

2.1 Multinomial Logit

In a MNL model, the probability of decision-maker n choosing alternative i is given by :

$$P(i) = \frac{e^{V_{ni}}}{\sum_{j \in C_n} e^{V_{nj}}} \tag{1}$$

where $V_{ni} = V(x_{in}, \beta)$, the systematic part of the utility of alternative i for decision maker n, is function of the alternative's attributes (x_{in}) and a vector of unknown parameters (β). For notation simplicity the scale parameter μ is omitted. The term C_n represents the full set of available alternatives from where the decision-maker can choose.

If C_n is large, the analyst might want to sample a smaller subset of alternatives D_n. The probability of constructing the subset D_n given that alternative i was chosen is denoted by $\pi(D_n|i)$. Following McFadden (1978) the probability of choosing alternative i given a subset D_n is:

$$P(i|D_n) = \frac{e^{\mu V_{ni} + \ln \pi(D_n|i)}}{\sum_{j \in D_n} e^{\mu V_{nj} + \ln \pi(D_n|j)}} \tag{2}$$

where the term $\ln \pi(D_n|i)$ works like an alternative-specific expansion factor. The positive conditioning property (McFadden, 1978) ensures that, if the probabilities $\pi(D_n|j)$ are positive and known for all alternatives $j \in D_n$, consistent estimates of the parameters β can be obtained through maximum log-likelihood estimation, following:

$$\max_{\beta} \sum_{n} \ln P(i|D_n)^{y_{in}} \tag{3}$$

where y_{in} assumes the value of one if n chose alternative i and zero otherwise.

The unbiased estimates resulting from solving (3) are possible thanks to the *iid* structure of the error terms in the MNL model. More complex models, allowing for correlation between alternatives (like the NL or CNL models), do not hold this property and therefore correction for sampling can not be achieved by just adding an alternative specific correction, as done in (2). This issue is reviewed in the next subsection.

2.2 MEV Models

Many random utility models (such as the MNL, NL and CNL models) can be expressed as particular cases of the (more general) Multivariate Extreme Value family of models (McFadden, 1978). The error-correlation structure in MEV models (also named Generalized Extreme Value models) is defined through the generating function $G(e^{V_1}, ..., e^{V_J})$, such that the probability of choosing alternative i is:

$$P_n(i) = \frac{e^{V_{in}} G_i}{G(e^{V_1}, ..., e^{V_J})} \tag{4}$$

where

$$G_i = \frac{\partial G(e^{V_{1n}}, e^{V_{2n}}, ..., e^{V_{J_n n}})}{\partial e^{V_{in}}} \tag{5}$$

The choice probability of (4) can be re-written in the form of a multinomial logit, but keeping the error-correlation structure defined by G (Ben-Akiva and Lerman, 1985).

$$P_n(i) = \frac{e^{V_{in} + \ln G_i}}{\sum_{j \in C_n} e^{V_{jn} + \ln G_j}} \tag{6}$$

Different functional forms for G generate different models. For example a MNL model is obtained if $G(y) = \sum_{j \in C_n} y_j^{\mu}$

Taking advantage of the closed form of equation (6), Bierlaire et al. (2008) proposed an estimator over a sample of alternatives for MEV models. The choice probabilities in this case are similar to those described by (2):

$$P_n(i|D_n) = \frac{e^{V_{in} + \ln G_i + \ln \pi(D_n|i)}}{\sum_{j \in D_n} e^{V_{jn} + \ln G_j + \ln \pi(D_n|j)}} \tag{7}$$

It is important to notice that, unlike the case of (2), the choice probabilities of (7) do not depend only on the utilities of the alternatives in subset D_n. This is caused by the term G_i which, with few exceptions like the MNL, depends on the utilities of the alternatives in the full choice set. Therefore, under sampling of alternatives, equation (7) can not be used for a consistent maximum likelihood estimation of the unknown parameters.

Consistent estimation of the unknown parameters requires an unbiased estimator of the derivative of the generation function (G_i). The feasibility and complexity of this estimator will depend on the functional form of G. In this paper we analyze the case of the MEV formulation for a Nested Logit model and the estimator originally proposed by Guevara and Ben-Akiva.

