Sampling of Alternatives in Random Regret Minimization Models

C. ANGELO GUEVARA
Faculty of Engineering and Applied Sciences
Universidad de los Andes, Chile
San Carlos de Apoquindo, 2200, Las Condes, Santiago, Chile
Tel:56-2-618-1364
Fax: 56-2-618-1642
aguevara@uandes.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.

Word Count: 5,420 + 3 Tables + 1 Figure = 6,420 'words'

Submitted for presentation only at TRB 2013

ABSTRACT

We propose a methodology to achieve consistency, asymptotic normality and efficiency, while sampling alternatives in Random Regret Minimization models. Our method is an extension of previous results for Logit and MEV models. We illustrate the methodology using Monte Carlo experimentation. Experiments show that the proposed methodology is practical, that it outperforms the uncorrected model, and that it yields acceptable results.

1 Introduction

When choice sets are very large, like is the case in many route- and destination-choice models, sampling of alternatives becomes necessary to ensure the practical feasibility of discrete choice-model formulation and estimation. 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. This estimator capitalizes on the fact that, due to its independently and identically distributed (or: *iid*) errors, the RUM-based Logit model exhibits the IIA-property. This property states that the ratio of choice probabilities of any two alternatives in a choice set depends only on the performance (or: utility) of these two alternatives and not on that of other, 'irrelevant' alternatives.

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 between 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 distribution (McFadden, 1978). More recently, Mixed MEV-models have been proposed which allow for even more flexibility in terms of the specification of error term distribution and related behavioral implications and substitution patterns (e.g., McFadden & Train, 2000). Over the years, estimators have been proposed based on sampled choice sets in the context of these more advanced models (Manski & Lerman, 1977; Garrow et al., 2005; Bierlaire et al., 2008).

Recently, a choice model has been approach 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 between 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 between vehicle types, destinations, modes, routes, departure times, and driving maneuvers; politicians' choices between policy options; patients choices between medical treatments; and tourists' choices between leisure activity-locations. Recent studies on RRM can be found in, for example, Chorus & de Jong (2011), Thiene et al. (2012), Boeri et al. (2012), Kaplan & Prato (2012), Hensher et al. (2012), and Bekhor et al. (2012).

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. 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 his point it should be noted that, because of the fact that the 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*.

However, Guevara and Ben-Akiva (2010) recently proposed a method to address sampling of alternatives in MEV models, which consists in expanding the components that get truncated because of the sampling. This paper extends the work of Guevara and Ben-Akiva (2010) 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, determines the correct expansion factors required in some relevant examples (section 3), and illustrates and studies the finite sample properties of the estimators using Monte Carlo experimentation (section 4).

2 Estimation and Sampling of Alternatives in Random Regret Models

Consider that the random regret RR_{in}, which an agent n retrieves from an alternative i, can be written as the sum of a systematic part R and a random error term ε, as shown in Eq. (1)

$$RR_{in} = R_{in} + \varepsilon_{in} = \sum_{j \neq i} \ln\left(1 + \exp\left(\beta\left(x_{jn} - x_{in}\right)\right)\right) + \varepsilon_{in}, \tag{1}$$

where the systematic regret depends on variables x and parameters β*. Note that for reasons of ease of communication, and without loss of general applicability, we consider in this paper the simplified case where alternatives are evaluated in terms of a single attribute or variable.

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

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

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

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

Term $\pi(i, D_n)$ is 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. (3)

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

where $\pi(i \mid D_n)$ is the conditional probability of choosing alternative i, given that the set D_n was 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 use the Total Probability theorem (see, e.g., Bertsekas and Tsitsiklis, 2002) to write the probability $\pi(D_n)$ of constructing the set D_n as shown in Eq. (4)

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

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

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

$$\pi(i \mid D_n) = \frac{\pi(D_n \mid i) P_n(i)}{\pi(D_n)} = \frac{\pi(D_n \mid i) P_n(i)}{\sum_{j \in D_n} \pi(D_n \mid j) P_n(j)} = \frac{\pi(D_n \mid i) \dfrac{e^{-R_{in}}}{\sum_{k \in C_n} e^{-R_{kn}}}}{\sum_{j \in D_n} \pi(D_n \mid j) \dfrac{e^{-R_{jn}}}{\sum_{k \in C_n} e^{-R_{kn}}}}$$

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

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 in the expression shown in Eq. (5) would yield consistent estimators of the model parameters.

