Sampling of alternatives in random regret minimization models

Guevara, C. Angelo; Chorus, Caspar G.; Ben-Akiva, Moshe E.

DOI
10.1287/trsc.2014.0573

Publication date
2016

Document Version
Accepted author manuscript

Published in
Transportation Science

Post print version of http://dx.doi.org/10.1287/trsc.2014.0573 (Informs)

SAMPLING OF ALTERNATIVES IN RANDOM REGRET MINIMIZATION MODELS

C. ANGELO GUEVARA (corresponding author)
Faculty of Engineering and Applied Sciences
Universidad de los Andes, Chile
Mons. Álvaro del Portillo 12.455, Las Condes, Santiago, Chile.
Tel: 56-2-2618-1364
Fax: 56-2-2618-1642
caguevara@miuandes.cl

CASPAR G. CHORUS
Faculty of Technology, Policy and Management
Delft University of Technology
Jaffalaan 5, 2628BX, Delft, the Netherlands
E: c.g.chorus@tudelft.nl

MOSHE E. BEN-AKIVA
Department of Civil and Environmental Engineering
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
mba@mit.edu

Key Words: Sampling of Alternatives, Random Regret Minimization.

ABSTRACT

Sampling of alternatives is often required in discrete choice models to reduce the computational burden and to avoid describing a large number of attributes. This approach has been used in many areas, including modeling of route choice, vehicle ownership, trip destination, residential location, and activity scheduling. The need for sampling of alternatives is accentuated for Random Regret Minimization (RRM) models, because, different from Random Utility Models (RUM), the regret function for each alternative depends on all the alternatives in the choice-set. In this paper we develop and test a method to achieve consistency, asymptotic normality and relative efficiency, while sampling alternatives in a class of models that includes RRM. The proposed method can be seen as an extension of the approach used to address sampling of alternatives in Multivariate Extreme Value (MEV) models. We illustrate the methodology using Monte Carlo experimentation and a case study with real data. Experiments show that the proposed method is practical, performs better than an uncorrected model, and results in estimates that are statistically equal to those obtained with a model considering all the alternatives.

1 INTRODUCTION

Various types of discrete choice models that are relevant in transport modeling involve huge choice-sets. This is the case, for example, for models of route choice, trip destination, residential location or activity scheduling. Two types of difficulties may arise when the choice-set is too large. The first is the computational burden of managing a large number of alternatives and the second is the need to gather the data to describe them. Both difficulties may arise in estimation and in forecasting. In this article we consider the former, proposing a solution method for Random Regret Minimization (RRM) models.

In the context of the classical Random Utility Maximization-based (RUM) Logit model (McFadden, 1974), a convenient method has been proposed (McFadden, 1978) to obtain a consistent estimator for model parameters with a sample of alternatives. This estimator capitalizes on the fact that, due to its independently and identically distributed (*iid*) errors, the RUM-based Logit model exhibits the IIA-property. McFadden´s (1978) result concerning the sampling of alternatives for Logit has been profusely used over the years. Examples abound in many areas such as route choice (see. e.g. Fosgerau et al. 2013; Frejinger et al., 2009), vehicle ownership (Berkovec & Rust, 1985), trip destination (Carrasco, 2008), residential location (Lee and Waddell, 2010) and activity based modeling (see. e.g., Daly et al 2013; Bradley et al, 2010; Bowman and Ben-Akiva, 2001).

Although very convenient from a modeler's perspective, this IIA-property is often considered to be restrictive in terms of the implied behavior of decision-makers. Over the past few decades, this observation has led to the development of a number of alternative discrete choice model forms whose errors are not *iid*. While still featuring closed form choice probabilities, these models do not exhibit the IIA property as they allow for correlation among the errors associated with different (subsets of) alternatives. A prominent example of this category is the Nested Logit model (Ben-Akiva, 1973), which was shown a few years after its inception to belong to the more general family of closed form choice models based on a Multivariate Extreme Value (MEV) distribution (McFadden, 1978). More recently, MEV Mixture models have been proposed which allow for even more flexibility in terms of the specification of the error term distribution and related behavioral implications and substitution patterns (e.g., McFadden & Train, 2000).

The problem of sampling of alternatives in non-Logit models has been only recently studied. Guevara and Ben-Akiva (2013a) and Guevara (2010) proposed a method to achieve consistent estimation while sampling alternatives in MEV models, providing examples for the Nested Logit and the Cross Nested Logit. The method consists in developing a proper correction of a term that gets truncated because of the sampling. Also, Guevara and Ben-Akiva (2013b) proposed a method for estimation while sampling alternatives in Logit Mixture models, showing also that a naïve approach, in which the kernel of the mixture is replaced by McFadden´s (1978) correction for Logit, does achieve consistent estimation. With this, Guevara and Ben-Akiva (2013b) provide theoretical support for previous empirical results suggesting the suitability of the naïve

approach for Logit Mixture models (McConnel and Tseng, 2000; Nerella and Bhat (2004); Azaiez, 2010; Lemp and Kockelman, 2012).

Recently, a choice model has been proposed that does not exhibit the IIA-property even though (when written in Logit-form) its errors are *iid*. This Random Regret Minimization (RRM) model (Chorus, 2010), which is the focus of this paper, is based on a regret minimization-based decision rule. The model postulates that when decision makers choose among alternatives, they try to avoid the situation where a non-chosen alternative outperforms a chosen one in terms of one or more attributes. This translates into a regret function for a considered alternative that by definition features all attributes of all competing alternatives. Since its introduction a few years ago, the RRM model has been successfully estimated and applied by various authors in the context of a variety of different choice contexts, involving – to name a few examples – travelers choices among vehicle types, destinations, modes, routes, departure times, and driving maneuvers; politicians' choices among policy options; patients choices among medical treatments; and tourists' choices among leisure activity-locations. An overview of recent studies empriically comparing RRM with RUM models can be found in Chorus et al. (2014).

One disadvantage of the RRM model, which was highlighted in Chorus (2012), is that runtimes may suffer from combinatorial explosion when choice-sets become very large. This issue of course is a direct result from the behavioral postulate, incorporated in the regret function, that every alternative is compared with every other alternative in the choice-set in terms of every attribute.

The combinatorial explosion of RRM, compared to Logit, is illustrated in Figure 1, which depicts estimation time (ordinates axis) as a function of the number of alternatives (J) in the choice-set (abscissas axis). The results were obtained from 10 Monte Carlo simulations for each value of J between 50 and 1000, in steps of 50. The estimation time of each simulation is depicted in grey with a small symbol (a dot for RRM and a triangle for Logit) and the average within the 10 repetitions for each J is depicted with a larger dark symbol. Both the RRM and the Logit models consider only one attribute and 1000 observations. Figure 1 shows that the average estimation time for RRM as a function of J is fitted almost perfectly by a quadratic function, reflecting the computational problems that arise with RRM models with large choice-sets. In turn estimation time for the Logit is almost flat with J.

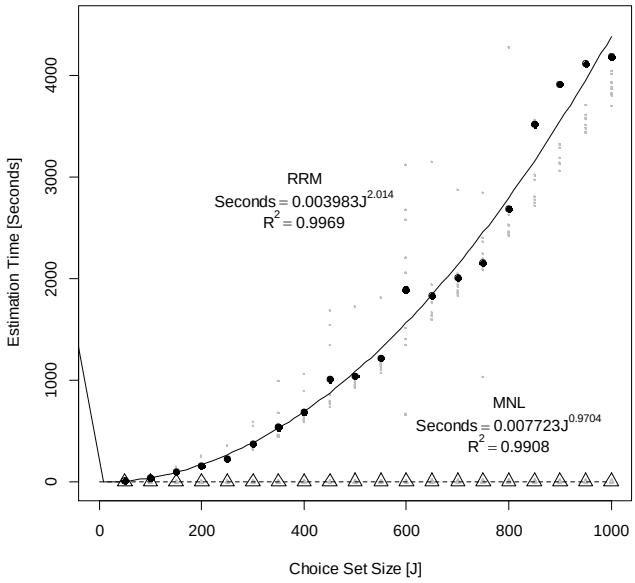


Figure 1: Estimation Time of a RRM and Logit Models as a Function of the Number of Alternatives (*J*)

As a consequence, finding a proper way to estimate RRM models on sampled choice-sets is an important condition for the model to be useful in the context of choice situations involving very large numbers of alternatives. At this point it should be noted that, since RRM model does not exhibit the IIA-property, McFadden's (1978) result does not apply. As mentioned, this is the case even when – such as is the case for RRM-based Logit models – errors are distributed *iid*.

This paper extends the work of Guevara and Ben-Akiva (2013a) by presenting an estimator for the RRM-based Logit model in the context of sampled choice-sets (section 2). Furthermore, it analyzes the conditions required for consistency, asymptotic normality and efficiency and determines the correct expansion factors required in some relevant examples (section 3). Then it illustrates the methods and studies the finite sample properties of the estimators using Monte Carlo experimentation (section 4) and real data (section 5). The article finishes summarizing the main results, their possible implications and suggestions for future lines of research.

2 ESTIMATION AND SAMPLING OF ALTERNATIVES IN RANDOM REGRET MINIZATION MODELS

We consider the RRM model proposed by Chorus (2010). The behavioral assumption behind the RRM model is that individual n chooses alternative i, within the choice-set C_n, if i minimizes the anticipated regret he or she may get from that decision. The regret is defined as a measure of how much worse is the chosen alternative i, regarding each attribute m, compared to all other alternatives $j \neq i$.

