# The estimation of Generalized Extreme Value models from choice-based samples

M. Bierlaire[*] D. Bolduc[†] D. McFadden[‡]

August 10, 2006

Report TRANSP-OR 060810
Transport and Mobility Laboratory
School of Architecture, Civil and Environmental Engineering
Ecole Polytechnique Fédérale de Lausanne
transp-or.epfl.ch

[*] Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne, Switzerland. Email: michel.bierlaire@epfl.ch

[†] Département d'économique, Université Laval, Québec

[‡] Econometrics Laboratory, University of California, Berkeley

## Abstract

We consider an estimation procedure for discrete choice models in general and Generalized Extreme Value (GEV) models in particular. It is based on a pseudo-likelihood function, generalizing the Conditional Maximum Likelihood (CML) estimator by Manski and McFadden (1981) and the Weighted Exogenous Sample Maximum Likelihood (WESML) estimator by Manski and Lerman (1977). We show that the property of Multinomial Logit (MNL) models, that consistent estimates of all parameters but the constants can be obtained from an Exogenous Sample Maximum Likelihood (ESML) estimation, does not hold in general for GEV models. We identify a specific class of GEV models with this desired property, and propose a new estimator for the more general case. This new estimator estimates the selection bias directly from the data. We illustrate the new estimator on pseudo-synthetic and real data.

# 1 Introduction

The estimation of discrete choice models is a very difficult task in the presence of selection bias, that is when the sampling strategy is based on the endogenous variable: the choice. These sampling techniques, also known as *choice-based*, are however commonly used in practice. In many cases, the analyst does not have a precise knowledge of the actual market shares in the population and therefore, cannot reflect them in the sampling scheme. Also, the analyst may want to oversample a specific alternative, in order to analyze a product with a small market share. Indeed, collecting sufficient data about such a product with a simple random sampling may require a prohibitively large sample size.

Various procedures have been considered in the literature. For instance, a Generalized Method of Moments for discrete choice models with choice-based sampling has been proposed by Imbens (1992). Nonparametric approaches are discussed by Morgenthaler and Vardi (1986) and Manski (1999), among others. We emphasize on the maximum likelihood paradigm because it is the most widely used setting for discrete choice modeling.

In the presence of selection bias, the Conditional Maximum Likelihood (CML) Estimation is known to produce consistent although inefficient estimators. As reported by Manski and Lerman (1977), McFadden has shown that CML estimation of the Multinomial Logit (MNL) model with a full set of Alternative Specific Constants (ASC) can equivalently be obtained from an Exogenous Sample Maximum Likelihood (ESML) estimation procedure after adjusting appropriately the constants.

In Sections 2 and 3, we summarize the issues of estimation under various sampling schemes, and propose an estimator based on a pseudo-likelihood functions, which generalizes the CML Estimator by Manski and McFadden (1981) and the Weighted Exogenous Sample Maximum Likelihood (WESML) Estimator by Manski and Lerman (1977). Generalized Extreme Value models are described in Section 4, where it is shown that ESML cannot be used as such for the estimation of non-MNL GEV models in the presence of selection bias, except for very specific instances. Such an instance is presented in Section 5. For the more general case, a new estimator for GEV models in the presence of selection bias is proposed in Section 6. Its performances are illustrated in Section 7 on pseudo-synthetic and real data.

## 2 Sampling

We are interested in the estimation of a discrete choice model where the choice set $\mathcal{C}$ is composed of $J$ alternatives. The independent, or exogenous variables are denoted by $x$, and are composed of socio-economic characteristics of the decision-maker as well as the attributes of the alternatives. The dependent, or endogenous variable, is discrete and represents the chosen alternative. We assume without loss of generality that the choice set of each individual is $\mathcal{C}$.

The discrete choice model gives the probability that a given alternative $i$ is selected, conditional to a choice context characterized by $x$:

$$\Pr(i|x,\theta) = P(i|x,\theta), \tag{1}$$

where $P$ is the choice model and $\theta$ is a vector of unknown parameters. The joint distribution of $(i,x)$ in the population is then given by

$$\Pr(i,x|\theta) = P(i|x,\theta)p(x), \tag{2}$$

where $p(x)$ is the proportion of the population with exogenous variable $x$.