Eq. (5) 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. (2), except for the term $\ln \pi(D_n \mid j)$, which is known as the sampling correction. The second is that the denominator depends only on the alternatives in D_n. These simplifications result from the cancellation of the denominators when dividing the probabilities of two alternatives, which is a convenient mathematical property of Logit that results from considering that the error is *iid* across alternatives.

However, Eq. (5) does not yet offer a practical solution for the sampling of alternatives in random regret models. This follows from the simple fact that, even though the denominator of the choice probability depends only on D_n, the argument R_i still depends on the full choice-set C_n.

In this paper, we adapt Eq. (5) to the problem of sampling of alternatives in random regret models by replacing R_i by an estimator that depends only on the subset D_n. The method we propose is a direct extension of the method proposed by Guevara and Ben-Akiva (2010) for addressing the sampling of alternatives in MEV models, which is develped in detail by Guevara (2010). 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.

The results on consistency, asymptotic normality and efficiency are summarized by the following theorem:

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

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

b) the choice model is RRM and $R_{in} = \sum_{j \neq i} \ln\left(1 + \exp\left(\beta\left(x_{jn} - x_{in}\right)\right)\right)$,

c) $\hat{R}_{in}(D_n)$ is an unbiased and consistent (in $\tilde{J}_n$) estimator of R_{in},

d) The variance of $\hat{R}_i(D_n)$ is bounded and decreases with $\tilde{J}_n$, which can be written as $Var(\hat{R}_{in}) = K_n / \tilde{J}_n$ where K_n is a scalar;

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

$$QL_{MEV,D} = \sum_{n=1}^{N} \ln \hat{\pi}(i \mid D_n) = \sum_{n=1}^{N} \ln \frac{e^{-\hat{R}_{in} + \ln \pi(D_n \mid i)}}{\sum_{j \in D_n} e^{-\hat{R}_{jn} + \ln \pi_n(D_n \mid j)}} \tag{6}$$

yields, under general regularity conditions, consistent estimators (in N) of the model parameters β^*, as $\tilde{J}_n$ increases with N at any rate. If $\tilde{J}_n$ increases faster than $\sqrt{N}$, the estimators of the model parameters will be consistent, asymptotically normal

$$\hat{\beta} \overset{a}{\sim} \text{Normal}\left(\beta^*, \mathbf{R}^{-1}\mathbf{W}\mathbf{R}^{-1}/N\right),$$

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

and as asymptotically efficient as the estimators obtained from the maximization of a quasi-log-likelihood shown on Eq. (6). Finally, if J_n is finite and the protocol is sampling without replacement, $\tilde{J}_n$ needs to increase only up to $\tilde{J}_n = J_n$ in order to achieve consistency and relative efficiency.

Draft Proof: Given that $\hat{R}_{in}$ is a consistent estimator of R_{in}, as $\tilde{J}_n$ increases, the Slutsky theorem guarantees that $\hat{\pi}(i \mid D_n)$will be a consistent estimator of $\pi(i \mid D_n)$, because it is continuous. Then, McFadden's consistency results for Logit guarantee that the maximization of the quasi-log-likelihood shown in Eq. (6) will result in the consistent estimation of the model parameters as *N* increases.

