Transportation Research Part B 58 (2013) 185–198

Contents lists available at ScienceDirect

Transportation Research Part B

journal homepage: www.elsevier.com/locate/trb

Sampling of alternatives in Logit Mixture models

C. Angelo Guevara [a,*], Moshe E. Ben-Akiva [b]

[a] Facultad de Ingeniería y Ciencias Aplicadas, Universidad de los Andes, Monseñor Álvaro del Portillo N°12.455, Las Condes, Chile
[b] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

ARTICLE INFO

Keywords:
Sampling of alternatives
Logit Mixture
Discrete choice

ABSTRACT

Employing a strategy of sampling of alternatives is necessary for various transportation models that have to deal with large choice-sets. In this article, we propose a method to obtain consistent, asymptotically normal and relatively efficient estimators for Logit Mixture models while sampling alternatives. Our method is an extension of previous results for Logit and MEV models. We show that the practical application of the proposed method for Logit Mixture can result in a *Naïve* approach, in which the kernel is replaced by the usual sampling correction for Logit. We give theoretical support for previous applications of the *Naïve* approach, showing not only that it yields consistent estimators, but also providing its asymptotic distribution for proper hypothesis testing. We illustrate the proposed method using Monte Carlo experimentation and real data. Results provide further evidence that the *Naïve* approach is suitable and practical. The article concludes by summarizing the findings of this research, assessing their potential impact, and suggesting extensions of the research in this area.

1. Introduction

Employing a strategy of sampling of alternatives is frequent in various transportation models that deal with large choice sets. This is the case, for example, of models of residential location, route choice, or activity-based models. Sampling of alternatives in such models may be needed, for example, to deal with the computational burden of managing large choice-sets, or because it is too costly to measure the attributes of all the alternatives. The problem of estimating a choice model with a sample of alternatives was resolved for the Logit model by McFadden (1978). However, the assumptions of the Logit are too restrictive for various choice problems, motivating the extension of McFadden's (1978) result to more flexible models.

The first extension of McFadden's (1978) result to non-Logit models was developed by Guevara and Ben-Akiva (2013), who studied the problem of sampling of alternatives in Multivariate Extreme Value (MEV) models. The MEV is a family of models that include the Logit and other closed-form models that allow some level of correlation among alternatives, such as the Nested Logit and the Cross Nested Logit. The method proposed by Guevara and Ben-Akiva (2013) achieves consistency, asymptotic normality and relative efficiency while sampling alternatives, and is based on the expansion of a term that gets truncated because of the sampling. If the researcher can sample a different set to perform the expansion of the truncated term, the method can be applied directly. In turn, when the researcher cannot re-sample, assumptions about the choice probabilities are needed. Guevara and Ben-Akiva (2013) illustrate the application of the method to different Monte Carlo experiments and to real data. Results show that the method is practical and yields acceptable results, even for relatively small sample sizes.

* Corresponding author. Tel.: +56 022 618 1364; fax: +56 022 618 1642.
E-mail addresses: caguevara@miuandes.cl (C.A. Guevara), mba@mit.edu (M.E. Ben-Akiva).

0191-2615/$ - see front matter
http://dx.doi.org/10.1016/j.trb.2013.08.011

A second extension of McFadden's (1978) result to non-Logit models was developed by Guevara et al. (in preparation), who studied the problem of sampling of alternatives in Random Regret Minimization (RRM) models. In RRM models, every alternative is compared to all other alternatives to build what is called a regret function (Chorus, 2010). Evaluation of this function quickly becomes difficult as the number of alternatives grows, motivating the need for sampling of alternatives. Guevara et al. (in preparation) show that a method, based on the expansion of the terms that get truncated because of the sampling, yields consistent, asymptotically normal and relatively efficient estimates. The authors also provide evidence showing that the method is practical and yields acceptable results for finite samples.

The extension of McFadden's (1978) result to Logit Mixture models is relevant because the Logit Mixture is fully flexible, in the sense that it can approximate any random utility model (McFadden and Train, 2000). The first studies in the area are the articles by McConnel and Tseng (2000) and Nerella and Bhat (2004), who provided Monte Carlo evidence suggesting the suitability of a *Naïve* approach for sampling of alternatives in a random coefficients model. In this *Naïve* approach the kernel of the model is simply replaced by McFadden's (1978) correction for Logit, ignoring the fact that the IIA assumption is formally broken in Logit Mixture models. A similar empirical result is reported by Azaiez, 2010, who additionally showed that the *Naïve* approach seemed to do better than an approximated method inspired in the approach used by Guevara and Ben-Akiva (2013) for MEV models. Lemp and Kockelman (2012) also provide evidence suggesting that the *Naïve* approach is suitable, but that its empirical efficiency depends on the sampling protocol considered. Seemingly contradicting all previous results, Chen et al. (2005) provide empirical evidence suggesting that the *Naïve* approach is not suitable for an error components Logit Mixture model. Finally, von Haefen and Domanski (2013) studied the problem of sampling of alternatives in a latent-class model, which is a special case of a Logit Mixture. The authors demonstrate how the expectation–maximization (EM) algorithm (see, e.g., Train, 2009) can be used with the Logit sampling correction to generate consistent estimates.

In this article, we study the conditions needed to achieve consistency, asymptotic normality and relative efficiency while sampling alternatives in Logit Mixture models. Our methodology can be seen as an extension of McFadden's (1978) result for Logit, and it builds on the methodology proposed by Guevara and Ben-Akiva (2013). We show that the proposed method in practice can be applied in three ways, one of which is the previously considered *Naïve* approach. We use Monte Carlo experimentation and real data to illustrate the different versions of the proposed method and to shed light on their finite properties.

The article is structured in seven sections. Following this introduction, Section 2 shows that the conditional probability of choosing an alternative, given that a certain subset was drawn, can be written as a Logit Mixture model in which the kernel is the product of two terms, one of which is McFadden's result for sampling of alternatives in Logit. It is then shown that the maximization of a quasi-log likelihood function based on this expression results in consistent estimators of the model parameters. However, this estimator is not practical since it still has a term that depends on the full choice-set. Next, in Section 3 we show that a proper approximation of the unfeasible term will result in consistent, asymptotically normal and relatively efficient estimators of the model parameters. In Section 4 we describe three possible methods to develop the approximation in practice. In Section 5 we provide Monte Carlo evidence to illustrate the application and to assess the performance of the estimators using the three methods. Section 6 reports the application of the method to real data on residential location and Section 7 summarizes the main conclusions, implications, and potential extensions of this research.

2. Consistency of an estimator for Logit Mixture models, conditional on a sampled choice set

In this section we show that the maximization of a modified log-likelihood function allows the consistent estimation of the parameters of a Logit Mixture model under sampling of alternatives. The derivation will be constructed as an extension of McFadden's (1978) result on sampling of alternatives for the Logit model. It will result in an impractical method that will still depend partially on the full choice set. This limitation will be addressed later in sections 3 and 4 by constructing feasible estimators inspired in the methods proposed by Guevara and Ben-Akiva (2013) for sampling on MEV models.

Consider the problem of modeling the probability that an agent n will choose an alternative i within the J_n elements of the set C_n. Agents are assumed to be rational, in the sense that they choose the alternative from which they retrieve the largest utility U_{in}, which is assumed to be composed by a systematic part V and a random part ε, as shown in:

$$U_{in} = V_{in} + \varepsilon_{in} = V(x_{in}, \beta_n) + \varepsilon_{in}. \tag{1}$$

The systematic part of the utility depends, usually linearly, on attributes x_{in} with a vector of coefficients β_n, which can be interpreted as agents' taste for the attributes. Taste is assumed to be heterogeneous among agents but, and without loss of generality, generic among alternatives.

If it is assumed that ε is distributed *iid* Extreme Value $(0, \mu)$, the probability that agent n will choose alternative i, given β_n, will correspond to the Logit model shown in:

$$L_n(i|\beta_n, x_n, C_n) = \frac{e^{\mu V(x_{in}, \beta_n)}}{\sum_{j \in C_n} e^{\mu V(x_{jn}, \beta_n)}}, \tag{2}$$

where μ is the scale of the distribution of the error terms. For identification, μ is normalized to equal 1.

However, the researcher does not know each β_n, only their density over a set of parameters θ. For example, if β follows a *Normal* distribution, θ will be the mean and variance of β. Therefore, the researcher can only specify the following expression for the probability that agent n will choose alternative i.

$$P_n(i|\theta, x_n, C_n) = \int L_n(i|\beta, x_n, C_n) f(\beta|\theta) d\beta = \int \frac{e^{V(x_{in},\beta)}}{\sum_{j\in C_n} e^{V(x_{jn},\beta)}} f(\beta|\theta) d\beta. \quad (3)$$

Consider now that the true choice set C_n is too large to be practical for estimation and that the researcher needs to sample a subset D_n with $\tilde{J}_n$ elements. This may be needed, for example, to reduce the computational burden or because it is too costly to identify all the alternatives. D_n must include the chosen alternative i. Otherwise, the model could not be estimated. Various sampling protocols can be used to build the set D_n. For example, one could first draw the chosen alternative and then sample a given number of additional alternatives with a fixed probability, or by importance sampling.