2.3 Sampling correction for the Nested Logit model

The MEV formulation of a Nested Logit model with M nests considers a generating function with the following functional form:

$$G = \sum_{m=1}^{M} \left(\sum_{i \in C_{mn}} e^{\mu_m V_{in}} \right)^{\frac{\mu}{\mu_m}} \tag{8}$$

where μ_m is the scale parameter for nest m and μ is the scale parameter for the higher level nest. C_{mn} is the full set of alternatives in nest m.

The logarithm of the first order derivative of (8) is:

$$\ln G_{in} = \left(\frac{\mu}{\mu_{m(i)}} - 1\right)\left(\ln \sum_{j \in C_{m(i)n}} e^{\mu_{m(i)} V_{jn}}\right) + \ln \mu + (\mu_{m(i)} - 1) V_{in} \tag{9}$$

where $m(i)$ is the nest containing alternative i.

Since the logsum depends on all the alternatives in the nest it needs to be approximated if the probability is to be calculated over a sub-sample D_{mn}. Guevara and Ben-Akiva (2010) proposed the following estimator of the logsum for nest $m(i)$:

$$\left(\ln \sum_{j \in C_{m(i)n}} e^{\mu_{m(i)} V_{jn}}\right) \approx \left(\ln \sum_{j \in D_{m(i)n}} w_{jn} e^{\mu_{m(i)} V_{jn}}\right) \tag{10}$$

where $D_{m(i)n}$ is a sub-sample of the alternatives in nest $m(i)$. The weights (w_{jn}) are calculated as follows:

$$w_{jn} = \frac{\widetilde{n}_{jn}}{E_n(j)} \tag{11}$$

where $\widetilde{n}_{jn}$ is the number of times alternative j was sampled and $E_n(j)$ is the probability for alternative j to be included in the sample, according to the sampling protocol. In their analysis, Guevara and Ben-Akiva use a sampling without replacement protocol for the sampling of alternatives, therefore making $\widetilde{n}_{jn}$ at most equal to one and $E_n(j)$ equal to the probability of drawing alternative j. They also propose to use the same sample for the elements in the logsum and the alternatives in the choice set, this means that the chosen alternative is always included in $D_{m(i)n}$.

Their approximation generates asymptotically unbiased estimates of the utility parameters and the scale parameters for each nest. However, for relatively small sample sizes, the approximated logsum is unable to reproduce the full logsum values and, therefore, the estimation results are biased.

The best results are obtained when $E_n(j)$ is calculated using the true choice probabilities; this implies the un-realistic assumption of the analyst being able to observe the probabilities before estimation. Other more realistic estimators of the true choice probabilities were tested, generating more biased coefficient estimates with the exception of an iterative estimator which generated results statistically equal to those obtained with the true probabilities.

In the following section we propose techniques to reduce the bias in the estimation results for MEV models with sampling of alternatives.

3 Techniques for bias-reduction

Two statistical techniques are proposed for an improved estimation of the logsum (10): importance sampling and bootstrapping.

3.1 Importance Sampling

Importance sampling allows to reduce the variance of an average that approximates an expectation. For this, (i) a *proposal distribution* needs to be chosen that defines how sampling has to take place and (ii) the average needs to be corrected for this sampling strategy. Essentially, a proposal distribution that favors large values (in absolute terms) is more likely to draw elements into the average that substantially contribute to the approximated expectation, and hence it leads to an improved reduction in variance.

The bias in the estimator (10) is monotonously increasing with the variance of the argument of the logarithm: For a zero variance, there is no bias at all. The larger the variance gets, the more the nonlinear form of the logarithm takes effect in distorting the distribution of its argument. It hence is desirable to estimate this argument with a low variance. In the following, we apply importance sampling for this purpose.

Defining for notational simplicity

$$z_{in} = \mu_{m(i)} V_{in} \tag{12}$$

and omitting the index n as from now, (10) becomes

$$\ln \sum_{j \in C_{m(i)}} e^{z_j}. \tag{13}$$

This expression can be rephrased as the logarithm of an expectation:

$$\ln \sum_{j \in C_{m(i)}} \frac{e^{z_j}}{g(j)} g(j) = \ln \left(\mathrm{E} \left\{ \frac{e^{z_j}}{g(j)} \middle| j \sim g(j) \right\} \right). \tag{14}$$

In order to compute the argument of the logsum, a proposal distribution $g(j)$ needs to be defined that is strictly positive for all $j \in C_{m(i)n}$. Based on a set of R independent and identically distributed samples generated from this distribution, (14) is then approximated by

$$\ln \sum_{j \in C_{m(i)}} e^{z_j} \approx \ln \left(\frac{1}{R} \sum_{r=1}^{R} \frac{e^{z_{j(r)}}}{g(j(r))} \right), \quad j(r) \sim g(j),\ r = 1 \ldots R. \tag{15}$$

A concrete version of the proposal distribution $g(j)$ will be described in Section 5.2 (equation 27).