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

The asymptotic distribution of the estimators of the model parameters that result from the maximization of the quasi-log-likelihood shown in Eq (6) can be derived using the two-stage approach employed by Train (2009, pp. 247-257) to analyze the asymptotic properties of simulation-based estimators. In a first stage, we will analyze the asymptotic distribution of the sample average of the score, which is defined as the gradient of the quasi-log-likelihood shown in Eq. (6). In a second stage we will use those results to derive the asymptotic distribution of the estimators of the model parameters. The derivation of this result is not detailed in this article, but is fully equivalent to the derivation described by Guevara (2010, Ch.5) for sampling of alternatives in MEV models. Following this derivation it can be shown that

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

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

Note that $\boldsymbol{\Omega}$ is usually defined as the "robust" or "sandwich" variance-covariance matrix of the estimators of the model parameters (see, e.g., Train, 2009, pp. 201). Berndt *et al.* (1974) proposed an estimator of $\boldsymbol{\Omega}$ that is known as the BHHH matrix, which, for the problem deployed here corresponds to the following expression:

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

These results imply that the estimators obtained by the maximization of the quasi-log-likelihood function shown in Eq. (6) will have the same asymptotic variance-covariance matrix as the estimators that would be obtained by using Eq. (5), that is, if the full choice-set *C* is available for the calculation of the expansion of the term R_{in}. Consequently, it can be affirmed that estimators obtained by maximizing Eq. (6) are efficient among all possible approximations of the choice probability described in Eq. (5).

3 Application of the method

3.1 Formulation

If a sample D_n is drawn from the true choice-set C_n, the only term that would be affected (and therefore needs to be approximated) is

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

R_{in} can be approximated by constructing an expanded sum. Then, to apply the method, the first step is to determine the expansion factors w_{jn} required to obtain an unbiased and consistent estimator

$$\hat{R}_{in} = \sum_{\substack{j \neq i \\ j \in D_n}} w_{jn} \ln\left(1 + \exp\left(\beta\left(x_{jn} - x_{in}\right)\right)\right).$$

For attaining unbiasedness and consistency, the expansion factors w_{jn} in $\hat{R}_{in}$ have to have the following form

$$w_{jn} = \frac{\tilde{n}_{jn}}{E\left(\tilde{n}_{jn}\right)}, \tag{7}$$

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. Note that if the sampling protocol is 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} depend on the sampling protocol used and, importantly, on whether or not the same subset is used to write the sampling correction $\ln \pi(D_n \mid i)$ in Eq. (5) and to build the expansion factors w_{jn}.

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 then 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 $\tilde{J}_m$ but this is not essential and can be easily generalized.

The expansion factors required depend on the protocol used for building $\tilde{D}_n$. In what follows we consider that the protocol is a simple random sample without replacement. In such a case the expansion factors in $\hat{R}_{in}$ are the following for each alternative *j*:

$$w_{jn} = \frac{\tilde{n}_{jn}}{E(\tilde{n}_{jn})} = \frac{1}{\tilde{J}/J} = \frac{J}{\tilde{J}} \tag{8}$$

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 so McFadden´s (1978) sampling correction can be determined. 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 McFadden´s (1978) 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. (6).

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

$$\pi_n(i \mid D_n, \tilde{D}_n) = \frac{e^{-\hat{R}_{in}(\tilde{D}_n)}}{\sum_{j \in \tilde{D}_n} e^{-\hat{R}_{jn}(\tilde{D}_n)}}$$

where $\hat{R}_{in} = \sum_{\substack{j \neq i \\ j \in \tilde{D}_n}} \frac{J}{\tilde{J}} \ln\left(1 + \exp\left(\beta\left(x_{jn} - x_{in}\right)\right)\right)$.

Therefore, a model estimated using the quasi-log-likelihood function built using this expression 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 directly applied in canned estimation software such as BIOGEME (Bierlaire, 2003) or ALOGIT (Daly, 1992) with minor modifications, making it more attractive for practitioners. Things become more troublesome when the researcher does not have full control of the data and is forced to use instead the same set D_n to build the term $\hat{R}_{in}$.