Define $\pi(i,D_n|\beta_n,x_n)$ as the conditional probability that agent n will choose alternative i and that the researcher will sample the set D_n, given the coefficients β_n and attributes x_n. By the Bayes theorem, this conditional probability can be rewritten as shown in:

$$\pi(i, D_n|\beta_n, x_n) = \pi(D_n|i, x_n) L_n(i|\beta_n, x_n, C_n) = \pi(i|\beta_n, x_n, D_n)\pi(D_n|\beta_n, x_n), \quad (4)$$

$\pi(i|\beta_n,x_n,D_n)$ in Eq. (4) is the conditional probability that the agent would choose alternative i, given that the set D_n was constructed by the researcher. $\pi(D_n|i,x_n)$ is the conditional probability that the researcher would construct the set D_n, given that alternative i was chosen by the agent.

Note that it is considered in Eq. (4) that $\pi(D_n|\beta_n,i,x_n) = \pi(D_n|i,x_n)$ because, after conditioning on i, it is assumed that the way the other elements on D_n are drawn does not depend on β_n. This assumption is not essential and can be generalized, but it is representative of most strategies for sampling of alternatives that can be applied in practice. For some special cases, this term may become even further simplified. For example, if the sampling protocol used was to draw the chosen alternative i and then to draw $\tilde{J}_n - 1$ alternatives with a fixed probability, then $\pi(D_n|i,x_n) = \pi(D_n|i)$. Furthermore, if that fixed probability is independent across alternatives $\pi(D_n|i) = \pi(D_n|j) \quad \forall j \in C_n$. In general, the conditional probability $\pi(D_n|i,x_n)$ will depend on x_n and will not be equal across alternatives.

Considering that the events of choosing each alternative in C_n are mutually exclusive and totally exhaustive, $\pi(D_n|\beta_n,x_n)$ can be rewritten using the Total Probability theorem (see, e.g., Bertsekas and Tsitsiklis, 2002) as shown in:

$$\begin{aligned}\pi(D_n|\beta_n, x_n) &= \sum_{j\in C_n} \pi(D_n|j, x_n) L_n(j|\beta_n, x_n, C_n) \\ &= \sum_{j\in D_n} \pi(D_n|j, x_n) L_n(j|\beta_n, x_n, C_n)\end{aligned}, \quad (5)$$

where the second equality in Eq. (5) holds because $\pi(D_n|j,x_n) = 0 \forall j \notin D_n$ since D_n must include the chosen alternative.

Combining Eq. (4) and Eq. (5), the following expression for the conditional choice probability is obtained

$$\pi(i|\beta_n, x_n, D_n) = \frac{\pi(D_n|i, x_n) L(i|\beta_n, x_n, C_n)}{\sum_{j\in D_n} \pi(D_n|j, x_n) L(j|\beta_n, x_n, C_n)} = \frac{e^{V(x_{in},\beta_n)+\ln \pi(D_n|i,x_n)}}{\sum_{j\in D_n} e^{V(x_{jn},\beta_n)+\ln \pi(D_n|j,x_n)}}, \quad (6)$$

where $\ln \pi(D_n|j,x_n)$ is termed the sampling correction.

Eq. (6) indicates that the conditional probability of choosing alternative i, given that a particular choice-set D_n was constructed, depends only on the alternatives in D_n. This is a consequence of the Independence of Irrelevant Alternatives (IIA) property, expressed in this case in the cancellation of the denominators when dividing the probabilities of two alternatives in the Logit kernel. This result holds because, given β_n, the model is a Logit.

McFadden (1978) demonstrated that the maximization of a quasi-log likelihood, constructed using the conditional choice probabilities shown in Eq. (6), yields consistent estimators if the Logit model has fixed coefficients, that is, if $\beta_n = \beta \forall n$. In what follows, we will extend McFadden's (1978) result for the case when the coefficients are random.

Consider now that the researcher does not know each β_n, but only their density over a set of parameters θ. Then, using the Bayes' theorem, and conditioning conveniently to retrieve the result shown in Eq. (6), $\pi(i|\theta,x_n,D_n)$ can be written as follows

$$\begin{aligned}\pi(i|\theta, x_n, D_n) &= \frac{\pi(i,D_n|\theta,x_n)}{\pi(D_n|\theta,x_n)} = \frac{1}{\pi(D_n|\theta,x_n)} \int \pi(i, D_n|\beta, x_n) f(\beta|\theta) d\beta \\ &= \frac{1}{\pi(D_n|\theta,x_n)} \int \pi(D_n|\beta, x_n)\pi(i|\beta, x_n, D_n) f(\beta|\theta) d\beta \\ &= \int \left(\frac{\pi(D_n|\beta,x_n)}{\pi(D_n|\theta,x_n)}\right) \frac{e^{V(x_{in},\beta)+\ln \pi(D_n|i,x_n)}}{\sum_{j\in D_n} e^{V(x_{jn},\beta)+\ln \pi(D_n|j,x_n)}} f(\beta|\theta) d\beta\end{aligned}. \quad (7)$$

Then, defining

$$W_n = \frac{\pi(D_n|\beta, x_n)}{\pi(D_n|\theta, x_n)} = \frac{\sum_{j\in D_n} L_n(j|\beta, x_n, C_n)\pi(D_n|j, x_n)}{\sum_{j\in D_n} P_n(j|\theta, x_n, C_n)\pi(D_n|j, x_n)}, \quad (8)$$

we will show that one can obtain consistent estimators of the distribution parameters θ by maximizing the quasi-log-likelihood function shown in Eq. (9), where i corresponds to the alternative chosen by agent n.

$$QL_{LM,D} = \sum_{n=1}^{N} \ln \int W_n \frac{e^{V(x_{in},\beta)+\ln \pi(D_n|i,x_n)}}{\sum_{j\in D_n} e^{V(x_{jn},\beta)+\ln \pi(D_n|j,x_n)}} f(\beta|\theta)d\beta. \quad (9)$$

However, Eq. (9) is not practical for the problem of sampling of alternatives in Logit Mixture models. Although the sum in the denominator of the kernel depends only on the alternatives in the set D_n, the term W_n still depends on the full choice-set, as shown in Eq. (8). We will solve this limitation later in sections 3 and 4, but we first need to show that the maximization of the quasi-loglikelihood shown in Eq. (9), yields consistent estimators of the model parameters.

Maximizing Eq. (9) is the same as maximizing Eq. (9) times $1/N$, which is in turn a sample analog of the expected value $E()$ of the log-likelihood constructed using the conditional probabilities shown in Eq. (7) over the population

$$\frac{1}{N}\sum_{n=1}^{N} \ln \int W_n \frac{e^{V(x_{in},\beta)+\ln \pi(D_n|i,x_n)}}{\sum_{j\in D_n} e^{V(x_{jn},\beta)+\ln \pi(D_n|j,x_n)}} f(\beta|\theta)d\beta \approx E\left(\ln \int W \frac{e^{V(x_i,\beta)+\ln \pi(D|i,x)}}{\sum_{j\in D} e^{V(x_j,\beta)+\ln \pi(D|j,x)}} f(\beta|\theta)d\beta\right),$$

where x, D and W are random variables that take values x_n, D_n, and W_n respectively.

The expected value depends on the joint density $f(i,x,D|\theta^*)$, where θ^* corresponds to a vector of population parameters of the distribution of β. Then

$$E() = \int \{\ln \phi_i(\theta)\} f(i,x,D|\theta^*) didDdx, \quad (10)$$

where

$$\phi_i(\theta) \equiv \int W \frac{e^{V(x_i,\beta)+\ln \pi(D|i,x)}}{\sum_{j\in D} e^{V(x_j,\beta)+\ln \pi(D|j,x)}} f(\beta|\theta)d\beta = \pi(i|\theta,x,D). \quad (11)$$

By the Bayes theorem we can re-write the joint density as $f(i,x,D|\theta^*) = \pi(i|\theta^*,x,D)\pi(D|\theta^*,x)f(x)$. To simplify the notation, we assume that the full choice-set C does not vary in the sample. This assumption is not essential and can be generalized. Under these conditions, the integration of alternatives i will be over all the elements in C and the integration of subsets D will be over all possible subsets $D \subseteq C$. Therefore, Eq. (10) can be rewritten as follows:

$$E() = \int [\sum_{i\in C}\sum_{D\subseteq C} \{\ln(\phi_i(\theta))\pi(i|\theta^*,x,D)\pi(D|\theta^*,x)\}]f(x)dx. \quad (12)$$

Then, replacing $\phi_i(\theta^*) \equiv \pi(i|\theta^*,x,D)$, the following expression for the expectation is obtained

$$E() = \int \left[\sum_{D\subseteq C} \pi(D|\theta^*,x) \left\{\sum_{i\in C} (\phi_i(\theta^*) \ln \phi_i(\theta))\right\}\right] f(x)dx. \quad (13)$$

Note that the only part of $E()$ in Eq. (13) depending on the arguments θ has the form

$$\sum_{i\in C} \phi_i(\theta^*) \ln \phi_i(\theta). \quad (14)$$

This expression has a maximum at $\theta = \theta^*$ because

$$\frac{\partial}{\partial\theta}\left[\sum_{i\in C}\phi_i(\theta^*)\ln\phi_i(\theta)\right]_{\theta=\theta^*} = \sum_{i\in C}\phi_i(\theta^*)\frac{1}{\phi_i(\theta^*)}\frac{\partial\phi_i(\theta)}{\partial\theta}|_{\theta=\theta^*} = \sum_{i\in C}\frac{\partial\phi_i(\theta)}{\partial\theta}|_{\theta=\theta^*} = 0,$$

where the last equality holds because

$$\sum_{i\in C}\phi_i(\theta) = 1.$$

Then, under general regularity conditions, this maximum is unique and the maximum of Eq. (9) converges in probability to the maximum of the true likelihood, and therefore it yields consistent estimators of the model parameters (Newey and McFadden, 1986).