3.2 Bootstrapping

Bootstrap methods were first proposed by Efron (1979) as simulation-based techniques for statistical inference. Bootstrapping is generally used to infer the properties of an estimator from a limited sub-sample of observations; this opens the possibility of measuring the bias of an estimator and correcting for it.

The bootstrapping technique approximates a given distribution by a limited set of samples and makes further inference about this distribution by re-sampling from this set of samples. Let θ_g be some statistic of $x \sim g$. The statistic is estimated from a set of R samples $x(r) \sim g$, $r = 1 \ldots R$, using the estimator

$$\hat{\theta}(x(1), \ldots, x(R)). \tag{16}$$

The bias of this estimator, which is

$$\mathrm{E}\{\hat{\theta}(x(1), \ldots, x(R)) | x(r) \sim g,\ r = 1 \ldots R\} - \theta_g, \tag{17}$$

can be estimated using the bootstrap estimator

$$\frac{1}{B} \sum_{b=1}^{B} \hat{\theta}(x(1, b), \ldots, x(R, b)) - \hat{\theta}(x(1), \ldots, x(R)). \tag{18}$$

where $x(1,b),\dots,x(R,b)$ is a set of independent and uniform re-samples from the original sample $x(1),\dots,x(R)$. Subtracting this bias from the original estimator $\hat{\theta}(x(1),\dots,x(R))$ results in the corrected estimator

$$2\hat{\theta}(x(1),\dots,x(R)) - \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}(x(1,m),\dots,x(R,m)). \tag{19}$$

In our case, the statistic under consideration is

$$\theta_g = \ln\left(E\left\{\frac{e^{z_j}}{g(j)}\middle| j \sim g(j)\right\}\right), \tag{20}$$

and the respective estimator is

$$\hat{\theta}(i(1),\dots,i(R)) = \ln\left(\frac{1}{R}\sum_{r=1}^{R}\frac{e^{z_{j(r)}}}{g(j(r))}\right), \quad j(r) \sim g(j),\ r = 1\dots R. \tag{21}$$

An improved version of (15) can therefore be obtained by application of the corrected estimator (19) using the definitions (20) and (21).

4 Experiment: bias-correction with bootstrapping

We perform a Monte Carlo simulation experiment similar to the one presented in Guevara and Ben-Akiva (2010). This consists in a nested logit model with 2 nests; the first containing 5 alternatives and the second containing 1000 alternatives. All the alternatives of the first nest are included in the estimation while a importance sampling protocol is applied in the second nest: for each observation the chosen alternative is included and an additional set of non-chosen alternatives is sampled without replacement from the full choice set.

The utilities are linear-in-parameters and depend on 2 variables, a and b, randomly generated from a uniform (-1,1) distribution. The values of the true parameters are set to $\beta_a = 1$, $\beta_b = 1$, $\mu_1 = 2$, $\mu_2 = 3$. Choices are simulated for 1000 observations using the true parameters and the complete choice set for both nests, following the probability distribution defined in (6).

Two experiments are performed. The first applies the approximated logsum proposed by Guevara and Ben-Akiva in order to have benchmark results for comparison purposes. The second experiments uses the same approximation for the logsum, but includes a bootstrap correction which is used in a second estimation instance.