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, or when the data available corresponds to a random sample 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 consistency and unbiasedness are the following

$$w_{ij} = \frac{1}{P_n(j) + \frac{\tilde{J} - 1}{J - 1}\left(1 - P_n(j)\right)}. \tag{9}$$

There is a crucial difference between Eq. (9) and Eq. (8). The expression shown in Eq. (9) 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 *W* of each alternative. Replacing choice probabilities by population shares in Eq. (9), the expansion factors implied by this procedure become the following:

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

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 are clearly available from different sources (Census data for spatial choice models or flow counts in route choice modeling), at least at the level of the nests. As in the *1_0* method (please see below), *Pop. Shares* can also be easily implemented in canned estimation software with minor modifications, making it more attractive for practitioners. The disadvantage is that the approximation may be too rough and may cause 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, considering that it takes value 1 for the observed chosen alternative, and 0 for the non-chosen ones. Replacing these assumptions in the example described in Eq. (8) the expansion factors in this case will be the following:

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

$w_{jn} = \frac{J_{m(j)} - 1}{\tilde{J}_{m(j)} - 1}$ if j is not chosen.

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 more attractive for practitioners. The disadvantage is that the approximation may be too rough and may cause important biases. This approach is studied using Monte Carlo experiments in Section 4.

4 Monte Carlo Experiment

A Monte Carlo experiment was performed to illustrate the application of the proposed method for achieving consistency, efficiency and asymptotic normality in the case of sampling of alternatives in Random Regret Minimization models. We analyze the efficacy and efficiency of each method in recovering the true coefficients of each model depending of the number of alternatives sampled.

The setting of this experiment is summarized by Figure 1. The true or underlying model is a Random Regret model where the true coefficient of the attribute equals 1, implying a regret model of the form $R_{in} = \sum_{\substack{j \neq i \\ j \in C_n}} \ln\left(1 + \exp\left(x_{jn} - x_{in}\right)\right)$. The choice is between 30 alternatives, there are 2000 observations and the attribute *x* is distributed Uniform(-1.5,1.5).

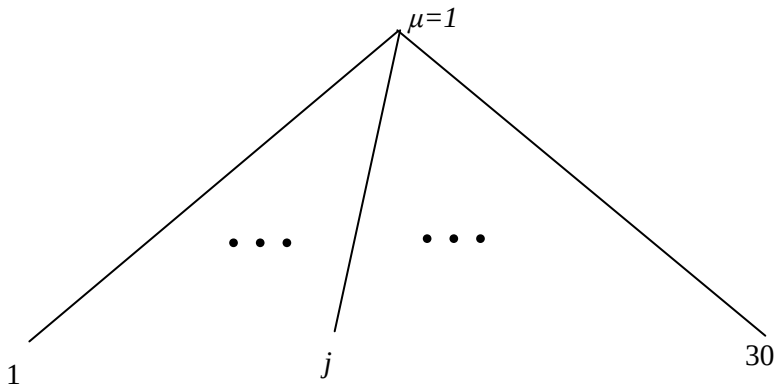

Figure 1 Structure of the Random Regret Model for the Monte Carlo Experiment

N=2,000 $J = 30$; $\tilde{J} = 5, 10, 15, 20$ and 25

The methodology used to implement the Random Regret model shown in Figure 1 for Monte Carlo experimentation was performed in several steps. First, the choice probability was calculated replacing the true value of the parameter (which equals 1) in Eq. (2). Then, these choice probabilities were used to build a discrete cumulative density function by alternative. Afterwards, a random number Uniform (0,1) was generated for each observation. Finally, the chosen alternative was determined as the inverse of the cumulative density function, evaluated for each random number.

The sampling protocol used to draw alternatives 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 $\tilde{J} = 5, 10, 15, 20$ and 25.