3. Asymptotic distribution of a feasible estimator for sampling of alternatives in Logit Mixture models

The quasi-loglikelihood shown in Eq. (9) is not practical because the term W_n depends on all the alternatives in the choice-set C_n. In this section we will show that if an approximation for W_n is properly constructed using the elements in D_n, one can achieve consistency, asymptotic normality and relative efficiency.

Consider that $\widehat{W}_n$ is an estimator of W_n that fulfils the following three requirements:

- $\widehat{W}_n$ is an unbiased estimator of W_n,

- $\widehat{W}_n$ is a consistent estimator of W_n as $\widetilde{J}_n$ grows,
- $\widehat{W}_n$ is feasible in the sense that it is constructed solely using alternatives in D_n.

Then, it can be shown that the maximization of the feasible quasi-log-likelihood function shown in Eq. (15)

$$QL_{LM,D}^{Feasible} = \sum_{n=1}^{N} \ln \int \widehat{W}_n \frac{e^{V(x_{in},\beta)+\ln \pi(D_n|i,x_n)}}{\sum_{j\in D_n} e^{V(x_{jn},\beta)+\ln \pi(D_n|j,x_n)}} f(\beta|\theta) d\beta, \tag{15}$$

provides, under general regularity conditions, consistent estimators $\hat{\theta}$ of the model parameters θ^*, as $\widetilde{J}_n$ grows with N at any rate. If $\widetilde{J}_n$ grows faster than $\sqrt{N}$, $\hat{\theta}$ will also be asymptotically normal, with the following parameters:

$$\hat{\theta} \overset{a}{\sim} Normal(\theta^*, \boldsymbol{\Omega}/N) = Normal(\theta^*, \mathbf{R}^{-1}\mathbf{M}\mathbf{R}^{-1}/N), \tag{16}$$

where $\mathbf{M} = Var\left(\frac{\partial \ln \phi(\theta^*)}{\partial \theta}\right)$, $\mathbf{R} = E\left(\frac{\partial^2 \ln \phi(\theta^*)}{\partial \theta \partial \theta'}\right)$, and $\phi()$ is defined as in Eq. (11).

$\boldsymbol{\Omega}$ is usually termed the "robust" or "sandwich" variance–covariance matrix (see, e.g., Train, 2009, p. 201). A feasible estimator for $\boldsymbol{\Omega}$ was proposed by Berndt et al. (1974).

Note that $\boldsymbol{\Omega}$ is also the variance–covariance matrix of the unfeasible estimator resulting from the maximization of Eq. (9). Therefore, it can be affirmed that the estimators resulting from the maximization of Eq. (15) will be relatively efficient, compared to any other estimator considering an approximation of W_n .

Finally, an additional condition has to be added when J_n is finite and the protocol to draw alternatives is sampling without replacement. Under those conditions, $\widetilde{J}_n$ cannot go to infinity with N to derive the asymptotic distribution. Instead, $\widetilde{J}_n$ would need to increase only up to $\widetilde{J}_n = J_n$ for achieving consistency, asymptotic normality, and relative efficiency, because then the feasible estimator in Eq. (15) would become equal to the one shown in Eq. (9).

The derivation of the asymptotic distribution in Eq. (15) is analog to the demonstration developed by Guevara and Ben-Akiva (2013, Appendix 1) for the feasible estimator while sampling alternatives in MEV models. Both demonstrations are also analog to the procedure used by Train (2009, pp. 247–257) to derive the asymptotic distribution of simulation-based estimators.

The demonstration consists of two stages. The first analyzes the asymptotic distribution of the sample average of the true score $g(\theta)$ and an approximation of it $\hat{g}(\theta)$. In the second stage this result is used to derive the asymptotic distribution of the estimators of the model parameters when considering the approximation.

The key difference between the derivation described by Guevara and Ben-Akiva (2013, Appendix 1) and the derivation needed in the case studied in this paper is in the definition of the score functions $g(\theta)$ and $\hat{g}(\theta)$. In this case it holds that

$$g(\theta) = \frac{1}{N}\sum_{n=1}^{N} \frac{\partial \ln \phi_n(\theta)}{\partial \theta} = \frac{1}{N}\sum_{n=1}^{N} \frac{\partial}{\partial \theta} \ln \int W_n \frac{e^{V(x_{in},\beta)+\ln \pi(D_n|i,x_n)}}{\sum_{j\in D_n} e^{V(x_{jn},\beta)+\ln \pi(D_n|j,x_n)}} f(\beta|\theta) d\beta. \tag{17}$$

for the unfeasible estimator, where i corresponds to the alternative chosen by agent n, and

$$\hat{g}(\theta) = \frac{1}{N}\sum_{n=1}^{N} \frac{\partial \ln \hat{\phi}_n(\theta)}{\partial \theta} = \frac{1}{N}\sum_{n=1}^{N} \frac{\partial}{\partial \theta} \ln \int \widehat{W}_n \frac{e^{V(x_{in},\beta)+\ln \pi(D_n|i,x_n)}}{\sum_{j\in D_n} e^{V(x_{jn},\beta)+\ln \pi(D_n|j,x_n)}} f(\beta|\theta) d\beta, \tag{18}$$

for the feasible approximation.The reader is referred to Guevara and Ben-Akiva (2013, Appendix 1) for further details on the derivation of the asymptotic distribution shown in Eq. (16).

4. Practical implementation of the feasible estimator for sampling of alternatives in Logit Mixture models

4.1. Introduction

The practical application of the proposed method for sampling of alternatives in Logit Mixture models requires constructing an estimator of W_n using solely the elements in the set D_n. In this section we propose and analyze three feasible estimators: *Population Shares*, *1_0* and *Naïve*. Then, in Section 5 we study their performance using Monte Carlo simulations.

Reconsider the term W_n in Eq. (8) with more detail.

$$W_n = \frac{\sum_{j\in D_n} L_n(j|\beta, x_n, C_n)\pi(D_n|j, x_n)}{\sum_{j\in D_n} P_n(j|\theta, x_n, C_n)\pi(D_n|j, x_n)} = \frac{\sum_{j\in D_n}\left\{\pi(D_n|j, x_n)\frac{e^{V(x_{in},\beta)}}{\sum_{j\in C_n} e^{V(x_{jn},\beta)}}\right\}}{\sum_{j\in D_n}\left\{\pi(D_n|j, x_n)\int \frac{e^{V(x_{in},\beta)}}{\sum_{j\in C_n} e^{V(x_{jn},\beta)}} f(\beta|\theta) d\beta\right\}}. \tag{19}$$

The problem is that both $L_n(j|\beta,x_n,C_n)$ and $P_n(j|\theta,x_n,C_n)$ depend on the full choice set C_n. One possible approach to avoid this limitation is to approximate the sums of exponentials by a term constructed from the expansion of the alternatives in D_n. This approach is inspired by the one used by Guevara and Ben-Akiva (2013) for the problem of sampling of alternatives in MEV models. Formally, the approximation proposed for W_n is

$$\widehat{W}_n = \frac{\sum_{j\in D_n}\left\{\pi(D_n|j,x_n)\frac{e^{V(x_{in},\beta)}}{\sum_{j\in D_n} w_{jn}e^{V(x_{jn},\beta)}}\right\}}{\sum_{j\in D_n}\left\{\pi(D_n|j,x_n)\int\frac{e^{V(x_{in},\beta)}}{\sum_{j\in D_n} w_{jn}e^{V(x_{jn},\beta)}}f(\beta|\theta)d\beta\right\}}, \tag{20}$$

where the w_{jn} are expansion factors. We need to define the expansion factors required to fulfill the assumptions needed for the validity of the approximation described in Section 3.

Consider first unbiasedness. Instead of finding proper expansion factors w_{jn} such that $\widehat{W}_n$ becomes an unbiased estimator of W_n, we can find expansion factors that would result in an unbiased estimator ($\widehat{B}_n$) of the sum of the exponentials (B_n) embedded both in $L_n(j|\beta,x_n,C_n)$ and $P_n(j|\theta,x_n,C_n)$, where

$$B_n = \sum_{j\in C_n} e^{V(x_{jn},\beta)}$$
$$\widehat{B}_n = \sum_{j\in D_n} w_{jn}e^{V(x_{jn},\beta)}. \tag{21}$$

Since W_n is continuous in the sum of the exponentials, the result described in Section 3 is applicable with a version of the score $g(\theta)$ and its approximation $\hat{g}(\theta)$, written explicitly as a function of the sum of the exponentials.