For example, if *m* refers to a price attribute, then $\beta_m < 0$. Therefore, if an agent *n* chooses alternative *i* he or she will perceive a price regret $\beta_m \{x_{jmn} - x_{imn}\}$ if $x_{jmn} < x_{imn}$, and zero otherwise, for each alternative *j* other than *i*. Formally, if *m* is the only attribute, the regret function R_{ijmn} can be summarized by the expression shown in Eq. (1).

$$R_{ijmn} = \max\left[0, \beta_m \{x_{jmn} - x_{imn}\}\right] \quad (1)$$

The regret function described in Eq. (1) is difficult to implement in practice for estimation because it is not differentiable. For that, Chorus (2010) proposes to approximate R_{ijmn} by the expression shown in Eq. (2)

$$R_{ijmn} = \max\left[0, \beta_m \{x_{jmn} - x_{imn}\}\right] \approx \ln\left(1 + \exp\left(\beta_m \{x_{jmn} - x_{imn}\}\right)\right), \quad (2)$$

which can be seen either as a plausible approximation (see Figure 1 in Chorus, 2010) or as the result of assuming unobserved heterogeneity in the regret function.

The regret function R_{in} of alternative *i* for agent *n* is completed by summing R_{ijmn} over all the attributes *m* and alternatives $j \neq i$ in the choice-set C_n, as shown in Eq. (3).

$$R_{in} = \sum_{\substack{j \in C_n \\ j \neq i}} \sum_{m=1}^{M} \ln\left[1 + \exp\left(\beta_m \{x_{jmn} - x_{imn}\}\right)\right], \quad (3)$$

Finally, it is considered that the individual seeks minimizing a random regret function $RR_{in} = R_{in} + \varepsilon_{in}$, where ε_{in} is a random term, whose negative is assumed to be independent and identically distributed (*iid*) Extreme Value (0,μ). Under those conditions, the probability that agent *n* will choose alternative *i* will correspond to the model shown in Eq. (4)

$$P_n(i) = \frac{e^{-\mu R_{in}}}{\sum_{j \in C_n} e^{-\mu R_{jn}}}, \quad (4)$$

where the scale parameter μ is indistinguishable from the overall regret scale, and therefore, for convenience, is normalized to equal 1.

Consider now that the researcher samples a subset D_n with $\tilde{J}_n$ elements from the true choice-set C_n that is considered by the decision maker. As was stated before, the sampling may be needed for reducing the computational burden and/or facilitate data collection. For estimation purposes, D_n must include the chosen alternative *i*, and then D_n is not independent of *i*. If *i* is not included in D_n, a probability measure constructed combining *i* and the elements in D_n may not be well defined, since it may be larger than one. Also, if *i* is not included in D_n, the likelihood may be unbounded, precluding model estimation.

Define $\pi(i, D_n)$ the joint probability that agent *n* would choose alternative *i* and that the researcher would draw the set D_n. Using the Bayes theorem, this joint probability can be rewritten as shown in Eq. (5)

$$\pi(i, D_n) = \pi(D_n \mid i) P_n(i) = \pi(i \mid D_n) \pi(D_n), \quad (5)$$

where $\pi(i \mid D_n)$ is the conditional probability of choosing alternative *i*, given that the set D_n was drawn, and $\pi(D_n \mid i)$ is the conditional probability that the researcher drew the set D_n, given that alternative *i* was chosen by the agent.

Since the events of choosing each one of the alternatives in C_n are mutually exclusive and totally exhaustive, we can write the probability $\pi(D_n)$ of constructing the set D_n as shown in Eq. (6)

$$\pi(D_n) = \sum_{j \in C_n} \pi(D_n \mid j) P_n(j) = \sum_{j \in D_n} \pi(D_n \mid j) P_n(j), \tag{6}$$

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

Substituting Eq. (6) and the choice probability $P_n(i)$ shown in Eq. (4) into Eq. (5), Eq. (7) is obtained by canceling and re-arranging terms.

$$\pi(i \mid D_n) = \frac{e^{-R_{in} + \ln \pi(D_n \mid i)}}{\sum_{j \in D_n} e^{-R_{jn} + \ln \pi(D_n \mid j)}} \tag{7}$$

The direct application of Mcfadden´s (1978) result on sampling of alternatives for Logit can be used to show that maximizing a log-likelihood based on the expression shown in Eq. (7) would yield consistent estimators of the model parameters.

Eq. (7) shows two things about the conditional probability $\pi(i \mid D_n)$. The first is that the form of the probability is very similar to Eq. (4), except for the term $\ln \pi(D_n \mid j)$, which is known as the sampling correction. The second is that the summation in the denominator is only over the alternatives in D_n.

However, Eq. (7) does not yet offer a practical solution for the sampling of alternatives in random regret models. The problem is that, even though the denominator of the choice probability depends only on D_n, the argument R_{in} still depends on the full choice-set C_n.

In this paper, we adapt Eq. (7) to the problem of sampling of alternatives in random regret minimzation models by replacing R_i by an estimator that depends only on the subset D_n. We analyze the conditions required for consistency, asymptotic normality and efficiency, determine the correct expansion factors required in some relevant examples, and illustrate the finite sample properties of the estimators using Monte Carlo experimentation and real data.

The results on consistency, asymptotic normality and efficiency are summarized by the following theorem, which is a generalization of the result of Guevara and Ben-Akiva (2013a) to a class of models that includes the RRM model: models whose utility function depends on the full choice set.

Theorem: Consider *N* observations, a choice-set C_n of cardinality J_n, and two subsets $D_n \subseteq C_n$ and $\widetilde{D}_n \subseteq C_n$. To simplify notation, we will assume that the cardinality of D_n and $\widetilde{D}_n$ is $\widetilde{J}_n$, but this is not essential and can be generalized. If:

a) the choice model is of the Logit form, in the sense that it can be written as

$$P_n(i) = \frac{e^{W_{in}(C_n)}}{\sum_{j \in C_n} e^{W_{jn}(C_n)}},$$

where $W_{in}(C_n)$ is any continuous and twice differentiable function of the attributes x_{jn} of all the alternatives in C_n, and a set of parameters β^*. Note that $W_{in}(C_n)$ includes, but is not limited to, the regret function defined in Eq. (3).

b) $\hat{W}_{in}(\tilde{D}_n)$ is an unbiased estimator of W_{in}

c) The variance of $\hat{W}_i(\tilde{D}_n)$ is bounded and decreases with $\tilde{J}_n$. Since $\hat{W}_i(\tilde{D}_n)$ is also unbiased, this also means that $\hat{W}_i(\tilde{D}_n)$ is also consistent;

d) $\pi(D_n \mid j) > 0 \quad \forall j \in D_n$ and $\pi(D_n \mid j) = 0 \quad \forall j \notin D_n$, which holds when the chosen alternative is included in D_n;

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

$$Q\hat{L}_D = \sum_{n=1}^{N} \ln \hat{\pi}(i \mid D_n, \tilde{D}_n) = \sum_{n=1}^{N} \ln \frac{e^{\hat{W}_{in}(\tilde{D}_n) + \ln \pi(D_n \mid i)}}{\sum_{j \in D_n} e^{\hat{W}_{jn}(\tilde{D}_n) + \ln \pi_n(D_n \mid j)}} \tag{8}$$

yields, under general regularity conditions, consistent estimators 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, and asymptotically normal:

$$\hat{\beta} \overset{a}{\sim} \text{Normal}(\beta^*, \mathbf{R}^{-1}\boldsymbol{\Omega}\mathbf{R}^{-1}/N) \tag{9}$$

where $\boldsymbol{\Omega} = Var\left(\frac{\partial \ln \pi_n(\beta^* \mid D)}{\partial \beta}\right)$ and $\mathbf{R} = E\left(\frac{\partial^2 \ln \pi_n(\beta^* \mid D)}{\partial \beta \partial \beta'}\right)$, where $\pi(i \mid D_n) = \frac{e^{W_{in}(C_n) + \ln \pi(D_n \mid i)}}{\sum_{j \in D_n} e^{W_{jn}(C_n) + \ln \pi_n(D_n \mid j)}}$. This variance-covariance matrix can be approximated by the BHHH estimator (Berndt, et al. 1974), using $\hat{\pi}(i \mid D_n, \tilde{D}_n)$, evaluated at the optimal values.

Note that the variance-covariance matrix attained with Eq. (8) is then the same that is attained by maximization of the impractical quasi-log-likelihood

$$QL_D = \sum_{n=1}^{N} \ln \pi(i \mid D_n) = \sum_{n=1}^{N} \ln \frac{e^{W_{in}(C_n) + \ln \pi(D_n \mid i)}}{\sum_{j \in D_n} e^{W_{jn}(C_n) + \ln \pi_n(D_n \mid j)}}. \tag{10}$$

This implies that the feasible estimator proposed in Eq. (8) is relatively efficient, in the sense that it yields estimators that are as asymptotically efficient as the estimators obtained when considering the full choice-set to calculate $W_{in}(C_n)$. They are not globally efficient because some efficiency is lost when sampling alternatives in Eq. (10).

Finally, if J_n is finite and the protocol is sampling without replacement, $\widetilde{J}_n$ needs to increase only up to $\widetilde{J}_n = J_n$ to achieve the asymptotic distribution shown in Eq. (9).

Proof: The proof is analog to the procedure used by Train (2009, pp. 247-257) for simulated maximum likelihood. In the appendix we provide a summarized demonstration, highlighting principal parts, and including a justification for the main assumptions that are required in the theorem.