4.1 Approximated logsum

First we attempt to reproduce the original results by using the approximation defined in (10). Also, to compare with the best possible results, we use the "true probabilities" approach for $E(j)$ as a benchmark. For this we use the following choice probability in the estimation:

$$P_n(i|D_n) = \frac{e^{V_{in} + \ln G_i'(D_{m(i)n}) + \ln \frac{|C_{m(i)}|}{|D_{m(i)n}|}}}{\sum_{j \in D_n} e^{V_{jn} + \ln G_j'(D_{m(j)n}) + \ln \frac{|C_{m(j)}|}{|D_{m(j)n}|}}} \tag{22}$$

where $G_i'(D_n)$ is the derivative of the generating function (9) but replacing the full logsum for the approximation defined in (10). The sampling correction is calculated as the number of alternatives in the full choice set ($|C_{m(i)}|$) over the sample size ($|D_{m(i)}|$). Since nest 1 considers all the available alternatives ($C_{1(i)} = D_{1(i)}, \forall i$) this correction is only applied in nest 2 where $C_{2(i)} \supseteq D_{2(i)}, \forall i$.

Table 1: Estimation results - Approximated logsum

parameter	average value	std-error	true value	t-test
β_a	0.831	0.052	1	3.226 *
β_b	0.848	0.054	1	2.788 *
μ_1	2.982	0.419	2	2.339 *
μ_2	3.646	0.189	3	3.428 *

* coefficients statistically different from the true parameters

Table 1 shows the results for the Monte Carlo experiment, using a sample of 10 alternatives for nest 2 (sample size = 1% of $|C_2|$) and estimating with the probabilities defined by (22). As expected, given the relatively small sample size, all the estimates are significantly different from the true parameters. This is shown by the t-test against the true values (a t-test bigger than 1.96 indicates a 95% probability of the estimate being different from the true parameter).

4.2 Bootstrapping

We repeat the experiment, but implementing a sequential estimation procedure: the first iteration considers a regular estimation using the approximated logsum. The second iteration repeats the estimation but incorporates the bootstrap correction (ρ_n) in the logsum. This correction is calculated with the parameters obtained from the first estimation (β^*, μ^*), following the method described in Section 3.2

$$\rho_n = \frac{1}{B}\sum_{b}^{B}\left(\ln \sum_{j\in D_{mn}^b} w_{jn} e^{\mu_m^* V_{jn}(\beta^*)}\right) - \left(\ln \sum_{j\in D_{mn}} w_{jn} e^{\mu_m^* V_{jn}(\beta^*)}\right) \tag{23}$$

where B is the number of re-sampling instances of the bootstrap estimator and D_{mn}^b defines the alternatives in the sample at each instance.

The parameters are re-estimated using the following choice probability:

$$P_n(i|D_n) = \frac{e^{V_{in}+\ln G_i'(D_{m(i)n})-\rho_n+\ln\frac{|C_{m(i)}|}{|D_{m(i)n}|}}}{\sum_{j\in D_n} e^{V_{jn}+\ln G_j'(D_{m(j)n})-\rho_n+\ln\frac{|C_{m(j)}|}{|D_{m(j)n}|}}} \tag{24}$$

The Monte Carlo experiment is performed with the same sample size used in Section 4.1 ($|D_2| = 10$); results are shown in Table 2. In this case the parameters were re-estimated after calculating the bootstrap correction for the bias, following the probability distribution defined in (24). The estimated parameters are closer and statistically equal to the true values; therefore the bias has been reduced with respect to the original estimation.

The results confirm the usefulness of the bootstrapping correction in the estimation of MEV models. However, the quality of the bootstrapped parameters depends on the quality (in terms of bias) of the first-instance

Table 2: Estimation results - Bootstrap correction

parameter	average value	std-error	true value	t-test
β_a	0.949	0.099	1	0.518
β_b	0.936	0.095	1	0.672
μ_1	2.505	0.732	2	0.690
μ_2	3.232	0.285	3	0.811

estimates. In the results shown in Table 2 the parameters used for the calculation of the bootstrap correction are those obtained in the first instance with the approximated logsum which, in average, have the values shown in Table 1. The unbiased estimates obtained after bootstrapping are only possible thanks to the relatively good original estimates, where the bias exists but is not extreme.

Therefore bootstrapping is an appropriate tool to reduce the bias of the estimates, but it necessarily requires a good initial approximation of the logsum. To illustrate this, Figures 1 and 2 show the evolution of the parameter values in two particular realizations of the Monte Carlo experiment. The realizations were selected as examples of a "good" and a "bad" starting point (which depends on the random sampling of alternatives). In both cases, the iterative estimation process converges very quickly to a stable result. In the case of a good starting point (Figure 1) the values are slightly shifted, but enough to reduce the bias significantly. In the case of the bad starting point (Figure 2), the bootstrapping and re-estimation technique is unable to move the values close enough to the true parameters.