The expansion factors w_{jn} needed for obtaining an unbiased estimator of the sum of the exponentials B_n depend on the sampling protocol used to draw D_n. Consider, for example, that the sampling protocol is the following: draw the chosen alternative, and then to draw $\widetilde{J}-1$ alternatives randomly. In such a case, it can be shown (see, Guevara, 2010, Chapter 5) that the expansion factors needed to achieve unbiasedness will be those shown in:

$$w_{jn} = \frac{1}{P_n(j|\theta,x_n,C_n)+\frac{\widetilde{J}-1}{J-1}[1-P_n(j|\theta,x_n,C_n)]}. \tag{22}$$

For other sampling protocols, the expansion factors would have to be different from those shown in Eq. (22), but they will necessarily have to depend on the choice probabilities $P_n(j|\theta,x_n,C_n)$, because the chosen alternative must be included in D_n for estimation.

The second condition needed for the application of the result shown in Section 3 is consistency. Since the expansion factors shown in Eq. (21) result in unbiased estimators, to achieve consistency it suffices to note that the variance of the estimated sum of the exponentials necessarily decreases with $\widetilde{J}$.

The final condition is feasibility. The expansion factors shown in Eq. (22) still depend on the choice probabilities and, consequently, on the full choice set. Following the approach used by Guevara and Ben-Akiva (2013) for MEV, we explore two possibilities to approximate the choice probabilities at this stage.

4.2. Population shares method

The first alternative is to replace the individual choice probabilities in Eq. (22) by an estimation of the population share of the respective alternative. For example,

$$H_j = \frac{1}{N}\sum_{n=1}^{n} y_{jn}$$

could be a sample estimator of the population share of alternative j, where y_{jn} that takes value *1* if j is the alternative chosen by n, and zero otherwise.

If the sampling protocol used to build D_n is to draw the chosen alternative and then to draw $\widetilde{J}-1$ alternatives randomly, the *Population Shares* method will correspond to consider the following approximation for the term W_n:

$$\widehat{W}_n^{Pop.shares} = \frac{\sum_{j\in D_n}\left\{\pi(D_n|j,x_n)\frac{e^{V(x_{in},\beta)}}{\sum_{j\in D_n}\left[\frac{1}{H_j+\frac{\widetilde{J}-1}{J-1}(1-H_j)}\right]e^{V(x_{jn},\beta)}}\right\}}{\sum_{j\in D_n}\left\{\pi(D_n|j,x_n)\int\frac{e^{V(x_{in},\beta)}}{\sum_{j\in D_n}\left[\frac{1}{H_j+\frac{\widetilde{J}-1}{J-1}(1-H_j)}\right]e^{V(x_{jn},\beta)}}f(\beta|\theta)d\beta\right\}}. \tag{23}$$

In Section 5 we will use Monte Carlo simulation to study the performance of this method on finite samples and, in Section 6, we will implement it with real data. It can be hypothesized that, despite the asymptotical validity of the *Population Shares* method, its performance in finite samples may be challenged for three reasons. First, *Population Shares* requires making a rather rough approximation of the individual choice probabilities. Second, the method does not make directly an approximation over W_n, but on a component of it. Third, significant numerical difficulties may arise when evaluating the $\widetilde{J}$ integrals that are needed to calculate the denominator in Eq. (23).

4.3. 1_0 Method

The second alternative to develop a practical estimator is to replace the individual choice probabilities in Eq. (22) by y_{jn}, which takes value *1*, if j is the alternative chosen by n, and zero otherwise.

If the sampling protocol used to build D_n is to draw the chosen alternative and then to draw $\widetilde{J}-1$ alternatives randomly, the *1_0* method will correspond to consider the following approximation for the term W_n:

$$\widehat{W}_n^{1\_0} = \frac{\sum_{j\in D_n}\left\{\pi(D_n|j,x_n)\dfrac{e^{V(x_{in},\beta)}}{\sum_{j\in D_n}\left[\dfrac{1}{y_{jn}+\frac{\tilde{J}-1}{J-1}(1-y_{jn})}\right]e^{V(x_{jn},\beta)}}\right\}}{\sum_{j\in D_n}\left\{\pi(D_n|j,x_n)\int\dfrac{e^{V(x_{in},\beta)}}{\sum_{j\in D_n}\left[\dfrac{1}{y_{jn}+\frac{\tilde{J}-1}{J-1}(1-y_{jn})}\right]e^{V(x_{jn},\beta)}}f(\beta|\theta)d\beta\right\}}. \tag{24}$$

In Section 5 we will use Monte Carlo simulation to assess the performance of this method on finite samples. It can by hypothesized that the *1_0* method may potentially suffer the same limitations of the *Population Shares* method.

4.4. Naïve method

Another alternative to develop a feasible method is to approximate directly the choice probabilities $L_n(j|\beta,x_n,C_n)$ and $P_n(j|\theta,x_n,C_n)$ in Eq. (19). This approximation can be done in different ways, including using an estimator of the population share of each alternative, or by considering that the probability is equal to *1* for the chosen alternative and zero otherwise. However, there is a better option in this case.

Consider, for the moment, that the researcher knows the true $P_n(j|\theta,x_n,C_n)$, the population choice probability for each individual and alternative. If such information happens to be available, $P_n(j|\theta,x_n,C_n)$ could be directly replaced to calculate exactly the denominator in Eq. (19). Then, the only missing component would be the kernel choice probabilities $L_n(j|\beta,x_n,C_n)$ in the numerator of Eq. (19).

But, if each $P_n(j|\theta,x_n,C_n)$ is known, they would make a very good approximation of $L_n(j|\beta,x_n,C_n)$. Such an approximation is far better than using a flat estimator of the *Population Shares* for all n, or by approximating each probability to 1 when the alternative is chosen and to zero otherwise. Furthermore, if $P_n(j|\theta,x_n,C_n)$ is used to approximate $L_n(j|\beta,x_n,C_n)$ in Eq. (19) and the denominator is known, we will provide an unbiased and consistent estimator of W_n directly, instead of approximating the sums of the exponentials embedded in it.

The final step of the analysis is practicality. Interestingly, if the $P_n(j|\theta,x_n,C_n)$ are used to approximate $L_n(j|\beta,x_n,C_n)$ in the numerator of Eq. (19), then $\widehat{W}_n$ becomes exactly equal to one.

$$\widehat{W}_n^{Naïve} = \frac{\sum_{j\in D_n}P_n(j|\theta,x_n,C_n)\pi(D_n|j,x_n)}{\sum_{j\in D_n}P_n(j|\theta,x_n,C_n)\pi(D_n|j,x_n)} = 1. \tag{25}$$

This implies that there is no need to know the true $P_n(j|\theta,x_n,C_n)$, and therefore, the *Naïve* method is feasible. Also, there is no need to calculate multiple integrals, as was the case for the approximations used in Eq. (23) and Eq. (24). Furthermore, the *Naïve* method is independent from the sampling protocol, unlike the methods *Population Shares* and *1_0*.

For the approximation shown in Eq. (25), the likelihood of the model becomes the same as that of the *Naïve* approach considered by McConnel and Tseng (2000), Nerella and Bhat (2004), Chen et al. (2005), Azaiez, 2010 and Lemp and Kockelman (2012). Consequently, it can be affirmed that this research provides a formal theoretical support for the use of the *Naïve* approach for estimation and sampling of alternatives in Logit Mixture models. The research shows that the *Naïve* approach for sampling of alternatives in Logit Mixture models implicitly considers an approximation that achieves consistency, asymptotic normality and relative efficiency.

In Section 5 we will use Monte Carlo simulation to study the performance of this method on finite samples and, in Section 6, we will implement it with real data. It can be hypothesized that the *Naïve* method performs better than the *Population Shares* and the *1_0* methods for three reasons. First, because the approximation implicit in the *Naïve* approach should be more precise than those considered by the alternative methods. Second, the *Naïve* approach should have a weaker dependence on $\widetilde{J}$ since it acts directly on W_n instead of on the sum of the exponentials embedded on it. Third, the *Naïve* approach should have better computational properties since it does not require the calculation of additional integrals.

5. Monte Carlo experimentation

5.1. Introduction

In this section we develop three Monte Carlo experiments, in which we apply the different versions of the proposed method for estimation with sampling of alternatives in Logit Mixture models. The purpose of these experiments is twofold: to illustrate the application of the versions of the method, and to shed some light on their behavior in finite samples.

The first experiment is a random coefficients model, analog to those considered by McConnel and Tseng (2000), Nerella and Bhat (2004), Azaiez, 2010 and Lemp and Kockelman (2012). The second experiment is an error component model, analog to the one considered by Chen et al. (2005). The final experiment studies the impact of the noise of the data. In this case the variance of the systematic utility is varied, while keeping fixed the variance of the error term. This experiment is analog to one of the experiments developed by Lemp and Kockelman (2012).

All models were estimated using the Maximum Simulated Likelihood approach, considering 500 random draws. To avoid chatter (McFadden and Train, 2000), the same 500 draws were used throughout the estimation. Estimation was performed using the BFGS 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 2.80 GHz and 12 GiB RAM.