Two cases are differentiated in the theorem: when J_n is infinite and when J_n is finite.

When J_n is infinite, it is relevant to answer the question of whether or not a choice model with an infinite choice-set is, in general and in particular for the RRM, relevant and well defined. Models with infinite choice-sets had been previously considered, among others, for spatial choice, labor demand and route choice. Examples of those are the works of McFadden (1976), Ben-Akiva and Watanatada (1981), Dagsvik (1989) and, more recently, Fosgerau et al. (2013). In addition, the validity of the RRM as J goes to infinity can be proven by induction. First, the RRM model is well defined for J=2. Then, if RRM is well defined for an arbitrary J, it can be shown that it will be well defined for J+1. Nevertheless, numerical limitations in the estimation and forecasting of a model with an infinite choice set would make the model intractable, which is of course precisely what motivates the need for sampling of alternatives.

It is pertinent also to clarify what the theorem achieves when J_n is finite. For finite J_n, and if the protocol is sampling with replacement, $\widetilde{J}_n$ will have to grow infinitely with N to achieve consistency. Instead, when the protocol is sampling without replacement, if $\widetilde{J}_n$ grows with N, it will eventually reach J_n. At that point the variance of $\hat{W}_{in}\left(\widetilde{D}_n\right)$ will be zero because $W_{in} = \hat{W}_{in}$ and the quasi-loglikelihood in Eq.(8) would become the same as the quasi-loglikelihood in Eq.(10), achieving consistency.

Despite that the theoretical results hold asymptotically, the Monte Carlo experiments in Section 4 show that for finite N (1000 in the example) and finite J_n (1000), $\widetilde{J}_n$ as small as 30 can result in proper estimators. Moreover, depending on the behaviour of $Var\left(\hat{W}_{in}\right)$ for small $\widetilde{J}_n$, the theorem sheds some light on the speed of convergence, which can be useful in practice. For example, if one would like to maintain the statistical properties attained when $\widetilde{J}_n = 30$ and $N = 1000$, but with N =2000, the theoretical result states that $\widetilde{J}_n$ would have to be, at least, 45, because

$$\widetilde{J}_n = 45 \approx 30\frac{\sqrt{2000}}{\sqrt{1000}}.$$

3 APPLICATION OF THE METHOD IN PRACTICE

3.1 Introduction

For the application of the theorem to the RRM model in practice, it is convenient to note first that what occurs if, in Eq. (3), the incumbent alternative is included in the regret function.

$$\tilde{R}_{in} = \sum_{j \in C_n} \sum_{m=1}^{M} \ln\left[1 + \exp\left(\beta_m \left\{x_{jmn} - x_{imn}\right\}\right)\right] = R_{in} + M \ln(2) \tag{11}$$

Eq. (11) implies that the inclusion of the incumbent alternative implies the addition of the same constant $M \ln(2)$ to all the alternatives, where M is the number of attributes. Since this same constant cancels out, considering the incumbent alternative in the regret function has no impact in the choice probability shown in Eq. (4). For the rest of the paper we will consider the definition of the regret function including the incumbent alternative, as in Eq. (11). This will facilitate the notation of the different versions of the practical application of the method. Also, to save notation, we will skip the tilde from $\tilde{R}_{in}$.

We propose the following $\hat{R}_{in}$ as a feasible approximation of R_{in}

$$\hat{R}_{in} = \sum_{j \in \tilde{D}_n} w_{jn} \sum_{m=1}^{M} \ln\left[1 + \exp\left(\beta_m \left\{x_{jmn} - x_{imn}\right\}\right)\right]. \tag{12}$$

The expansion factors w_{jn} in $\hat{R}_{in}$ needed for attaining unbiasedness, as required by the theorem, would have to have the following form

$$w_{jn} = \frac{\tilde{n}_{jn}}{E\left(\tilde{n}_{jn}\right)}, \tag{13}$$

where $\tilde{n}_{jn}$ corresponds to the number of times alternative j is included in the sample for agent n, and $E\left(\tilde{n}_{jn}\right)$ is its expected value (see Guevara and Ben-Akiva, 2013a, Appendix B for a demonstration of an equivalent case). Note that if the protocol used to draw alternatives is sampling without replacement, $\tilde{n}_{jn} = 1$ and $E\left(\tilde{n}_{jn}\right)$ corresponds to the probability of sampling alternative j.

The expansion factors w_{jn} would depend on the sampling protocol used and, importantly, on whether or not the subset D_n used to write the sampling correction $\ln \pi\left(D_n \mid j\right)$ in Eq. (7) is the same as the subset $\left(\tilde{D}_n\right)$ used to build the expansion factors w_{jn}.

We what follows we will explore three methods to construct in practice the expansion factors shown in Eq. (13). These methods are analog to some of the approaches explored by Guevara and Ben-Akiva (2013a, 2013b) for the problem of sampling of alternatives in MEV and Logit Mixture, respectively.

3.2 Expansion Factors when Re-sampling is Possible

Consider first the case when the researcher has full control of the data and is able to sample a set D_n from C_n to build the sampling correction $\ln \pi(D_n \mid i)$, and to sample a different set $\tilde{D}_n$ from C_n to construct the expansion factors w_{jn} needed to build $\hat{R}_{in}$. To save notation we will consider that both D_n and $\tilde{D}_n$ have the same cardinality $\tilde{J}$ for all individuals, but this is not essential and can be generalized.

The expansion factors required depend on the protocol used for building $\tilde{D}_n$. In what follows we consider as an example that the protocol is a simple random sample without replacement. Note that the chosen alternative does not need to necessarily be in $\tilde{D}_n$. The sampling in this case is random from all the elements in C_n. This is crucial for the simplicity and practicality of applying this version of the method.

In such a case the expansion factors in $\hat{R}_{in}$ that are needed to achieve a an unbiased estimator of R_{in}, are the following for each alternative *j*:

$$w_{jn} = \frac{\tilde{n}_{jn}}{E(\tilde{n}_{jn})} = \frac{1}{\tilde{J}/J} = \frac{J}{\tilde{J}} \tag{14}$$

To describe the likelihood function required to estimate the model we need to specify also the sampling protocol used to build the set D_n, to be able to determine McFadden´s (1978) sampling correction. Consider, for example, that the protocol used in this case is the following. In a first step, the chosen alternative for each observation is included. Then, non-chosen alternatives are randomly sampled, without replacement, to make a total of $\tilde{J}$. Under this setting, it can be shown that the sampling correction will correspond to

$$\ln \pi_n(D \mid i) = \ln \binom{J-1}{\tilde{J}-1},$$

a term that, for this particular sampling protocol, is constant across alternatives and, therefore, cancels out in the calculation of the quasi-log likelihood function shown in Eq. (8).

To summarize, given the particular sampling protocols for D_n and $\tilde{D}_n$ described, the conditional probability of choosing alternative *i*, given that the sets D_n and $\tilde{D}_n$ were drawn, can be approximated by

$$\hat{\pi}_n\left(i \mid D_n, \tilde{D}_n\right)^{\text{Resampling}} = \frac{e^{-\sum_{j \in D_n} \frac{J}{\tilde{J}} \sum_{m=1}^{M} \ln\left[1+\exp\left(\beta_m \left\{x_{jmn} - x_{imn}\right\}\right)\right]}}{\sum_{k \in D_n} e^{-\sum_{j \in D_n} \frac{J}{\tilde{J}} \sum_{m=1}^{M} \ln\left[1+\exp\left(\beta_m \left\{x_{jmn} - x_{kmn}\right\}\right)\right]}}. \tag{15}$$

Therefore, according to the theorem, a model estimated using the quasi-log-likelihood function built using Eq. (15) will result in consistent and asymptotically normal estimators of the model parameters and the variance-covariance matrix of the estimators can be obtained using the BHHH estimator. This estimation tool is practical because it

can be applied in canned estimation software such as BIOGEME (Bierlaire, 2003) or ALOGIT (Daly, 1992) with minor modifications, making it very attractive for practitioners.

Finally, note that the intuition behind Eq. (15) is direct. If, for example, 10 out of 1000 alternatives are sampled randomly to build $\tilde{D}_n$, the regret function has to be calculated with the 10 alternatives, and then amplified by 100 to correct for the fact that regret is otherwise underestimated due to the smaller (sampled) choice set. Besides, note that since for this particular sampling protocol the expansion factor is the same for all the alternatives, the $w_{j_n} = J/\tilde{J}$ term comes out of the sum and becomes indistinguishable with the overall utility scale.

Things become more troublesome when the researcher is forced to use instead the same set D_n to build the term $\hat{R}_{in}$. We will discuss this in the next section.

3.3 Expansion Factors when Re-Sampling is Not Possible

Consider now that the researcher does not have full control of the data and is not able to sample two sets D_n and $\tilde{D}_n$. This can occur when the researcher is using a database previously processed and for which he or she does not have access to the original source, for example, because of privacy concerns.

If the protocol used to build D_n (and therefore also $\tilde{D}_n$) was to draw first the chosen alternative and then to sample $\tilde{J}-1$ alternatives randomly, the expansion factors required to attain unbiasedness are the following (see Guevara and Ben-Akiva, 2013a, Appendix C for a demonstration of an equivalent case).

$$w_{ij} = \frac{1}{P_n(j) + \frac{\tilde{J}-1}{J-1}\left(1-P_n(j)\right)}. \qquad (16)$$

There is a crucial difference between Eq. (16) and Eq. (14). The expression shown in Eq. (16) depends on the choice probabilities, which are unknown beforehand in an application with real data. To avoid this limitation in practice, we postulate two methods called *Pop.Shares* and *1_0*.