Figure 1: Estimation iterations (good initial point)

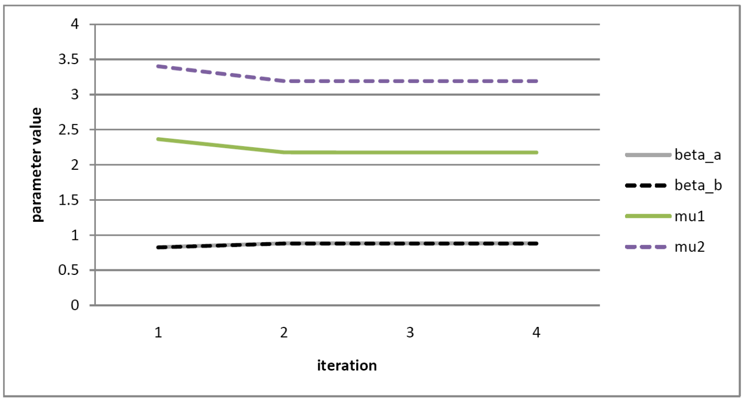
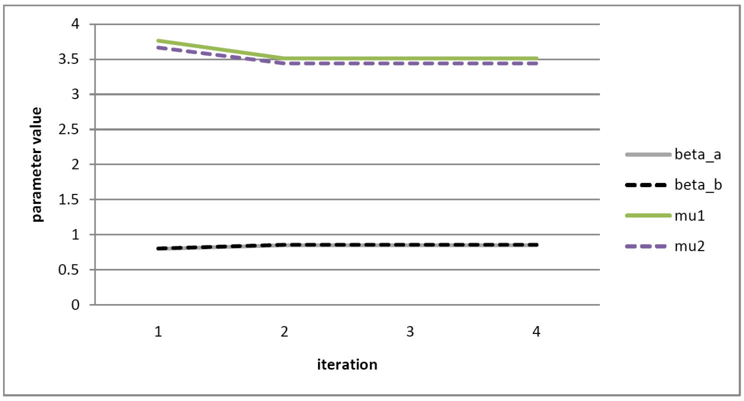


Figure 2: Estimation iterations (bad initial point)



These results indicate that bootstrapping alone will not solve the bias problems of an estimator: a good estimator is required beforehand. The next section proposes a method to obtain unbiased estimators for MEV models under sampling of alternatives.

5 Experiment: importance sampling for logsum estimation

The relevance of the initial estimates for the bootstrapping procedure motivates the search of a new strategy for sampling of alternatives in MEV models. As explained in Section 3.1, the quality of the the estimator for the logsum will depend on the sampling protocol. A new experiment is performed and the approximated logsum approach is compared with a new methodology described in Section 5.2

The experiment is performed with synthetic data built over a real dataset from a stated preferences survey to evaluate a high-speed train in Switzerland (Bierlaire et al., 2001). The original dataset considers three possible alternatives: Car (C), Train (T) and High-speed Train (HS). We estimate a model over the original dataset in order to have proper true values for the parameters in the experiment. The model assumes two nests: an "innovative" nest including only the high-speed train and a "traditional" nest including both car and train. Utilities are linear in parameters with some alternative-specific parameters. Results for the estimation over the original

Table 3: Estimation results over original dataset

parameter	affected V	value	std-error	t-test
β_{cost}	C-T-HS	-0.849	0.122	-6.96
β_{time_C}	C	-1.760	0.148	-11.84
β_{time_T}	T-HS	-1.840	0.173	-10.65
$\beta_{headway}$	T-HS	-0.496	0.227	-2.19
μ_1(innovative)	-	1*	-	-
μ_2(traditional)	-	1.55	0.201	2.76**

* fixed parameter

** t-test against 1

data are presented in Table 3.

Synthetic data is generated by generating new alternatives based on the original ones, introducing a multiplicative disturbance in the attributes with a uniform distribution (0.5,1.5). We generate 4 new High-speed-based alternatives, 49 new car-based alternatives and 49 new train-based alternatives. Therefore, in our synthetic data, the innovative nest has 5 alternatives and the traditional nest has 100 alternatives. Since the number of alternatives in each nest is different from the original problem we arbitrarily define new true values for the scale parameters: $\mu_1 = 2$ and $\mu_2 = 4$. Simulation of choices is performed over the synthetic dataset using the true choice probabilities defined by (6).