Under this setting we estimated the model using four different methods. The first model corresponds to a *Truncated* version of the problem were only the elements in the subset D_n are used to built the term $\hat{R}_{in} = \sum_{\substack{j \neq i \\ j \in D_n}} \ln\left(1 + \exp\left(\beta\left(x_{jn} - x_{in}\right)\right)\right)$.

The second method is termed *True Probabilities*. In this case, the true probabilities, which are known in this Monte Carlo Experiment, are used to build the expansion factors, as shown in Eq. (8). This estimator is not practical since in reality choice probabilities are unknown, but is reported to be compared with its approximated versions *Pop.Shares* and *1_0*, which were described in Section 3.3.

Finally, we considered the *Re-samplin,* 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 model was generated 100 times, and the estimation methods were applied considering different values for $\tilde{J}$. For each model estimated we report the following statistics to assess the efficacy and efficiency of each method in estimating the model coefficients.

Bias: Difference between average estimator and the true value of each respective parameter. The smaller the Bias, the better is the method in terms of small sample efficacy in recovering the true values of the model.

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 estimators. This statistical test can be used to test the null hypothesis that 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.

Table 1 reports the results for the *Truncated* model. It can be noted that the results are remarkably poor with the truncated model, which is to be expected since it neglects the fact that the random regret function gets truncated because of the sampling of alternatives. Even for $\tilde{J}$ as large as 25, the bias is around 47%, the t-test that the mean bias is zero is rejected with very large confidence, and there is not even one realization for which the estimator is within a 75% confidence interval.

Table 1 Assessment of Estimators Obtained with the Truncated Model

	Truncated			
$\tilde{J}$	**Bias**	**RMSE**	**t-test**	**Count**
5	5.933	35.331	16.733	0
10	4.045	16.432	15.484	0
15	2.463	6.099	13.203	0
20	1.256	1.597	9.244	0
25	0.477	0.234	5.864	0

Table 2 reports the results of the proposed methodology when the same choice set is used to calculate the sampling correction and to build the expansion of the truncated random regret model. First is shown the model considering the true probabilities, and then the two approximations that are feasible with real data, *methods 1_0* and *Population shares*.

Table 2 Assessment of Methods When Resampling is not Possible

	True Probabilities				1_0				Population Shares			
$\tilde{J}$	**Bias**	**RMSE**	**t-test**	**Count**	**Bias**	**RMSE**	**t-test**	**Count**	**Bias**	**RMSE**	**t-test**	**Count**
5	0.403	0.170	4.687	0	3.116	9.800	10.207	0	-0.419	0.177	13.249	0
10	0.180	0.038	2.473	5	0.587	0.358	5.129	0	-0.222	0.051	5.491	0
15	0.121	0.018	2.069	17	0.235	0.059	3.534	0	-0.121	0.016	3.009	2
20	0.078	0.009	1.519	36	0.105	0.014	1.968	20	-0.058	0.005	1.401	37
25	0.045	0.004	0.912	52	0.045	0.004	0.917	51	-0.016	0.002	0.347	74

Results show that the method proposed performs substantially better than the truncated model for all values of $\tilde{J}$. For the unfeasible *True Probabilities* model, the bias is below 10% from a sample of 20 alternatives. Concordantly, for the same number of alternatives sampled, the t-test that the mean estimator is equal to its true value cannot be rejected at a 95% level of significance. Equivalently, the count for this value of $\tilde{J}$ is 36. As it might be expected, for relatively small values of $\tilde{J}$ the statistics of the feasible methods *1_0* and *Population Shares*, are somehow below to those obtained with the *True Probabilities* method, which they are approximating. However, their values appear to be reasonably good. When $\tilde{J}$ becomes as large as 20, the *Population Shares* method even appears to perform better than the *True Probabilities* model on all criteria.