Method Pop.Shares:

One way to approximate the choice probabilities needed for the calculation of the expansion factors is to use the population shares H_j of each alternative. Replacing choice probabilities by population shares in Eq. (16), the expansion factors implied by this procedure become the following:

$$w_{jn} = \frac{1}{H_j + \frac{\tilde{J}-1}{J-1}\left(1-H_j\right)} \qquad \forall n = 1,\cdots,N; \forall j \in C_n .$$

An advantage in this case is that the expansion factors w_{jn} can be directly calculated without incurring additional computational costs. Although the true population shares are not available in a real application, good approximations of them may be available from different sources (e.g., census or flow counts), or directly from the sample, provided it is random. In case the H_j has to be gathered from the sample, it could be calculated as

$$H_j \approx \frac{\sum_n y_{jn}}{N},$$

where y_{jn} equals 1 if individual *n* chooses alternative *j*, and zero otherwise.

To summarize, given the particular sampling protocol described for D_n, the conditional probability of choosing alternative *i*, given that the set D_n was drawn, can be approximated by

$$\hat{\pi}_n\left(i \mid D_n\right)^{\text{Pop.Shares}} = \frac{e^{-\sum_{j\in D_n}\left\{\frac{1}{\frac{\sum_n y_{jn}}{N}+\frac{\tilde{J}-1}{J-1}\left(1-\frac{\sum_n y_{jn}}{N}\right)}\right\}\sum_{m=1}^{M}\ln\left[1+\exp\left(\beta_m\left\{x_{jmn}-x_{imn}\right\}\right)\right]}}{\sum_{k\in D_n} e^{-\sum_{j\in D_n}\left\{\frac{1}{\frac{\sum_n y_{jn}}{N}+\frac{\tilde{J}-1}{J-1}\left(1-\frac{\sum_n y_{jn}}{N}\right)}\right\}\sum_{m=1}^{M}\ln\left[1+\exp\left(\beta_m\left\{x_{jmn}-x_{kmn}\right\}\right)\right]}}. \quad (17)$$

The *Pop. Shares* method could be easily implemented in canned estimation software with minor modifications, making it attractive for practitioners. The disadvantage is that the approximation $H_j \approx P_n\left(j\right)$ may be too rough, potentially causing important biases. This approach is studied using Monte Carlo experiments in Section 4.

Method 1_0:

Another approach to avoid the need for the choice probabilities is to approximate them by considering that they take value 1 for the observed chosen alternative, and 0 for the non-chosen ones. Replacing these assumptions in the example described in Eq. (16) the expansion factors in this case would be the following:

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

$w_{jn} = \frac{J_n - 1}{\tilde{J}_n - 1}$ if *j* is not chosen.

It should be noted that the *1_0* approach to address the problem of estimation while sampling alternatives is the same as the one that was implicitly considered, by Frejinger et al. (2009) and by Lee and Waddell (2010), in a different context.

To summarize, given the particular sampling protocol described for D_n, the conditional probability of choosing alternative *i*, given that the set D_n was drawn, can be approximated by

$$\hat{\pi}_n(i \mid D_n)^{1-0} = \frac{e^{-\sum_{j\in D_n}\left\{y_{jn}+(1-y_{jn})\frac{J_n-1}{\tilde{J}_n-1}\right\}\sum_{m=1}^{M}\ln\left[1+\exp\left(\beta_m\{x_{jmn}-x_{imn}\}\right)\right]}}{\sum_{k\in D_n} e^{-\sum_{j\in D_n}\left\{y_{jn}+(1-y_{jn})\frac{J_n-1}{\tilde{J}_n-1}\right\}\sum_{m=1}^{M}\ln\left[1+\exp\left(\beta_m\{x_{jmn}-x_{kmn}\}\right)\right]}}. \quad (18)$$

The advantages and disadvantages of this procedure are similar to those of the *Pop.Shares* method: it can be directly implemented without using additional information and without incurring additional computational costs. Additionally, this method can be easily implemented in canned estimation software with minor modifications, making it attractive for practitioners. The disadvantage is that this approximation may be too rough and may cause important biases. This approach is studied using Monte Carlo experiments in Section 4.

4 MONTE CARLO EXPERIMENTS

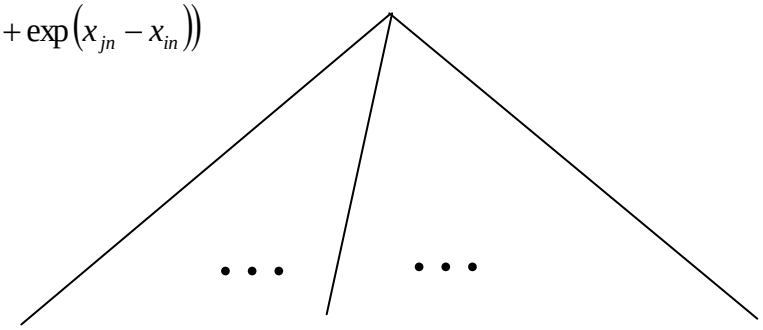
4.1 Introduction

In this section we report three Monte Carlo experiments that serve the purpose of illustrating the application of the different versions of the sampling method outlined in the previous sections. A secondary purpose of the experiments is to shed light on the relative performance of the variations of the method in finite samples, but having in mind that the results in this respect are only valid in the context of the experiments considered.

4.2 Assessment of different versions of the method

In the first experiment we analyze the empirical finite sample properties of each version of the method in recovering the true parameters of the model, depending on the number of alternatives sampled $\tilde{J}$. The setting of this experiment is summarized by Figure 2. The true or underlying model is a RRM model with 1000 alternatives and 1000 observations, with a single attribute *x* distributed Uniform (-1,1) and with parameter $\beta = 1$. The motivation for considering a single attribute in this experiment, was for being able to estimate the true model considering a number of alternatives as large as 1000 to be used as a benchmark.

$$R_{in} = \sum_{j\in C_n} \ln\left(1+\exp\left(x_{jn}-x_{in}\right)\right)$$



1 j 1000

Figure 2 Structure of the Random Regret Model for the Monte Carlo Experiment

$N = 1000$ $J = 1000$; $\tilde{J} = 5,\ 15, 30$ and 50

The methodology used to implement the RRM model shown in Figure 2 for the Monte Carlo experimentation consists of several steps. First, the choice probability was calculated using the true value of the parameter ($\beta = 1$) in Eq. (3). Then, these choice probabilities were used to build a discrete cumulative distribution 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 distribution function, evaluated for each random number.

The sampling protocol used to draw alternatives D_n from the choice-set C_n in this experiment was the following. First, the chosen alternative for each observation was included. Then non-chosen alternatives were randomly sampled, without replacement, to make a total of $\tilde{J} = 5, 15, 30$ and 50. The sampling protocol used to draw alternatives $\tilde{D}_n$, when it was considered to be different from D_n, was a simple random sample of $\tilde{J}$ alternatives from C_n.

Under this setting we estimated the model using five different methods. The first method corresponds to the *True* model, a model where all alternatives are considered in the choice-set. This model acts as a benchmark, both in terms of the maximum quality that can be attained for the estimators, and of the maximum estimation time.

The second estimation method corresponds to a *Truncated* version of the problem where only the elements in the subset D_n are used to build the term $\hat{R}_{in}^{Truncated} = \sum_{j \in D_n} \ln\left(1 + \exp\left(\beta\left(x_{jn} - x_{in}\right)\right)\right)$. This method acts as a benchmark in terms of the minimum quality that can be attained for the estimators.

The third estimation method considered is *Re-sampling*, method in which an alternative set $\tilde{D}_n$ is sampled to build the term $\hat{R}_{in}$. In this application, $\tilde{D}_n$ was drawn as a random sample without replacement, so that the expansion factors are calculated as $w_{jn} = \frac{J}{\tilde{J}}$. The quasi-loglikelihood considered in this case is the one shown in Eq. (15).

The fourth estimation method considered is *Pop. Shares*. In this case $\tilde{D}_n = D_n$, the expansion factors are calculated using the sample shares as an approximation of the choice probabilities, and the quasi-loglikelihood is the one shown in Eq. (17).

The final estimation method considered is *1_0*. In this case again $\tilde{D}_n = D_n$ the expansion factors are calculated using the observed choice as an approximation of the choice probabilities, and the quasi-loglikelihood considered in this case is the one shown in Eq. (18).

The model was generated 100 times, for different values for $\tilde{J}$. For each repetition of the model we regenerated the attribute *x*, the choices and the sets D_n and $\tilde{D}_n$. Estimation was performed using the BFGS (Fletcher, 1980) algorithm coded in the *optim* package of the open-source software R (R Development Core Team, 2008), on an IBM eServer with a CPU Intel Xeon X5560 of 2.80GHz and 12 GiB RAM.

For each model estimated we report the following statistics to assess the empirical finite sample properties of each method in estimating the model coefficient β.

Bias: Difference between average estimator within the 100 repetitions and the true value of the parameter. The Bias should tend to zero if the mean of the sampling distribution is equal to the true value.

Root Mean Squared Error (RMSE): Square root of the sum of the sampling variance and the square of the bias. The smaller the RMSE, the better is the method in terms of small sample efficiency.

t-test: Ratio between the bias and the sampling standard deviation of the average of the estimators. This statistical test can be used to test the null hypothesis that the mean of the sampling distribution is equal to its respective true value.