5.2. Random coefficients experiment

The true or underlying model in this Monte Carlo experiment is a random coefficients Logit with $N = 1000$ observations and $J = 1000$ alternatives for all observations ($C_n = C$). The systematic utility considers a single attribute that is distributed $Uniform(-2,1)$ for the first 500 alternatives and $Uniform(-1,2)$ for the second half. This uneven distribution across alternatives aims to build an experiment as simple as possible, while avoiding particular results that may arise from full symmetry. The specification also considers a generic linear taste coefficient β, which is distributed *Normal* with mean $\mu_\beta = 1.5$ and standard deviation $\sigma_\beta = 0.8$ across the 1000 observations. The random utility is completed by adding to the systematic utility an error term that is distributed Extreme Value with location parameter 0 and scale 1. The choice y_{jn} was defined by simulating the choice probability shown in Eq. (2), evaluated for each observation.

The experiment is completed by considering a particular sampling protocol to build subsets of alternatives D_n from the choice-set C, for each observation. This sampling protocol first draws the chosen alternative for each observation, and then samples $\tilde{J} - 1$ alternatives randomly without replacement. We repeated the experiment sampling with $\tilde{J} = 5$, 30 and 50 alternatives from the 1000 alternatives available for each observation.

The three versions of the method for estimation while sampling alternatives in Logit Mixture models were applied to the experiment. The first version of the method studied was *Population Shares*. In this case, W_n is approximated using Eq. (23). The second method considered was *1_0*. In this case, W_n is approximated using Eq. (24). The third method used is *Naïve*, in which W_n is approximated by 1, as shown in Eq. (25).

The experiment was repeated 100 times using different seeds for the generation of random variables. These repetitions were used to build four statistics to evaluate the efficacy and efficiency of each method in recovering the true values of the model parameters: the mean $\mu_\beta = 1.5$ and standard deviation $\sigma_\beta = 0.8$ of the random coefficient β.

Table 1 reports the summary statistics for the random coefficients model. The rows correspond to the results obtained with each method (*Population Shares*, *1_0* and *Naïve*) and the columns report the results for various sample sizes $\tilde{J} = 5, 30, 50$. The summary statistics reported are the following:

Bias: Calculated as the difference between the average estimator, across the 100 repetitions, and the true value of each parameter. A smaller bias implies better finite-sample efficacy.

Mean Squared Error (MSE): Calculated as the sum of the sampling variance and the square of the bias. A smaller MSE implies a more efficient method.

t-Test: Calculated as the ratio between the absolute value of the bias and the sampling standard deviation of the estimators. This statistic is used to test the null hypothesis that the mean of the sampling distribution is equal to its respective true value.

Table 1
Estimation results for the random coefficients experiment.

Method	$\tilde{J}$	5				30				50			
	Stat.	Bias	MSE	*t*-Test	Count	Bias	MSE	*t*-Test	Count	Bias	MSE	*t*-Test	Count
Pop. Shares	$\hat{\mu}_\beta$	−0.09318	0.01624	1.072	56	−0.06934	0.008114	1.206	42	−0.04073	0.005499	0.6572	62
	$\hat{\sigma}_\beta$	−0.08028	0.03379	0.4855	75	−0.02335	0.01010	0.2390	73	−0.02105	0.009764	0.2181	73
1_0	$\hat{\mu}_\beta$	0.1492	0.02966	1.736	28	0.09045	0.01299	1.305	43	0.04482	0.006670	0.6565	65
	$\hat{\sigma}_\beta$	1.690	2.888	9.446	0	0.2759	0.08595	2.783	4	0.1417	0.02924	1.479	35
Naïve	$\hat{\mu}_\beta$	−0.03998	0.01292	0.3757	73	−0.01762	0.004607	0.2688	70	−0.01379	0.004599	0.2077	70
	$\hat{\sigma}_\beta$	−0.1175	0.03949	0.7332	71	−0.03597	0.01037	0.3775	74	−0.02749	0.009926	0.2871	73

100 repetitions. $J = N = 1000$. Population parameters $\mu = 1.5$; $\sigma = 0.8$.

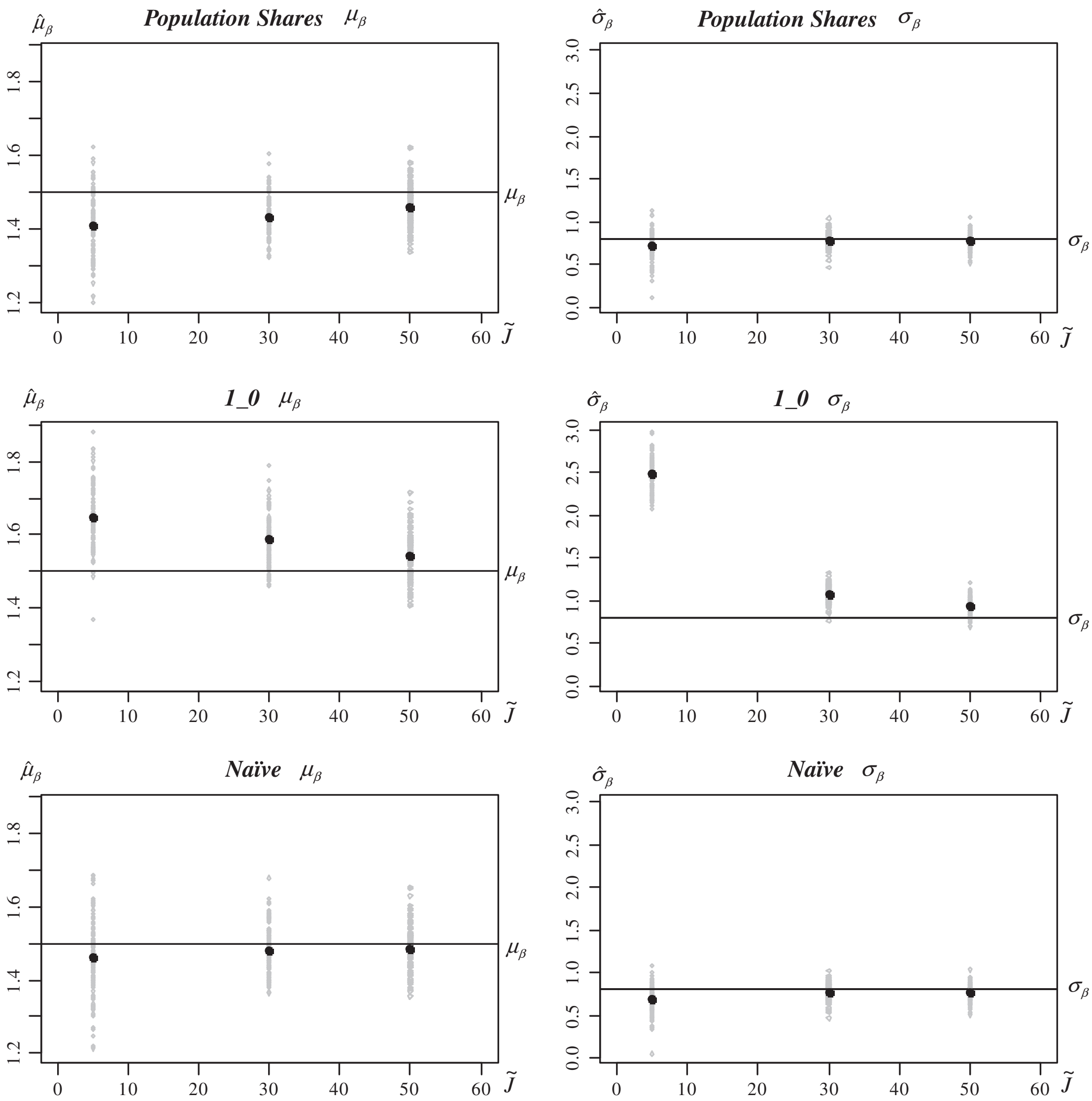

Fig. 1. Sampling distribution, random coefficients experiment.

Count: Calculated as the number of times the estimator of each repetition is within a 75% confidence interval of the true value. The interval is constructed using the sampling variance from the repetitions. This statistic is often called the empirical coverage and serves to assess the shape of the sampling distribution. The closer to 75 this statistic is, the closer its empirical distribution is to its theoretical sampling distribution.

Additionally, Fig. 1 portrays the detailed sampling distribution obtained from the 100 repetitions. The plots on the left report the results for the mean μ_β, and the plots on the right depict the results for the standard deviation σ_β. The two upper plots describe the results obtained with the *Population Shares* method. The plots in the middle illustrate the results of the method *1_0*, and the lower plots depict the results obtained with the *Naïve* method. The abscissa of each plot corresponds to the sample size $\tilde{J}$ and the ordinate corresponds to the respective estimator. For each $\tilde{J}$, the black large dot corresponds to the sample average of the respective estimator, and the grey smaller dots correspond to each one of the 100 repetitions. Finally, a horizontal line is drawn on each plot to remark the true value of the respective parameter.