Two experiments were performed. The first using Guevara and Ben-Akiva's approximation for the logsum and the second using importance sampling for the logsum estimator. In both cases the sample size for nest 2 was of 10 alternatives (10% of $|C_2|$).

5.1 Approximated logsum

The method proposed by Guevara and Ben-Akiva is applied to the synthetic dataset. Estimation is done using the probability described by (22) and alternatives for both the choice set and the logsum are randomly sampled without replacement and including the chosen alternative. Results for this experiment are shown in Table 4.

Table 4: Estimation results: approximated logsum

parameter	average value	std-error	true value	t-test
β_{cost}	-1.033	0.149	-0.849	1.237
β_{time_C}	-2.382	0.302	-1.760	2.055 *
β_{time_T}	-2.264	0.295	-1.840	1.439
$\beta_{headway}$	-0.742	0.119	-0.496	2.069 *
μ_1(innovative)	1.507	0.269	2	1.838
μ_2(traditional)	3.431	0.294	4	1.938

* coefficients statistically different from the true parameters

Results show two parameters that are biased with respect to the true ones: β_{time_C} and $\beta_{headway}$. The scale parameters, although not statistically biased, have values which are far from the true ones and t-tests that are close to 1.96.

5.2 Importance sampling for the logsum estimator

The approximated logsum described in Section 2.3 (equation 10) and used in the previous experiment utilizes the same alternatives that were sampled for the choice set: the chosen alternative and a set of alternatives that are randomly sampled, without replacement, from the full choice set. However meaningful for the alternatives in the choice probability, this sampling procedure is not the best for the estimation of the logsum.

The bias of the parameter-estimates will depend on the bias of the estimated logsum. Importance sampling of alternatives should generate a better estimate than random sampling, as explained in Section 3.1.

We propose a sequential estimation procedure that keeps the sampling protocol for the alternatives in the choice probability, but considers an importance sampling protocol for the alternatives in the approximated logsum.

In the first instance, since the choice probabilities are unknown, the alternatives to be included in the logsum are randomly sampled (with replacement) from the full choice set. The alternatives of the choice set are sampled following the same protocol described in Section 5.1. The parameters are estimated using choice probabilities following:

$$P_n(i|D_n) = \frac{e^{V_{in} + \ln G'_i(L_{m(i)n}) + \ln \frac{|C_{m(i)}|}{|D_{m(i)n}|}}}{\sum_{j \in D_n} e^{V_{jn} + \ln G'_i(L_{m(j)n}) + \ln \frac{|C_{m(j)}|}{|D_{m(j)n}|}}} \tag{25}$$

where $L_{m(i)n}$ is the set of sampled alternatives (from the nest containing i: $m(i)$) for the logsum and $\ln G'_i(L_{m(i)n})$ is the expression described in (9) but using the following approximated logsum:

$$\ln \sum_{j \in L_{m(i)n}} w'_{jn} e^{\mu_{m(i)} V_{jn}} \tag{26}$$

Given the sampling protocol for L_{mn}, the weights (w'_{jn}) are calculated, in the first instance, as the full set size $(|C_{m(i)}|)$ over the sample size $(|D_{m(i)}|)$.

The parameters obtained in the first estimation $(\beta^*, \ \mu^*)$ are used to calculate the importance sampling distribution that will generate the sample for the logsum in the second estimation. The probability of sampling an alternative from a particular nest m is defined as a Multinomial Logit:

$$g_n(i|m) = \frac{e^{V_{ni}(\beta^*, \mu^*)}}{\sum_{j \in C_m} e^{V_{nj}(\beta^*, \mu^*)}} \tag{27}$$

The new sample of alternatives to estimate the logsum (L'_{mn}) is done following $g_n(i|m)$.

A new estimation is performed, similar to the first one but replacing the approximated logsum for:

$$\ln \sum_{j \in L'_{mn}} \frac{1}{|L'_{mn}| \cdot g_n(j|m)} \cdot e^{\mu_{m(j)} V_{jn}} \tag{28}$$

Estimations are repeated until a stable value is achieved for all parameters.