Count: Number of times the estimator of each repetition is within a 75% confidence interval of the true value constructed using the sampling variance from all the repetitions. This statistic is usually termed the empirical coverage. The larger this statistic is, the better the performance of the method. The closer to 75 this statistic is, the closer its empirical distribution is to its theoretical sampling distribution.

Together with these statistics, we report in Table 1 the respective $\tilde{J}$, the estimation time in minutes (Time), and the number of times -within the 100 repetitions- that the model was not estimable because of an error in the optimization procedure (Error).

Table 1: Statistical Analysis of $\hat{\beta}$ for Different Methods to Estimate RRM Model with Sampling of Alternatives, Varying $\tilde{J}$

Method	Bias	RMSE	t-test	Count	$\tilde{J}$	Time [min]	Error
True	0.005105	0.08092	0.06322	76	1000	64.61	0
Truncated	347.0	390.4	1.939	12	5	0.002615	0
	284.4	288.6	5.794	0	15	0.01639	0
	270.8	273.1	7.703	0	30	0.05473	0
	259.0	260.2	10.15	0	50	0.1447	0
Resampling	0.3848	0.6276	0.7763	0	5	0.002875	68
	-0.01060	0.4702	0.02255	88	15	0.01984	0
	0.06234	0.5419	0.1158	94	30	0.07124	0
	0.02292	0.3844	0.05972	97	50	0.1841	0
Pop Shares	0.06136	0.8994	0.06838	0	5	0.003164	4
	-0.06717	0.3184	0.2158	75	15	0.02044	0
	-0.01191	0.3216	0.03705	90	30	0.07146	0
	-0.01601	0.1993	0.08058	85	50	0.1813	0

1_0	363.9	414.1	1.841	0	5	0.004880	19
	292.5	296.7	5.923	0	15	0.02359	0
	287.8	290.0	8.155	0	30	0.09929	0
	287.9	289.0	11.30	0	50	0.3138	0

J=1000; *N*=1000; β=1; 100 repetitions; One attribute, distributed U(-1,1)

The estimation results are also summarized in Figure 3. The abscissa corresponds to the $\tilde{J}$ and the ordinate depicts the estimator $\hat{\beta}$ of the single model parameter. The value of $\tilde{J}=1000$ is not presented in scale, and the values of $\hat{\beta}$ are limited to those between 0.0 and 2.0. The true value of $\beta=1.0$ is highlighted with a horizontal line. The estimators obtained for each method and repetition, are drawn in grey with the respective symbols detailed in the legend of Figure 3 for each method. The average of the estimators, within the 100 repetitions, is marked with a larger symbol for each method.

The estimators of the *True* model, the one estimated using $\tilde{J}=1000$, are depicted with a dot in Figure 3. As expected, this model performs well. The average of the 100 repetitions is almost equal the true value of $\beta=1.0$, and each repetition is close and symmetrically around it. This is reaffirmed by the statistics deployed in Table 1. The Bias is about 0.5%. The RMSE is about 8% and t-tests are far below the critical value of 1.984 to erroneously reject the null hypothesis that $\beta=1.0$. Also, none of the 100 repetitions failed, and the empirical coverage was 76, almost equal to its nominal value of 75. Finally, the estimation time was, in average, about 1 hour per repetition.

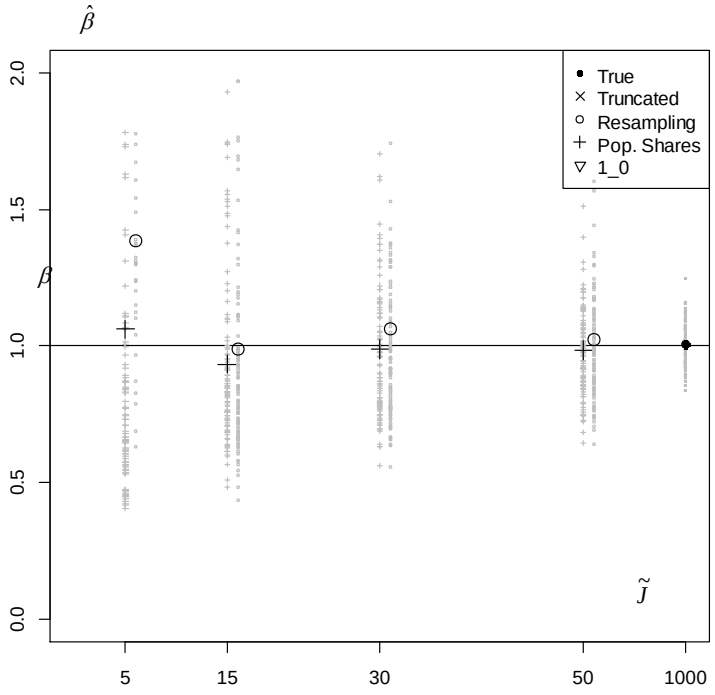


Figure 3: Estimators for Different Methods. 100 Repetitions for Various ($\tilde{J}$)

The estimators of the *Truncated* method, the one estimated ignoring the impact of the truncation of the regret function caused by the sampling, are depicted with an 'x' in Figure 3. As expected, although the Bias decreases with $\tilde{J}$ (see Table 1), the results are still very poor for $\tilde{J}=50$. Not even one estimator falls in the range of 0.0-2.0 depicted in Figure 3. Table 1 shows that the Bias is above 25000% compared to the true value, and so does the RMSE. Also, for all values of $\tilde{J}$, the t-test is above 1.984, the threshold for erroneously rejecting with 95% confidence the null hypothesis that the mean of the sampling distribution is equal to its respective true value. It is interesting that the best value of the t-test occurs for $\tilde{J}=5$, which can be explained by noting that the sample variance is larger for such small $\tilde{J}$. Finally, none of the 100 repetitions of the estimation procedure failed and the estimation time was, on average, less than 10 seconds for $\tilde{J}=50$. As a conclusion and completely in line with expectations, the Truncated method performs extremely poor in all aspects for small $\tilde{J}$, although it can be noted that results improve as $\tilde{J}$ grow, slightly, but steadily.

The estimators of the *Resampling* method, which is obtained by maximizing the quasi-loglikelihood shown in Eq.(15), are depicted with a circumference in Figure 3. This estimation method performs acceptably with $\tilde{J}$ as small as 30. From that point, the Bias is below 6% and the t-test is far below the critical value for rejecting the null hypothesis that β is equal to its true value. The RMSE is not as small as with the *True* model, but 600 times below the *Truncated* one. Also, it is interesting that 68 out of 100 of the repetitions failed for $\tilde{J}=5$, but none failed for larger $\tilde{J}$. This may be explained because the fundamental part of the method is to gather a proper estimate of the regret function with a reduced number of alternatives, and maybe with $\tilde{J}=5$ the estimator of R_{in} is so poor that it results in the estimation procedure becoming unbounded or undefined. Another possible explanation is that there might be a limitation of the estimation procedure BFGS in this context.

The estimation time of the *Resampling* method took about 11 seconds in average for $\tilde{J}=50$, which is very similar to the *Truncated* method, and about 350 times smaller than that of the *True* model. Finally, the Count for the *Resampling* method is higher than the nominal value of 75. This may reflect that 100 repetitions may not be enough, in this case, for providing a proper account of the sampling distribution, or that the finite sample distribution is not well behaved. As a conclusion, these results suggest that, although the *Resampling* method works asymptotically, various finite sample properties, particularly the Bias, are acceptable with $\tilde{J}$ as small as 30 out of 1000. However, statistical testing with finite samples should be treated with care since results suggest that the t-tests may have low power. Further investigation in this final issue is needed.

The estimators of the *Pop. Shares* method, the one obtained by maximizing the quasi-loglikelihood shown in Eq.(17), are depicted with a cross in Figure 3. This method performs as well as the *Resampling* method. For some values of $\tilde{J}$, *Pop. Shares* is superior and for others *Resampling* is superior. Failed estimations occur also only for $\tilde{J}=5$, but now in only 4 out of 100 repetitions, which suggests that this method is more

robust with regard to this respect. Estimation times are also of the same order of magnitude as for the *Resampling* method. As with the *Resampling* method, the Count in this case is larger than its nominal value. As a conclusion, the results suggest that the *Pop. Shares* method works as well as the *Resampling* method for finite samples.

Finally, the estimators of the *1_0* method, the one obtained by maximizing the quasi-loglikelihood shown in Eq.(18), are depicted with an inverted triangle in Figure 3. Results obtained with this method are very poor, in fact almost as poor as the results obtained with the *Truncated* method. As a conclusion, although the *1_0* method works asymptotically, the finite sample properties in this application are far from acceptable. The *Resampling* and the *Pop. Shares* methods both showed substantially better results.

4.3 Sensitivity to the variance of *x*

The second experiment is devised to analyze the relative performance of the methods when changing the variance of the data. The experiment is equivalent to the one described in the previous section in various aspects. There is also only one attribute *x*, the true parameter is $\beta = 1$, and there are 1000 observations. In turn, the true choice-set in this case has 500 alternatives for all individuals, and 30 alternatives are sampled. The attribute *x* is distributed Uniform $(-x_{\lim}, +x_{\lim})$, where $x_{\lim}$ varies from 0.2 to 3.0, in steps of 0.2. The data were generated 100 times and the estimators of the five methods are reported in Figure 4 and Table 2. In Table 2, only the results for $x_{\lim}$ = 0.2, 1.0 and 3.0 are reported.