In addition to making a relative comparison of the versions of the method, based on Fig. 1 and the statistics reported in Table 1, we will consider specific thresholds to asses the absolute suitability of each estimator. The *t*-test is required to be below $t_{5\%/2,100-1} = 1.984$. The Bias and MSE are required to be below 0.075 for μ_β and 0.04 for σ_β, which is equivalent to consider the Bias and the MSE to be below 5%, relative to the respective true value of each estimator. Finally, for the empirical coverage, the acceptable discordance with the nominal value will be defined as 10%, which corresponds to Counts above 67.

Table 1 and Fig. 1 show that results obtained using the method *1_0* are poor, largely dominated by *Population Shares* and *Naïve*, which are very similar among them. When only five alternatives are sampled for the *1_0* method, the *t*-statistic of σ_β is

Table 2
Estimation time [minutes] for the random coefficients experiment.

$\widetilde{J}$	5	30	50
Pop. Shares	3.183 (0.6540)	37.93 (12.17)	84.45 (16.90)
1_0	3.085 (0.6318)	42.66 (11.87)	97.34 (23.90)
Naïve	0.7682 (0.1615)	2.967 (0.5417)	4.962 (0.8914)

Average minutes for 100 repetitions. Standard error in parenthesis. 100 repetitions. $J = N = 1000$.

far greater than $t_{5\%/2,100-1} = 1.984$, and the *t*-statistic of μ_β is rejected with 90% confidence. Equivalently, the bias, MSE and Count are poor, particularly for σ_β. Notably, not even one realization of σ_β is within a 75% confidence interval when $\widetilde{J} = 5$. Statistics for the *1_0* method get better as $\widetilde{J}$ grows, but all are still clearly inferior to those of the *Population Shares* and *Naïve* methods. These results suggest that the impact of using the observed choice as a rough approximation of the choice probabilities to calculate the expansion factors in Eq. (22) was not suitable in this context.

Results for the *Population Shares* method are significantly better. First, the *t*-test is always below $t_{5\%/2,100-1} = 1.984$. The MSE is small. The threshold of 5% relative MSE is satisfied for all values of $\widetilde{J}$. The Bias is also well-behaved. The 5% threshold is satisfied when $\widetilde{J} = 30$ for μ_β and σ_β. The only potential limitation is with the empirical coverage, specially for the estimator of μ_β. When $\widetilde{J} = 5{,}30$, only about 50 out of 100 realizations are within a 75% confidence interval. The 10% threshold is satisfied for σ_β when $\widetilde{J} = 5$, but it is not satisfied for μ_β, even when $\widetilde{J} = 50$. These results suggest that using the *Population Shares* as a rough approximation of the choice probabilities to calculate the expansion factors in Eq. (23) yields somewhat acceptable results. However, relatively large sample sizes may be required for developing proper hypothesis testing with finite samples.

Finally, Table 1 and Fig. 1 show that the *Naïve* method behaves slightly better than the *Population Shares* method. In this case, all the criteria are fulfilled when $\widetilde{J} = 30$. When $\widetilde{J} = 5$ the results fail only the criterion for the Bias for σ_β. These results suggest that the approximation $L_n(j|\beta, x_n, C_n) \approx P_n(j|\theta, x_n, C_n)$ used to build the *Naïve* estimator shown in Eq. (25), yields good results, even with small sample sizes. This conclusion is consistent with previous results reported by McConnel and Tseng (2000), Nerella and Bhat (2004), Azaiez, 2010 and Lemp and Kockelman (2012).

In addition to evaluating efficacy and efficiency, it is also critical to assess the practicality of the different versions of the methods. The *Population Shares* and *1_0* versions require ad hoc implementations, while the *Naïve* method can be implemented directly in canned software like BIOGEME (Bierlaire, 2003) or ALOGIT (Daly, 1992). The *Population Shares* and *1_0* versions also require the evaluation of more integrals to calculate the denominators in Eqs. (23) and (24). This larger complexity may impact efficacy, efficiency, and estimation time.

Table 2 reports the average estimation time obtained from the application of the different methods for the 100 repetitions. Results show that estimation time for the *Naïve* method is smaller than the other methods for all values of $\widetilde{J}$ investigated. When $\widetilde{J} = 50$, the estimation time of the *Naïve* method is approximately 5 min, while it approaches or surpasses an hour and a half in the other methods. Results in Table 2 also show that the estimation time grows with $\widetilde{J}$, and that it does it at different rates, for each method. For the *Naïve* method, each additional $\widetilde{J}$ translates to approximately 0.1 additional minutes in estimation time for all the ranges of $\widetilde{J}$ studied. For the other methods, each additional $\widetilde{J}$ translates to approximately 1.4 min in estimation time, when $\widetilde{J}$ is between 5 and 30, and the rate grows to about 2.3 min when $\widetilde{J}$ is between 30 and 50. The differences in estimation time are explained by the need to calculate integrals to evaluate the terms $\widehat{W}_n$ for the *Population Shares* and the *1_0* methods, which are shown on Eq. (23) and Eq. (24), respectively.

These results suggest that both the *Naïve* and *Population Shares* methods yield acceptable results for a random coefficients model, but the former seems to be more robust. Furthermore, these results show that the *Naïve* approach is several times faster and easier to apply than the other methods.

5.3. Error components experiment

The true or underlying model in this Monte Carlo experiment is an error components Logit with $J = 1000$ alternatives and $N = 1000$ observations. As in the previous experiment, the systematic utility considers a single attribute that is distributed *Uniform*$(-2,1)$ for the first 500 alternatives and *Uniform*$(-1,2)$ for the second half. In this case, the specification considers a linear taste coefficient $\beta = 1.5$ that is fixed across observations. There is also an error component, shared by the alternatives between 501 and 1000, which is distributed *Normal* with mean zero and standard deviation $\sigma = 0.8$. This specification of the error components model allows for correlation among the alternatives between 501 and 1000. In this sense, it is similar, but not equal, to consider a Nested Logit model in which the second half of the alternatives belong to a nest (Walker et al., 2007).

Table 3 summarizes the statistics for the different methods. As in the random coefficients experiment, in this case the *1_0* method is significantly poor. Although the fit improves with $\widetilde{J}$, the criteria for the Bias, MSE, *t*-test and Count are far from being fulfilled, even for $\widetilde{J} = 50$, and are particularly poor for σ .

Table 3
Estimation results for the error components experiment.

Method	$\widetilde{J}$	5				30				50			
	Stat.	Bias	MSE	*t*-Test	Count	Bias	MSE	*t*-Test	Count	Bias	MSE	*t*-Test	Count
Pop. Shares	$\hat{\beta}$	−0.03754	0.005360	0.5973	67	−0.01536	0.003794	0.2575	70	−0.006116	0.003338	0.1065	71
	$\hat{\sigma}$	−0.4587	0.3266	1.345	42	−0.02442	0.1568	0.06180	73	−0.004522	0.1244	0.01282	80
1_0	$\hat{\beta}$	0.4013	0.1699	4.274	0	0.08332	0.01089	1.327	46	0.04662	0.005825	0.7714	62
	$\hat{\sigma}$	3.41	11.67	13.00	0	0.7568	0.6305	3.148	2	0.4260	0.2498	1.630	31
Naïve	$\hat{\beta}$	−0.007112	0.005242	0.09871	76	−0.008314	0.003655	0.1388	74	−0.004756	0.003469	0.08101	73
	$\hat{\sigma}$	−0.2210	0.1846	0.5997	71	−0.1020	0.1066	0.3288	81	−0.07183	0.1021	0.2307	78

100 repetitions. $J = N = 1000$. Population parameters $\beta = 1.5$; $\sigma = 0.8$.

The *Population Shares* method behaves much better in this experiment. From $\widetilde{J} = 30$ the relative Bias is below 5%, *t*-test is below $t_{5\%/2,100-1} = 1.984$ and relative Count is below 10%. The only criteria that is not met is that of the relative MSE, which is still 16% for $\widetilde{J} = 50$. However, this limitation in efficiency is not severe and of second importance, compared to the Bias.

The *Naïve* method also behaves acceptably well in this case, but it is slightly inferior to the *Population Shares* method. The main limitation is that the relative Bias, the most relevant measure, is still 9% for σ when $\widetilde{J} = 50$. These results cannot be catalogued as severe, but suggest that there is a qualitative difference with the results obtained with the *Population Shares* method.

The finding that, in this case, the standard deviations seem to be harder to estimate with the *Naïve* method, is concordant, to some extent, with the results reported by Chen et al. (2005). However, in our case we find that the problem is much less severe. This discrepancy in the results might be explained by the fact that Chen et al. (2005) considered only a single realization of the model, instead of repeating the experiment various times. Consequently, it could be the case that the particular realization considered by Chen et al. (2005) may have been especially poor by chance.

These results suggest that although the *Naïve* method yields acceptable results for an error components model, the *Population Shares* method is more robust in this context, at least regarding the Bias and the empirical coverage of the mean.

5.4. Noise variation

The final experiment assesses the impact of the variation of the relative noise of the model. This experiment is constructed by varying the Random Coefficients experiment described in Section 5.2 for $\widetilde{J} = 5$. The only component that is changed in this case is the variance of the single attribute *x* of the systematic utility. This affects the relative noise of the model because the scale of the Extreme Value error is maintained for all cases.