We consider general stratified sampling strategies (Manski and McFadden, 1981) where the population is partitioned into G collectively exhaustive groups, defined in terms of combinations of both exogenous and endogenous variables. Individuals are randomly selected within each group. If $N_P$ is the total number of individuals in the population, and $N_s$ the total number of individuals in the sample, then a sampling strategy is characterized by $H_g$, $g = 1, \ldots, G$, the proportion of each group in the sample. In this case, the probability for a given individual belonging to group g to be in the sample is

$$r_g = \frac{H_g N_s}{W_g N_P} \tag{3}$$

where $W_g$ is the proportion of individuals in group g in the population.

If $\mathcal{C}_g$ is the set of alternatives relevant to group g, and $X_g$ the set of exogenous variables[1] relevant to group g, we have from (2)

$$W_g = \int_{x \in X_g} \sum_{i \in \mathcal{C}_g} \Pr(i, x|\theta)dx = \int_{x \in X_g} \sum_{i \in \mathcal{C}_g} P(i|x, \theta)p(x)dx, \tag{4}$$

and thus the probability $r_g$ depends on the unknown parameters $\theta$, which significantly complicates model estimation.

Sampling strategies can be classified into four categories.

1. A Simple Random Sampling (SRS) is obtained when only one group is considered, and every individual has the same probability to be selected. In this case, $H_g = W_g$ and $r_g = r = N_s/N_P$ is clearly independent from $\theta$.

2. An Exogenous Stratified Sampling (XSS) is obtained when the groups are characterized only by exogenous variables $x$. In this case, all alternatives are relevant to each group g, that is $\mathcal{C}_g = \mathcal{C}$ and

$$\begin{array}{rcl} W_g & = & \int_{x \in X_g} \sum_{i \in \mathcal{C}_g} \Pr(i, x|\theta)dx \\ & = & \int_{x \in X_g} \left(\sum_{i \in \mathcal{C}} P(i|x, \theta)\right) p(x)dx \\ & = & \int_{x \in X_g} p(x)dx, \end{array}$$

[1] We assume, without loss of generality, that the independent variables are all continuous, in order to simplify the formulas. In practice, it is usually a combination of continuous and discrete variables.

does not depend on $\theta$. Consequently, the probability $r_g$ for an individual to be selected in the sample is also independent from $\theta$.

3. An Endogenous Stratified Sampling (ESS), also called Choice-Based Sampling, is obtained when the groups are characterized by the endogenous variables only, that is the chosen alternative. In this case, $W_g$ does not simplify, and consequently depends on unknown parameters $\theta$. And so does the probability for an individual to be selected in the sample.

4. An Exogenous and Endogenous Stratified Sampling (XESS) is obtained when the groups are characterized by both the exogenous and endogenous variables. Again, $W_g$ depends on $\theta$.

Denoting $g(i, x)$ the group containing individuals with exogenous variable $x$ and choice $i$, the probability for an individual to be in the sample, conditional on $i$ and $x$ is

$$r_{g(i,x)}(\theta) = \Pr(s|i, x, \theta) \tag{5}$$

where we denote by $s$ the event of being in the sample, and $r$ is defined by (3).

For the sake of completeness, we also consider the context where the choice set $\mathcal{C}$ contains a large number of alternatives. It may be convenient to analyze choice *as if* it were limited to a subset $\mathcal{B} \subseteq \mathcal{C}$, in order to limit data collection and computation. We assume that for an observation with configuration $(i, x)$, the analyst draws a subset $\mathcal{B}$ with probability $\pi(\mathcal{B}|i, x)$. We further assume that any set $\mathcal{B}$ that is drawn with positive probability contains the chosen alternative $i$, and at least one non-chosen alternative. We also assume a *positive conditioning property* that $\pi(\mathcal{B}|i, x) > 0$ for the observed choice $i$ implies $\pi(\mathcal{B}|j, x) > 0$ for each $j \in \mathcal{B}$. It means that $\mathcal{B}$ could have been drawn conditioned on any of its elements as the observed choice. Finally, we assume without loss of generality that $r_{g(i,x)} > 0$ when $\pi(\mathcal{B}|i, x) > 0$.