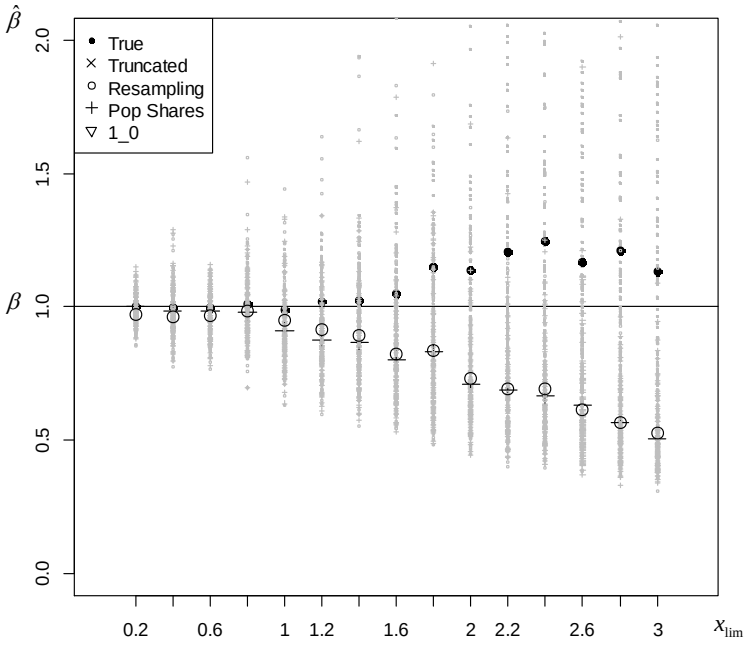


Figure 4: Estimators for Different Methods. 100 Repetitions
$J = 500; N = 1000; \tilde{J} = 30$ **;** $\boldsymbol{X}$ **follows** $U(-x_{\lim}, x_{\lim})$

The working hypotheses are fourfold. (1) A larger variance *x* will reflect in a larger variance of the statistic that is being estimated (R_{in}) by the proposed method, which implies that a larger size $\tilde{J}$ of the sampled choice set would be needed to attain the same level of error at a given confidence level. (2) A larger variance of *x* would also imply a larger variance of the choice probability, implying that a larger *N* would be needed to maintain the statistical properties. This would impact both the true model and the estimation with sampling of alternatives. (3) If the variance is too large, this may eventually cause numerical problems in the estimation methods. (4) Finally, an increase in the variance of the attribute will increase the level of information, improving the efficiency of the estimator.

Figure 4 and Table 2 show that, just as for the experiments reported in Section 4.2, the methods *Truncated* and *1_0* have a poor performance. No realizations in the range 0.0-2.0 are observed for both methods. Furthermore, the *Resampling* and *Pop. Shares* methods show very similar performance as reported in the previous sub-section. For $x_{\lim}$=1.0 and smaller, both methods perform very well, with biases below 9%, small RMSE and t-tests bellow the critical value to erroneously reject the null hypothesis that the coefficient is equal to its true value. Things become worse for larger $x_{\lim}$, both for *Resampling* and for the *Pop. Shares* method. This can be explained by a mixture of hypothesis 1 and 2. The fact that the Bias grows for $x_{\lim}$>1.6, even for the true model, suggest that from that point onwards, the second hypothesis is more relevant, which means that a larger *N* is required to maintain good statistical properties.

Table 2: Statistical Analysis of $\hat{\beta}$ for Different Methods to Estimate RRM
$J = 500; N = 1000; \tilde{J} = 30$ **,** $\boldsymbol{X}$ **follows** $U(-x_{\lim}, x_{\lim})$

Method	$x_{\lim}$	Bias	RMSE	t-test	Count	Time[min]	Error
True	0.2	0.001229	0.034632	0.035514	75	17.39	0
	1.0	-0.010732	0.071602	0.151598	77	17.21	0
	3.0	0.130163	0.504423	0.267089	79	18.54	0
Truncated	0.2	143.9	144.4	11.75	0	0.09732	0
	1.0	125.6	126.1	11.94	0	0.1113	0
	3.0	52.73	52.94	11.27	0	0.1108	0
Resampling	0.2	-0.02873	0.05970	0.5490	62	0.0686	0
	1.0	-0.05144	0.3231	0.1613	97	0.07934	0
	3.0	-0.4710	0.5009	2.765	4	0.1007	0
Pop Shares	0.2	0.00008789	0.05354	0.001641	76	0.0677	0
	1.0	-0.08756	0.1569	0.6725	60	0.08052	0
	3.0	-0.4961	0.5148	3.605	3	0.09911	0
1_0	0.2	226.5	226.9	18.15	0	0.1136	0
	1.0	142.3	142.7	13.34	0	0.1374	0

	3.0	58.23	58.42	12.38	0	0.1355	0

There is no support for hypothesis 3 for the range of values of $x_{\lim}$ analyzed, as none of the estimations failed. There is also no support for hypothesis 4 for the range of $x_{\lim}$ considered. For all cases the RMSE grows with $x_{\lim}$, which suggests that the effect in terms of efficiency is offset by the other effects. For smaller $x_{\lim}$ (not reported) the adjustment got slightly deteriorated in a similar way for both the *Resampling* and the *Pop. Shares* method.

As a conclusion, results suggest that the variance of the data impacts the $\tilde{J}$ that is needed to attain a certain statistical quality of the estimators. This implies that it is not possible to suggest a proper $\tilde{J}$ for all contexts – for example, as a given fraction of *J*. In Section 4.4 we propose a method to choose the number of alternatives to be sampled in practice.

4.4 Selection of $\tilde{J}$ in practice

The third Monte Carlo experiment was devised to illustrate how one may decide which $\tilde{J}$ to use in a practical application. In general, as was highlighted in the previous sub-section, it is not possible to provide a recommendation for $\tilde{J}$ as a fraction of *J*. The $\tilde{J}$ needed will depend, among other things, on the distribution of the data, the number of attributes, the true value of the parameters, the number of observations *N*, the optimization procedure, and the computing capabilities. The choice of a proper value for $\tilde{J}$ involves a trade-off between estimation time and quality of the estimators. The larger $\tilde{J}$, the longer it will take to estimate the model, but the better the estimates will be.

In a practical application, the researcher will have a single database. To assess the fit of the model for a given $\tilde{J}$, the researcher can sample *R* sets $D_r(\tilde{J})$ and $\tilde{D}_r(\tilde{J})$, obtaining a respective series of $\hat{\beta}_r$. With this, the following two statistics can be calculated: $\bar{\hat{\beta}} = \frac{1}{R}\sum_{r=1}^{R}\hat{\beta}_r(\tilde{J})$ and $\hat{\sigma}_{\hat{\beta}} = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}\left(\hat{\beta}_r - \bar{\hat{\beta}}_r\right)}$.

$\hat{\sigma}_{\hat{\beta}}$ can be seen as an estimator of the noise of the estimation parameter, which is equivalent to the concept of noise described for the estimation of the score in the demonstration shown in the Appendix. The larger the $\tilde{J}$, the smaller the $\hat{\sigma}_{\hat{\beta}}$ will be. Eventually, when $\tilde{J}$ =*J*, $\hat{\sigma}^2_{\hat{\beta}} = 0$. $\hat{\sigma}_{\hat{\beta}}$ is a measure that the researcher may want to constrain and trade-off with estimation time, when choosing the $\tilde{J}$ to be used in practice.

If the researcher is able to estimate the model with the full choice-set to obtain $\hat{\beta}(C)$, $\bar{\hat{\beta}}$ can be used to estimate what can be defined as the sampling bias $\bar{\hat{\beta}} - \hat{\beta}(C)$. Note that this bias is not the same as the one we considered in the context of the previous experiments. In those experiments, the bias was calculated with respect the true value of the parameter. In this case, the bias is calculated with respect to the estimator obtained

when considering the full choice-set and for a given dataset. This notion of bias is equivalent to the concept of bias described for the estimation of the score in the Appendix. The larger $\tilde{J}$, the smaller the sampling bias will be. Eventually, when $\tilde{J}=J$, the sampling bias will be zero. $\bar{\hat{\beta}} - \hat{\beta}(C)$ is thus a measure that the researcher may want to constrain and trade-off with estimation time when choosing the $\tilde{J}$ to be used in practice.

If the researcher is not able to estimate the model with the full choice-set to obtain $\hat{\beta}(C)$, $\bar{\hat{\beta}}$ can still be used directly to choose the proper $\tilde{J}$, by checking its stability. This is analog to the way that number of draws has to be chosen when estimating a model by simulated maximum likelihood, as suggested by Chiou and Walker (2007).

To illustrate this procedure we report a Monte Carlo experiment in which the true model has 1000 alternatives, there are 1000 observations and a single attribute *x* distributed Uniform (-1,1) with parameter β=1. The estimation of the true model for a particular realization of *x*, results in $\hat{\beta}(C)$=0.9322 and is performed in about 57 minutes.

Table 3 summarizes the statistics obtained using the *Pop. Shares* method for various $\tilde{J}$, considering *R*=30 repetitions for each $\tilde{J}$. It should be remarked that the repetitions in this case as not the same as in the previous experiments where *x*, the choices and the choice-sets were regenerated each time. In this case, the only thing that changes across repetitions is $D_r(\tilde{J}) = \tilde{D}_r(\tilde{J})$. For completeness, we also include in Table 3 the number of times, within the 30 repetitions, that the optimization procedure failed.