The first case corresponds to a relatively *Large Noise*. This is achieved by reducing the variance of the attribute, compared to the reference case described in Section 5.2. The attribute in this case is generated *Uniform*$(-1.5, 0.5)$ for the first 500 alternatives and *Uniform*$(-0.5, 1.5)$ for the second half. Consequently, the variance of the attribute for all alternatives and observations is in this case 0.08333.

The second case is *Middle Noise*, which corresponds to the model described in Section 5.2. In this case, the attribute is generated *Uniform*$(-2, 1)$ for the first 500 alternatives and *Uniform*$(-1, 2)$ for the second half. Consequently, the variance of the attribute for all alternatives and observations in this case is 0.7500.

The third case corresponds to *Small Noise*. In this case, the attribute is generated *Uniform*$(-3, 2)$ for the first 500 alternatives and *Uniform*$(-2, 5)$ for the second half. Consequently, the variance of the attribute for all alternatives and observations in this case is 2.083.

The results are reported in Table 4. As with the previous experiment, method *1_0* is inferior and the *Population Shares* and *Naïve* methods are similarly well behaved. The impact of the variation of the noise, both for *Population Shares* and *Naïve*

Table 4
Estimation results for the noise variation $\widetilde{J} = 5$.

Method	Stat.	Large noise				Middle noise				Small noise			
		Bias	MSE	*t*-Test	Count	Bias	MSE	*t*-Test	Count	Bias	MSE	*t*-Test	Count
Pop.		*Shares*	$\hat{\mu}_\beta$			−0.07707	0.013585	0.8814	58	−0.09318	0.016238	1.0719	56
										−0.093290	0.016386	1.0643	53
	$\hat{\sigma}_\beta$	−0.0931	0.0550	0.433	75	−0.08028	0.0338	0.48553	75	−0.053851	0.0148	0.49282	68
1_0	$\hat{\mu}_\beta$	0.2513	0.0766	2.169	15	0.14922	0.02966	1.736	28	0.06330	0.010431	0.7898	59
	$\hat{\sigma}_\beta$	2.21	4.91	11.59	0	1.6899	2.8877	9.446	0	1.1520	1.3444	8.734	0
Naïve	$\hat{\mu}_\beta$	−0.01265	0.012191	0.1153	75	−0.03998	0.012923	0.3757	73	−0.050481	0.012550	0.5048	65
	$\hat{\sigma}_\beta$	−0.1240	0.0655	0.554	79	−0.11750	0.0395	0.73324	71	−0.088492	0.0196	0.81627	56

100 repetitions. $J = N = 1000$. Population parameters $\mu = 1.5$; $\sigma = 0.8$; $\widetilde{J} = 5$.

methods, is equally small. In general, as the relative noise decreases, the Bias of the mean μ_β slightly increases and the Bias of the standard deviation σ_β slightly decreases. As the relative noise decreases, MSE of σ_β slightly decreases, and it generally increases for μ_β. Finally, the shape of the sampling distribution becomes slightly worse as the noise decreases. This can be noted in that, as the noise decreases, the Count differs more from its nominal value 75, and the *t*-tests become larger.

These results suggest that the variation of the noise had little or no impact in the relative efficiency or efficacy among the different versions of the methods. This might be explained by that, differently from Lemp and Kockelman (2012), the methods that we study do not differ in their use of the attributes of the model.

6. Application to real data

The section applies the method to real data. For this application, we revisit the Lisbon database of residential location used by Guevara and Ben-Akiva (2013). As with the Monte Carlo experiment, the purpose of this application to real data is twofold: to illustrate the implementation of the proposed method and to shed light on its finite-sample properties.

The database consists of 63 observations of chosen dwellings among a choice-set of 11,501 residences. The dwellings belong to the municipalities of Lisbon, Odivelas and Amadora, all within the Lisbon Metropolitan Area in Portugal. Details on the construction of this database can be found in Guevara, 2010.

Guevara and Ben-Akiva (2013) used this database to estimate a Nested Logit model considering a single nest for the 3483 dwellings that belonged to the municipalities of Odivelas and Amadora. In the present article we develop an analog model in which we capture the correlation among the dwellings from those two municipalities by considering an error component term that is shared only by those alternatives.

The systematic utility of the model depends linearly on four attributes: (1) dwelling price, which interacts with three income levels; (2) the distance to head-of-the-household's workplace; (3) the logarithm of the dwelling area; and (4) the logarithm of dwelling age +1. Because of the omission of quality attributes that are likely to be correlated with dwelling's price, we correct for endogeneity using the control-function method. As in Guevara and Ben-Akiva (2012, 2013), we use, as instrumental variables for price, the average price of similar dwellings beyond 500 m and within 5 km.

We first estimate the *Full Model*, considering all 11,501 alternatives. Then we subsequently sample a different number of alternatives. The Full Model serves as a benchmark to compare the results obtained when sampling alternatives. The protocol used to sample alternatives was to draw first the chosen alternative and then to sample randomly up to complete $\widetilde{J}_{OA}$ from Odivelas–Amadora, and the same $\widetilde{J}_L$ from Lisbon. We considered the following set of sample sizes for $\widetilde{J}_{OA} = \widetilde{J}_L$: 5, 30, 50 and 100.

Table 5 reports the estimation results. We only report the results obtained with the *Population Shares* and the *Naïve* methods. The *1_0* results are not reported for the sake of space and because, just as in the Monte Carlo experiment, the *1_0* method showed the poorest results in this case study.

Table 5
Lisbon's error component model while sampling alternatives.

Variables	Full model	$\widetilde{J}_{OA}=\widetilde{J}_L=5$		$\widetilde{J}_{OA}=\widetilde{J}_L=30$		$\widetilde{J}_{OA}=\widetilde{J}_L=50$		$\widetilde{J}_{OA}=\widetilde{J}_L=100$	
		Pop. Shares	*Naïve*	*Pop. Shares*	*Naïve*	*Pop. Shares*	*Naïve*	*Pop. Shares*	*Naïve*
Dwelling price (in 100,000 €)	−2.935	−2.655	−2.655	−3.082	−3.049	−3.062	−3.034	−2.962	−2.952
	(1.246)	(0.7256)	(0.7256)	(1.278)	(1.286)	(1.508)	(1.521)	(1.307)	(1.297)
Dwelling. price * [Inc. > 2,000 €/M]	0.8877	1.036	1.036 (0.5705)	0.8796	0.8698	0.8719	0.8655	0.8572	0.8541
	(0.7363)	(0.5705)		(0.7606)	(0.7611)	(0.7774)	(0.7811)	(0.7928)	(0.7894)
Dwelling price * [Inc. > 5,000 €/M]	0.8138	0.7483	0.7483	0.7737	0.7737	0.9017	0.8992	0.8426	0.8421
	(0.3049)	(0.3387)	(0.3387)	(0.3723)	(0.3691)	(0.3871)	(0.3841)	(0.2887)	(0.2880)
Distance to workplace (km)	−0.2715	−0.2263	−0.2263	−0.2927	−0.2897	−0.2785	−0.2765	−0.2698	−0.2690
	(0.1008)	(0.05274)	(0.05274)	(0.1015)	(0.1064)	(0.09912)	(0.1046)	(0.09271)	(0.09316)
Log (dwelling area [m^2])	2.380	1.652	1.652 (0.7329)	2.409	2.369	2.592	2.555	2.419	2.408
	(1.263)	(0.7329)		(1.278)	(1.291)	(1.549)	(1.566)	(1.307)	(1.298)
Log [dwell. age (years)+1]	−0.4740	−0.5302	−0.5302	−0.5236	−0.5189	−0.4880	−0.4845	−0.4905	−0.4893
	(0.1512)	(0.1303)	(0.1303)	(0.1771)	(0.1790)	(0.1915)	(0.1931)	(0.1666)	(0.1657)
$\hat{\delta}$ control-funct. Aux. variable	1.125	0.8708	0.8708	1.251	1.233	1.242	1.226	1.125	1.120
	(0.6511)	(0.3528)	(0.3528)	(0.6369)	(0.6414)	(0.7833)	(0.7899)	(0.6421)	(0.6381)
$\sigma_{O\text{-}A}$ Odiv.–Amadora	0.8629	1.956E−06	−1.276E−07	1.515	1.401	1.305	1.212	0.9355	0.9028
	(2.547)	(1.882E−04)	(1.337E−05)	(2.377)	(2.427)	(2.978)	(3.069)	(2.230)	(2.234)
Log-likelihood	−560.01	−110.46	−110.46	−218.32	−218.35	−250.20	−250.23	−294.25	−294.26
Est. time [hrs]	11.76	0.1156	0.01151	1.682	0.04099	4.850	0.08185	17.95	0.1655
Sample size *N*	63	63		63		63		63	
Choice-set size $\widetilde{J}$	11,501	10		60		100		200	

Error component for Amadora and Odivelas. Include sampling correction and correction for endogeneity with control-function method $\hat{\delta}$. Sample $\widetilde{J}_{OA}$ alts. from Odivelas–Amadora and $\widetilde{J}_L$ from Lisbon municipality. €/M: Euros per month. Robust standard errors in parenthesis. Inc.: monthly income.