Additionally, Table 2 shows that the consistency of the method proposed depends on the value of $\tilde{J}$. Although the estimator of the regret function is consistent and unbiased, the fact that log-likelihood function is nonlinear, implies that for a given $\tilde{J}$ there is going to be a bias that will never disappear, even if *N* goes to infinity. In practice, this implies that the researcher should test the stability of the estimators of the model parameters as a function of $\tilde{J}$. If the estimators for different values of $\tilde{J}$ are statistically equal, one can be confident that the finite sample bias due sampling of alternatives is negligible. Otherwise, $\tilde{J}$ should be increased until attaining stability. This is equivalent to the need for testing for the stability of Logit Mixture's estimators as a function of the number of draws, in the simulated maximum-likelihood framework (Walker, 2001).

Finally, Table 3 reports the results for the case when re-sampling is possible. Comparing the results of Table 3 with those from Table 1 and 2, it can be noted that, for this experimental setting, being able to resample a choice-set to expand the regret function results in better results. With the same data, the bias is now below 10% for $\tilde{J}$ equal to 15. The t-test is also below the 95% confidence critical value and the count is 44 for this smaller value of $\tilde{J}$.

Table 3 Assessment of Methods When Resampling is Possible

	ReSampling			
$\tilde{J}$	**Bias**	**RMSE**	**t-test**	**Count**
5	0.631	0.416	4.632	0
10	0.240	0.066	2.623	3
15	0.076	0.009	1.300	44
20	0.021	0.003	0.427	72
25	0.010	0.002	0.208	77

Finally, it should be recalled that the Monte Carlo results are by no means a full description of the small sample properties of the estimators, but only a partial description that is valid only for the examples analyzed, and should be understood as an illustration of the behavior of the models. This implies that the results showed here in which the resampling method outperformed the *method 1_0* and *Pop.Shares*, neither the bias and other statistics obtained, may be simply transferred to other applications. To explore general applicability of our results, further investigation is required, particularly regarding the utilization of real data.

5 Conclusion

This article proposes a novel 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 Random Regret Minimization models (RRM). In light of the fact that runtimes of RRM models increase more than linearly 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 is 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 (with *iid* errors), does not exhibit the IIA property, McFadden's classical 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 one developed by Guevara and Ben-Akiva (2010) 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 model parameters and in the estimated shares when no correction is applied. In addition, 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, two 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. 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.

Acknowledgments

Funding for this research came in part from Fondecyt, Chile, through grant N°11110131. All Monte Carlo and real data experiments were generated and/or estimated using the open-source software R (R Development Core Team, 2008). Support from The Netherlands Organization for Scientific Research (NWO), in the form of VENI-grant 451.10.001, is gratefully acknowledged by the second author.

References

Bekhor, S., Chorus, C.G., Toledo, T., 2012. A Stochastic User Equilibrium formulation for the Random Regret Minimization-based route choice model. *Transportation Research Record* (in press)

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.

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

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

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

Bierlaire M, 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.

Boeri, M., Longo, A., Doherty, E., Hynes, S., 2012. Site choices in recreational demand: A matter of utility maximization or regret minimization? *Journal of Environmental Economics and Policy*, **1**(1), 32-47

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., de Jong, G.C., 2011. Modeling experienced accessibility for utility-maximizers and regret-minimizers. *Journal of Transport Geography*, **19**, 1155-1162

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

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

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

Guevara C.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. and Ben-Akiva 2010. Sampling of Alternatives in Multivaraite Extreme Value (MEV) Models. World Conference in Transportation Research, Lisbon, Portugal.

Hensher, D.A., Greene, W.H., Chorus, C.G., 2012. Random Regret Minimization or Random Utility Maximization: An exploratory analysis in the context of automobile fuel choice. *Journal of Advanced Transportation* (in press).

Kaplan, S., Prato, C.G., 2012. The application of the random regret minimization model to drivers' choice of crash avoidance maneuvres. *Transportation Research Part F* (in press)

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

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

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

Thiene, M., Boeri, M., Chorus, C.G., 2012. Random Regret Minimization: Exploration of a new choice model for environmental and resource economics. *Environmental and Resource Economics*, **51**(3), 413-429

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

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