Table 3: Practical Determination of $\tilde{J}$

$\tilde{J}$	**Sampling Bias**	$\bar{\hat{\beta}}$	$\hat{\sigma}_{\hat{\beta}}$	**Time [Seconds]**	**Error**
5	-0.08610	0.8461	0.3090	0.1663	4
15	0.09320	1.025	0.6257	1.180	0
30	-0.03381	0.8984	0.1526	3.792	0
50	-0.02142	0.9107	0.1278	10.29	0
100	-0.001991	0.9302	0.0893	45.40	0

J=1000; N=1000; 30 Repetitions; Population Shares Method
$\hat{\beta}(C)$=0.9322; Time (C)=57 minutes

To choose a proper $\tilde{J}$, the researcher would have to make a decision on the acceptable sampling bias, shift in $\bar{\hat{\beta}}$, noise $\hat{\sigma}_{\hat{\beta}}$; and on a desirable estimation time. An additional criterion would also be to consider not having errors in the estimation procedure. These criteria could be accomplished, for example, for $\tilde{J}=30$, because it has a sampling bias below 5%, $\hat{\sigma}_{\hat{\beta}}$ of about 15% , estimation time below 5 seconds, the shift in $\bar{\hat{\beta}}$ about 1%, and not a single failed estimation within the 30 repetitions.

4.5 Conclusion

The Monte Carlo experiments show that the method we proposed for sampling of altrnatives in the context of RRM models, is practical. Results suggest that, in the context of the experiments, *Pop. Shares* and *Resampling* versions of the method seem to provide acceptable results for samples of alternatives as small as 30 out of 1000 alternatives. They also illustrate that the sample size that is needed for obtaining a given level of quality of the estimates, depends on many features, including the distribution of the data. This means that it is not possible to provide a simple criterion for the choice of the proper sample size, such as that $\tilde{J}$ has to be some fraction of the true *J*. In turn, the approach described in Section 4.4 is recommended, to decide based on the trade-off between estimation time and goodness of fit.

Besides, the experiments suggest that, although the finite sample bias can be small for small values of $\tilde{J}$, the power of the t-tests with finite samples may be low. This can be inferred from the observation that empirical coverage tended to be larger than its nominal value for most cases. This issue should be analyzed in further research.

Finally, it should be remarked that the relative assessment of the methods reported in this section is only valid for the experiments reported, and does not represent a complete description of the finite sampling properties of the estimators.

5 APPLICATION WITH REAL DATA

Finally, in this section we revisit a real data experiment used by Chorus (2010) to demonstrate the RRM model. The data concern revealed parking choices and was collected by Van der Waerden et al. (2008) at the campus of Eindhoven University of Technology. The choice set consists of 14 parking lot alternatives and 350 cases (which is the sample used for estimation by Chorus, 2010).

The choice model considers four attributes of the parking lots. The first is NR_SPACES, which corresponds to the number of spaces available at each parking lot. The second is ROOM_MANEUV, which is a dummy that takes value 1 if the parking lot has extra space for making maneuvers. The third attribute is RIGHT_OF_WAY, which is a dummy that takes value 1 if the driver has right-of-way when leaving the parking lot. The fourth and last attribute is DISTANCE, which corresponds to the distance between agent's workplace and the parking lot, and is discretized as follows: equals 1 when the distance is approximately 100 meters; equals 2 when distance is approximately 300 meters; and equals 3 when it is approximately 500 meters.

Although the choice set may not seem particularly large (J=14), as a proof of concept of the method we preferred not to generate a pseudo-synthetic experiment with a larger choice set (as in Bierlaire et. al, 2008) but using the real data as they were. The reason is that a pseudo-real dataset will not really offer fundamentally new insights compared to the experiments described in Section 4; additionally, this model will allow illustrating the behavior of the method with various attributes and providing additional support to the statement that the choice of the proper $\tilde{J}$ cannot be specified as a given fraction of *J*.

Table 3 summarizes the estimators obtained for the true RRM model of parking lot choices, using all the 14 alternatives available. These results are the same as the ones reported by Chorus (2010). In what follows we will use the *Resampling* method to estimate the model with sampling of alternatives, varying $\tilde{J}$ from 2 to 14. The model was estimated 30 times for each $\tilde{J}$.

We report in Figure 5 the average estimators $\bar{\hat{\beta}}$ within the 30 repetitions. The values of $\beta(C)$, which is the corresponding parameter attained with the true model as reported in Table 3, are depicted with a dashed line. We also report a bandwidth of 10% deviation from each $\beta(C)$. As expected, all $\bar{\hat{\beta}}$ get closer to $\beta(C)$ as $\tilde{J}$ grows. However, the speed of convergence is heterogeneous. On the one hand, $\bar{\hat{\beta}}$ for NR_SPACES is within the 10% bandwidth form $\tilde{J}$ =2. On the other hand, for DISTCANCE, this occurs only as $\tilde{J}$ =13 out of 14.

Table 3: RRM True Model of Parking Lot Choices

	$\hat{\beta}$	s.e
NR_SPACES	0.08671	(0.01430)
ROOM_MANEUV	0.09066	(0.02750)
RIGHT_OF_WAY	0.03387	(0.02763)
DISTANCE	-1.444	(0.4517)
L(0)	-923.7	
L($\hat{\beta}$)	-404.7	
ρ^2	0.5619	
N	350	
J	14	

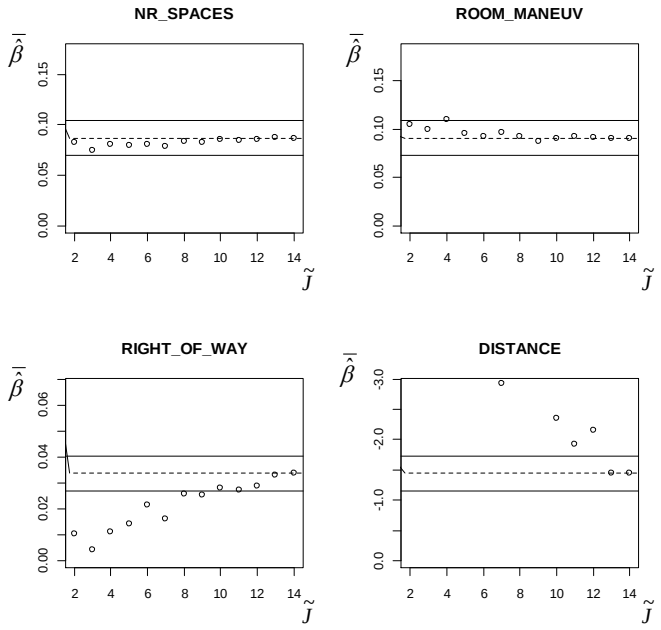


Figure 5: $\bar{\hat{\beta}}$ as a Function of $\tilde{J}$ for the RRM of Parking Lot Choice

In Figure 6 we report the standard deviation $\hat{\sigma}_{\hat{\beta}}$ for each $\tilde{J}$. Note that $\hat{\sigma}_{\hat{\beta}} = 0$ for $\tilde{J}$ =14. This value is depicted with a dashed line. As expected, all $\hat{\sigma}_{\hat{\beta}}$ shrinks as $\tilde{J}$ grows. However, as with $\bar{\hat{\beta}}$, the behavior is heterogeneous. $\hat{\sigma}_{\hat{\beta}}$ for NR_SPACES is always below 0.04, while for DISTANCE, it only occurs for $\tilde{J}$ =14.

The heterogeneity in $\bar{\hat{\beta}}$ and $\hat{\sigma}_{\hat{\beta}}$ illustrates that, when choosing $\tilde{J}$ in a model with various attributes the researcher would have to consider some type of norm to account for the degree of convergence of the full vector of parameters. A robust strategy could be to consider the convergence of the worst behaved parameter. In addition, the fact that for one of the parameters a somehow reasonable convergence is attained only for $\tilde{J}$ =93% of *J* serves to illustrate that the choice of $\tilde{J}$ cannot be settled as a fixed fraction of *J*.

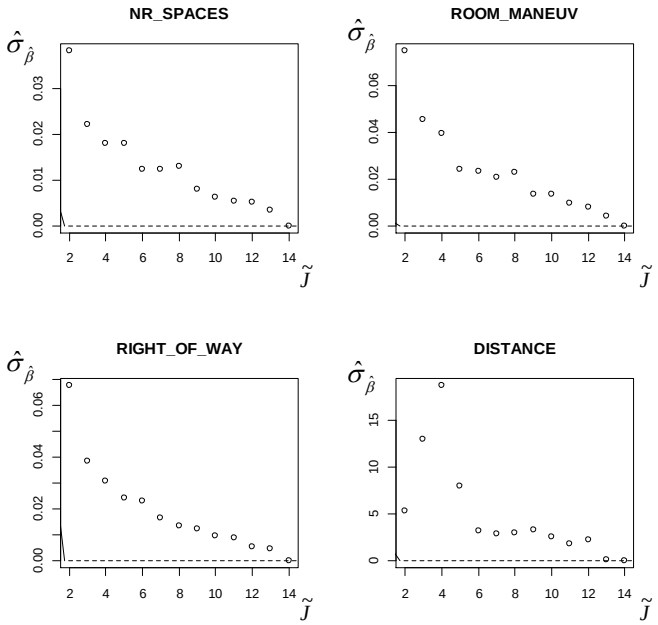


Figure 6: $\hat{\sigma}_{\hat{\beta}}$ as a Function of $\tilde{J}$ for the RRM of Parking Lot Choice

6 CONCLUSION

This article proposes a method to obtain of consistent, asymptotically normal, and efficient estimators (i.e., efficient relative to any other estimator using the same sample) for the problem of sampling of alternatives in the context of Random Regret Minimization models (RRM). In light of the fact that runtimes of RRM models increase almost quadratically with choice-set size, finding a proper way to estimate RRM-models on sampled choice-sets is a crucial condition to ensure that the RRM approach remains a feasible and attractive alternative for Random Utility Maximization-models (RUM) in the context of (very) large choice-sets. Given that the RRM-model, even when written in Logit form (i.e., with *iid* errors), does not exhibit the IIA property, McFadden's (1978) result cannot be applied to obtain a proper correction term when choice-sets are sampled. To overcome this situation, a tailor-made correction approach for RRM-models is presented in this paper, which is a direct extension of the approach developed by Guevara and Ben-Akiva (2013a) to address a similar problem in RUM-based MEV models.