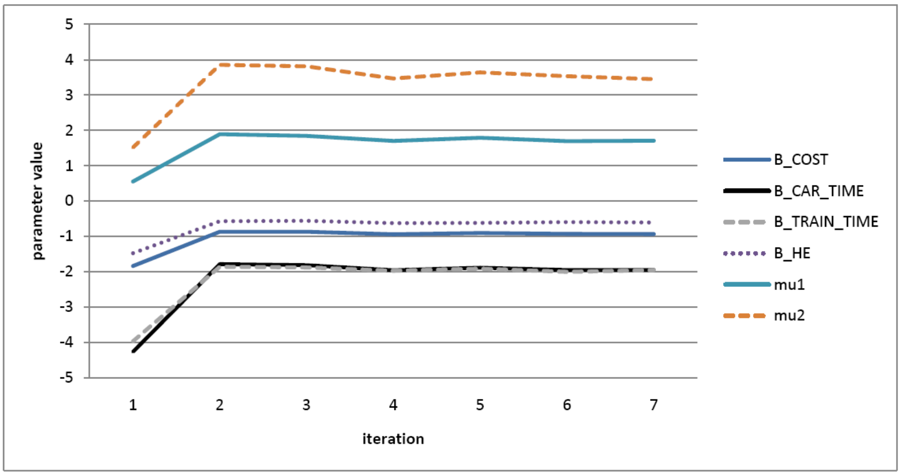
Table 5 shows the results for the proposed methodology. The importance sampling procedure generates unbiased estimates for all the parameters, outperforming the results of the approximated logsum shown in Table 4. Also, the bias of the final estimates does not depend strongly on the bias of

Table 5: Estimation results: importance sampling

parameter	average value	std-error	true value	t-test
β_{cost}	-0.863	0.069	-0.849	0.204
β_{time_C}	-1.805	0.148	-1.760	0.300
β_{time_T}	-1.791	0.160	-1.840	0.309
$\beta_{headway}$	-0.590	0.066	-0.496	1.428
μ_1(innovative)	2.052	0.187	2	0.280
μ_2(traditional)	3.984	0.338	4	0.046

the original estimates. As seen in Figure 3, in a particular realization of the Monte Carlo experiment, the estimated parameters in the first iteration are far from the true values. Despite this, in the following estimations, good estimates are achieved. This implies that the importance sampling method also outperforms the bootstrapping method which, as explained in Section 4.2, requires good initial estimates.

Figure 3: Evolution of the estimates



6 Conclusions and further research

This paper proposed two method to reduce the bias of the parameters when estimating MEV models under sampling of alternatives. The bootstrap method allows to reduce the bias in the parameters of any estimator, but depends on the quality of the estimator itself. The importance sampling method allows to estimate unbiased parameters with a relatively small sample size.

The main contribution of the importance sampling method is a better sampling distribution for the elements in the logsum. The use of a different sample for the choice set and for the logsum allows to reach better results while still having an estimator that is consistent with choice theory.

Both methods reach unbiased estimates with small samples sizes, further research will analyze the effect of different sample sizes in the quality of the estimates. The combination on the two methods proposed on this paper is also part of further research. The bias (if any) of the estimates in the importance sampling method can be reduced by applying bootstrapping.

References

Ben-Akiva, M. E. and Lerman, S. R. (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press, Cambridge, Ma.

Bierlaire, M., Axhausen, K. and Abay, G. (2001). Acceptance of modal innovation: the case of the Swissmetro, *Proceedings of the 1st Swiss Transportation Research Conference*, Ascona, Switzerland. www.strc.ch.

Bierlaire, M., Bolduc, D. and McFadden, D. (2008). The estimation of generalized extreme value models from choice-based samples, *Transportation Research Part B: Methodological* **42**(4): 381–394.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife, *The Annals of Statistics* **7**(1).

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

Guevara, C. and Ben-Akiva, M. E. (2010). Sampling of alternatives in multivariate extreme value (mev) models, *Proceedings of the 12th World Conference on Transport Research*.

McFadden, D. (1978). Modeling the choice of residential location, *in* A. Karlqvist (ed.), *Spatial Interaction Theory and Residential Location*, North-Holland, Amsterdam, pp. 75–96.