The estimators of each parameter for the *Full Model* are shown in the first column of Table 5, together with the respective robust standard error in parenthesis. It can be noted that all the coefficients have the expected sign. Dwelling price is negative, but it becomes less negative as income grows. Distance to the workplace is negative, dwelling's area is positive and dwelling's age is negative. Also, the coefficient of the auxiliary variable $\hat{\delta}$ of the control function method is positive and significant. This implies that the model did suffered endogeneity, and that the omitted quality is positively correlated with price (Guevara, 2010). Finally, the estimator of σ_{OA} is 0.8629. This parameter accounts for the correlation among the alternatives in Odivelas and Amadora. This term has low significance, arguably because of the small number of observations available.

We now compare the estimators of the *Full Model*, with those obtained when sampling alternatives. We first consider the deviation of each estimator relative to the respective estimator obtained for the *Full Model*. When $\tilde{J}_{OA} = \tilde{J}_L = 5$, the relative deviation of the terms other than σ_{OA}, goes from 10% to 31% for both the *Naïve* and *Population Shares* methods, and it jumps to 100% for σ_{OA}. Results improve as $\tilde{J}_{OA} = \tilde{J}_L$ grow, but only when $\tilde{J}_{OA} = \tilde{J}_L = 100$, the estimators reach a more than acceptable level, compared with the estimators obtained with the *Full Model*. In that case the relative deviation of the terms other than σ_{OA}, goes from 0% to 4% both for the *Naïve* and the *Population Shares* method. Also, the relative deviation for σ_{OA} is 5% for the *Naïve* method, and 8% for the *Population Shares* method.

We conclude that, in this case study, the *Naïve* method is slightly better than the *Population Shares* method for all sample sizes, and that when $\tilde{J}_{OA} = \tilde{J}_L = 100$, the estimators obtained are reasonably close to those of the *Full Model*.

The conclusion is qualitatively the same when analyzing the deviation between the robust standard error obtained while sampling alternatives, and the robust standard error obtained with the *Full Model*. The *Naïve method* behaves slightly better than the *Population Shares* method for all sample sizes considered and, when $\tilde{J}_{OA} = \tilde{J}_L = 100$, the estimators of the standard errors are reasonably close to those of the *Full Model*, although this final statement means accepting deviations of up to 12%.

There are two points to note regarding the log-Likelihood achieved by each model. First, the log-likelihood is not comparable between models estimated with different $\tilde{J}$. Second, the differences in log-likelihood observed between the *Population Shares* and the *Naïve* methods are negligible, and possibly purely attributable to sampling or estimation error.

The final statistic for comparison is estimation time. Table 5 shows that the estimation time of the *Full Model* of this case study was about 12 h. When $\tilde{J}_{OA} = \tilde{J}_L = 5$, both methods take significantly less time. The *Naïve* method takes less than one minute and the *Population Shares* method takes about 7 min. Estimation time for the *Naïve* method grows almost linearly with $\tilde{J}$ at a rate of almost 3 s for each additional $\tilde{J}$. Consequently, even for $\tilde{J} = 100$, estimation time is under 10 min, and significantly shorter than that of the Full Model. The growth rate with $\tilde{J}$ of estimation time for the *Population Shares* method is more than linear. Indeed, the estimation time for $\tilde{J} = 100$ is about 18 h, almost 6 h more than the estimation time of the Full Model. As a result, the potential benefits of sampling of alternatives using the *Population Shares* method vanished, possibly because of the need for calculating additional integrals in the evaluation of Eq. (23).

7. Conclusion

This research provides a formal demonstration of the consistency, asymptotic normality and relative efficiency of a feasible estimator for sampling of alternatives in Logit Mixture models. This methodology is an extension of McFadden's (1978) result for Logit, and it builds on the methodology proposed by Guevara and Ben-Akiva (2013) for MEV models. This extension for Logit Mixture models is relevant because Logit Mixture is fully flexible, since it can approximate any random utility model (McFadden and Train, 2000).

We show that a feasible version of the proposed method is equivalent to consider a *Naïve* form of the conditional likelihood, in which the kernel of the Logit Mixture model is replaced by McFadden's (1978) sampling correction for Logit. It can be stated that the main contribution of this research is in the provision of theoretical support for previous empirical works suggesting the suitability of the *Naïve* approach. We show that the *Naïve* approach yields consistent estimators, and describe how to do proper hypothesis testing with it.

This investigation suggests that the *Naïve* approach should be preferred for estimation while sampling alternatives in Logit Mixture model because: (1) it yields consistent estimators; (2) it achieves relative asymptotic efficiency; (3) it can be applied in canned software; (4) proper hypothesis testing could be performed using robust standard errors, which are already available in canned software; (5) it is independent of the sampling protocol considered and does not require constructing an estimator for the choice probabilities; and (6) it is much faster than other feasible methods, with empirical evidence suggesting that its estimation time is linear in the choice-set sample size.

Three possible extensions can be identified for this research. The first line is to perform an analytical study of the finite-sample properties of the estimators investigated in this paper. This would help in identifying the conditions under which each method achieves better efficacy, efficiency, and estimation time.

A second line of research is in the practical and theoretical study of the sample sizes required for proper estimation. Preliminary evidence suggests that it is not a matter of the ratio $\tilde{J}/J$ as suggested by Nerella and Bhat (2004), but it possibly depends on the variance of the data and the error one is willing to accept for the estimators. Two practical strategies might be followed to this respect. The first would be to increase $\tilde{J}$ until the estimators have changes that are small enough, similar to the procedure suggested by Chiou and Walker (2007) for determining the proper number of draws for estimation of the Logit Mixture with maximum simulated likelihood. The second approach would be to study the error obtained by sampling multiple times for a given $\tilde{J}$.

A final line of research would be to study the performance of the proposed methods when the model is estimated using different procedures or under other contexts. Estimation methods that are potentially attractive to be analyzed are, for example, the MAMCL (Bhat, 2011), EM (Train, 2009), and Hierarchical Bayes (Train, 2009). Other contexts to be considered are, for example, panel data (Cherchi and Guevara, 2012a), or models with multiple random coefficients (Cherchi and Guevara, 2012b).

Acknowledgments

This publication was made possible in part by CONICYT, FONDECYT/Iniciación 11110131 and the Singapore National Research Foundation under the Future Urban Mobility research group of the Singapore-MIT Alliance for Research and Technology. All Monte Carlo and real data experiments were generated and/or estimated using the open-source software R (R Development Core Team, 2008). We would also like to thank the editorial contributions of Darinka Guevara and Katherine Rosa.

References

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

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. In: Proceedings of the 3rd Swiss Transportation Research Conference, Ascona, Switzerland.

Bhat, C.R., 2011. The maximum approximate composite marginal likelihood (MACML) estimation of multinomial probit-based unordered response choice models. Transportation Research Part B 45 (7), 923–939.

Chen, Y., Duann, L., Hu, W., 2005. The estimation of discrete choice models with large choice-set. Journal of the Eastern Asia Society for Transportation Studies 6, 1724–1739.

Cherchi, E., Guevara, C.A., 2012a. Maximum simulated likelihood and expectation–maximization methods to estimate random coefficients logit with panel data. Transportation Research Record: Journal of the Transportation Research Board 2302, 65–73.

Cherchi, E., Guevara, C.A., 2012b. A Monte Carlo experiment to analyze the curse of dimensionality in estimating random coefficients models with a full variance–covariance matrix. Transportation Research Part B 46 (2), 321–332.

Chiou, L., Walker, J.L., 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.

Daly, A., 1992. ALOGIT 3.2 User's Guide. Hague Consulting Group, The Hague.

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., 2013. Sampling of alternatives in multivariate extreme value (MEV) models. Transportation Research Part B 48, 31–52.

Guevara, C.A., Ben-Akiva, M., 2012. Change of scale and forecasting with the control-function method in logit models. Transportation Science 46 (3), 425–437.

Guevara, C.A., Chorus, C., Ben-Akiva, M.E., 2013. Sampling of Altenatives in Random Regret Minimization Models. Presented at the 92nd Transportation Research Board Annual Meeting, Washington, DC.

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: Karlquist, Lundqvist, Snickers, Weibull (Eds.), Spatial Interaction Theory and Residential Location. North Holland, Amsterdam, pp. 75–96.

Mcfadden, D., Train, K., 2000. Mixed MNL models for discrete response. Journal of applied Econometrics 15 (5), 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.

Newey, W., McFadden, D., 1986. Large sample estimation and hypothesis testing. In: Engle, McFadden (Eds.), Handbook of Econometrics, vol. 4 (36), pp. 2111–2245.

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, second ed. Cambridge University Press, New York, NY.

von Haefen, R.H., Domanski, A., 2013. Estimating Mixed Logit Models with Large Choice Sets. Working Paper. Department of Agricultural and Resource Economics, North Carolina State University, USA.

Walker, J., Ben-Akiva, M., Bolduc, D., 2007. Identification of parameters in normal error component logit-mixture (NECLM) models. Journal of Applied Econometrics 22 (6), 1095–1125.