In line with expectations, Monte Carlo experiments showed that sampling of alternatives causes a significant bias in the estimators of the RRM-model parameters and in the estimated shares when no correction is applied. In addition, these experiments as well as an application on real data show that the proposed method for correcting the terms that get truncated because of the sampling, performed reasonably well. In cases where the researcher has full control of the data and 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, one practical approximation methods showed reasonably good results.

The sample size required to obtain good estimators while sampling alternatives in Random Regret models will vary on a case-by-case basis and cannot be expressed as a percentage of the cardinality of the true choice-set. Using synthetic and real data, we show that in general, an appropriate strategy to determine if the size of the sample of alternatives is large enough is to test the stability of the estimators with different number of alternatives sampled and to analyze the sampling bias and noise.

ACKNOWLEDGMENTS

Funding for this research came in part from Fondecyt, Chile, through grant N°11110131, and the Singapore National Research Foundation under the Future Urban Mobility research group of the Singapore-MIT Alliance for Research and Technology. Support from The Netherlands Organization for Scientific Research (NWO), in the form of VIDI-grant 016-125-305, is gratefully acknowledged by the second author. We are also grateful for the valuable comments of two anonymous referees and of the participants of TRB XCII and TRISTAN VIII, where preliminary versions of this paper were presented. All Monte Carlo and real data experiments were generated and/or estimated using the open-source software R (R Development Core Team, 2008).

REFERENCES

Azaiez, I., 2010, Sampling of Alternatives for Logit Mixture Models. Master Thesis. Transport and Mobility Laboratory, EPFL, Switzerland.

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., & Watanatada, T. (1981). Application of a continuous spatial choice logit model. Structural analysis of discrete data with econometric applications, 320-343.

Berkovec, J., & Rust, J. (1985). A nested logit model of automobile holdings for one vehicle households. *Transportation Research Part B: Methodological*, 19(4), 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.

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., 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.

Bowman, J. L., & Ben-Akiva, M. E. (2001). Activity-based disaggregate travel demand model system with activity schedules. *Transportation Research Part A: Policy and Practice*, 35(1), 1-28.

Bradley, M., Bowman, J. L., & Griesenbeck, B. (2010). SACSIM: An applied activity-based model system with fine-level spatial and temporal resolution. *Journal of Choice Modelling*, 3(1), 5-31.

Carrasco, N. (2008). Deciding where to shop: disaggregate random utility destination choice modeling of grocery shopping in canton zurich (Doctoral dissertation, Master Thesis, IVT, ETH Zurich, Zurich).

Chiou, L., and J. Walker, 2007. Masking identification of discrete choice models under simulation methods. *Journal of Econometrics* **141.2**, 683-703.

Chorus, C.G., 2010. A new model of Random Regret Minimization. *European Journal of Transport and Infrastructure Research*, **10**(2), 181-196

Chorus, C.G., 2012. *Random regret-based discrete choice modeling: A tutorial*. Springer Briefs in Business, Springer, Heidelberg, Germany

Chorus, C.G., van Cranenburgh, S., Dekker, T., 2014. Random Regret Minimization for consumer choice modeling: Assessment of empirical evidence. *Journal of Business Research* (conditionally accepted for publication)

Daly, A., Hess, S. & Dekker, T. 2013. Sampling of alternatives for spatial choice modelling, paper presented at the 13th triennial WCTR conference, Rio de Janeiro

Daly, A.J., 1992. *ALOGIT 3.2 User's Guide*, Hague Consulting Group, The Hague, the Netherlands

Dagsvik, J. K. (1989). The generalized extreme value random utility model for continuous choice (No. 1989-41). Tilburg University, Center for Economic Research.

Fletcher, R. (1980). Practical methods of optimization. John Wiley & Sons.

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

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

Guevara C.A. 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.

Guevara, C. A., & Ben-Akiva, M. E. (2013a). Sampling of alternatives in Multivariate Extreme Value (MEV) models. Transportation Research Part B: Methodological, 48, 31-52.

Guevara, C. A., & Ben-Akiva, M. E. (2013b). Sampling of alternatives in Logit Mixture models. Transportation Research Part B: Methodological, 58, 185-198.

Lee, B. H., & Waddell, P. (2010). Residential mobility and location choice: a nested logit model with sampling of alternatives. *Transportation*, 37(4), 587-601.

Lemp, J., Kockelman, K., 2012. Strategic sampling for large choice-sets in estimation and application. *Transportation Research Part A* 46 (3), 602–661.

McConnel, K., Tseng, W., 2000. Some preliminary evidence on sampling of alternatives with the random parameters logit. *Marine Resource Economics* 14 (4), 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.

McFadden, D., 1976. The mathematical theory of demand models. In: Stopher, P., Meyburg, A. (Eds.), Behavioral Travel Demand Models. Lexington Books, pp. 305–314.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. Chapter 4 in Frontiers in Econometrics, 105-142.

McFadden, D., Train, K.E., 2000. Mixed MNL models for discrete response. *Journal of Applied Econometrics*, **15(**5), pp. 447-470

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.

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

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

Van der Waerden, P., Borgers, A. and Timmermans, H.J.P. 2008. Modeling parking choice behavior in business areas. Paper presented at the 87th annual meeting of the Transportation Research Board, Washington D.C.

Appendix

The demonstration of the theorem is analog to the two step procedure used by Train (2009, pp. 247-257) to derive the asymptotic distribution of simulation-based estimators. The first step consists in the derivation of the distribution of the approximated score

$$\hat{g}(\beta)=\frac{1}{N}\sum_{n=1}^{N}\frac{\partial \ln \hat{\pi}_n(\beta \mid D_n)}{\partial \beta}=\frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial \beta}\ln\frac{e^{\hat{W}_{in}(\tilde{D}_n)+\ln \pi(D_n|i)}}{\sum_{j\in D_n}e^{\hat{W}_{jn}(\tilde{D}_n)+\ln \pi_n(D_n|j)}},$$

relative to the true score

$$g(\beta)=\frac{1}{N}\sum_{n=1}^{N}\frac{\partial \ln \pi_n(\beta \mid D_n)}{\partial \beta}=\frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial \beta}\ln\frac{e^{W_{in}(C_n)+\ln \pi(D_n|i)}}{\sum_{j\in D_n}e^{W_{jn}(C_n)+\ln \pi_n(D_n|j)}}.$$

The second step is to derive the distribution of $\hat{\beta}$, noting that $\hat{\beta}$ is the root of the equation $\hat{g}(\hat{\beta})=0$.

In what follows we provide a summarized account of the first step, just to highlight why $\hat{W}_{in}(\tilde{D}_n)$ needs to be an unbiased estimator of W_{in}, and why the variance of $\hat{W}_{in}$ needs to be bounded and decrease with $\tilde{J}$, what also means that $\hat{W}_{in}(\tilde{D}_n)$ is a consistent estimator of W_{in}. The reader is referred to Train (2009, pp. 247-257) or Guevara and Ben-Akiva (2013a), for further details.

To simplify the notation we will assume that $\tilde{D}_n=\tilde{D}$ for all *n*. Consider $\hat{g}(\beta)$ in the vicinity of the true values β^* in the following form

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

The first term $A_1=g(\beta^*)$ is the statistic that is being approximated by $\hat{g}(\beta^*)$. The second term A_2 corresponds to the bias of the estimator of $g(\beta^*)$ and the third term A_3 is the noise of the approximation.

The noise (A_3) corresponds to the deviation of $\hat{g}(\beta^*)$ from its expected value, which will depend on a particular draw of the alternatives to construct the choice-set $\tilde{D}$. Since $\hat{W}_i(\tilde{D}_n)$is bounded and decreases with $\tilde{J}_n$, we can claim that the same occurs with the variance of the noise. This can be expressed as $Var(A_{3n})=S_n/\tilde{J}$, where S_n is the variance of A_3 for a given *n* when $\tilde{J}=1$. Then, by 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:

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

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

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

and the noise will vanishes as *N* increases, even if $\tilde{J}$ is fixed.

The bias term A_2 can be studied by taking a second order Taylor's approximation of $\hat{W}_{in}\left(\tilde{D}_n\right)$ around $\hat{W}_{in}\left(\tilde{D}_n\right)=W_{in}$. Noting that $\hat{g}_n(\beta,W_{in})=g_n(\beta)$, it follows that

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

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

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

The fact that $Var(W_n(\beta))$ is bounded and decreases with $\tilde{J}$ can be captured by the expression $Var\left(\hat{W}_n(\beta)\right)=K_n/\tilde{J}$, where K_n is a scalar. Then, the expected value of the bias A_2 can be rewritten as $A_2=\frac{Z}{\tilde{J}}$, 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.

In summary, 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 $g(\beta)$.