We denote $R(i, x, \mathcal{B}, \theta)$ the probability that a population member with configuration $(i, x)$ is sampled, and is assigned the truncated choice set $\mathcal{B}$, that is

$$R(i, x, \mathcal{B}, \theta) = \Pr(s, \mathcal{B}|i, x, \theta) = r_{g(i,x)}(\theta)\pi(\mathcal{B}|i, x). \tag{6}$$

# 3 Estimation

Estimation issues in the presence of various sampling protocols are analyzed in details by Manski and McFadden (1981), Cosslett (1981) and Ben-Akiva and Lerman (1985, chapter 8).

Namely, Manski and McFadden (1981, Eq. (1.41)) suggest to use the Conditional Maximum Likelihood (CML) Estimator which produces consistent estimates. As compared to the full maximum likelihood estimation, efficiency is lost but consistency is retained. The objective function is given by

$$\mathcal{L}(\theta) = \sum_{n=1}^{N} \ln \Pr(i_n|x_n, \mathcal{B}_n, s, \theta), \tag{7}$$

where $i_n$ and $x_n$ are the observed dependent and independent variables for observation $n$ in the sample, and $\mathcal{B}_n$ the truncated choice set considered for that observation.

Denoting $p(x)$ the proportion of the population with exogenous variable $x$, we write the joint probability $\Pr(i, x, \mathcal{B}, s, \theta)$ in two ways:

$$\Pr(i, x, \mathcal{B}, s, \theta) = \Pr(i|x, \mathcal{B}, s, \theta) \Pr(\mathcal{B}, s|x, \theta)p(x)$$

and

$$\Pr(i, x, \mathcal{B}, s, \theta) = \Pr(\mathcal{B}, s|i, x, \theta) \Pr(i|x, \theta)p(x).$$

Therefore,

$$\Pr(i|x, \mathcal{B}, s, \theta) = \frac{\Pr(\mathcal{B}, s|i, x, \theta) \Pr(i|x, \theta)p(x)}{\Pr(\mathcal{B}, s|x, \theta)p(x)}$$

where

$$\Pr(\mathcal{B}, s|x, \theta) = \sum_{j \in \mathcal{B}} \Pr(\mathcal{B}, s|j, x, \theta) \Pr(j|x, \theta),$$

that is, using (1) and (6),

$$\Pr(i|x, \mathcal{B}, s, \theta) = \frac{R(i, x, \mathcal{B}, \theta)P(i|x, \theta)}{\sum_{j \in \mathcal{B}} R(j, x, \mathcal{B}, \theta)P(j|x, \theta)},$$

and (7) becomes

$$\mathcal{L} = \sum_{n=1}^{N} \ln \frac{R(i_n, x_n, \mathcal{B}_n, \theta) P(i_n | x_n, \theta)}{\sum_{j \in \mathcal{B}_n} R(j, x_n, \mathcal{B}_n, \theta) P(j | x_n, \theta)}, \tag{8}$$

which is a generalized version of Manski and McFadden (1981, Eq. (1.41)). Suppose now that $R(i, x, \mathcal{B}, \theta)$ can be written as a product

$$R(i, x, \mathcal{B}, \theta) = Q(i, x, \mathcal{B}) S(i, x, \mathcal{B}, \theta), \tag{9}$$

where Q is a term that contains no unknown parameters and can be calculated from the sampling protocol and data, and S is a term that may contain unknown parameters. The partition of R into Q and S need not correspond to the partition of R into $r$ and $\pi$, as defined by (6).

In this case, the pseudo-likelihood function

$$\widehat{\mathcal{L}} = \sum_{n=1}^{N} Q(i_n, x_n, \mathcal{B}_n)^{-1} \ln \frac{S(i_n, x_n, \mathcal{B}_n, \theta) P(i_n | x_n, \theta)}{\sum_{j \in \mathcal{B}_n} S(j, x_n, \mathcal{B}_n, \theta) P(j | x_n, \theta)} \tag{10}$$

has the same maximizers as (8) (as Q does not depend on $\theta$) and, therefore, its maximization provides consistent estimates of the parameters. It reduces to the CML estimator by Manski and McFadden (1981) when $Q = 1$, and to the Weighted Exogenous Sample Maximum Likelihood (WESML) Estimator proposed by Manski and Lerman (1977) when $S = 1$.

In the following, we derive this pseudo-likelihood function for Generalized Extreme Value models, and in particular the term

$$\frac{S(i, x, \mathcal{B}, \theta) P(i | x, \theta)}{\sum_{j \in \mathcal{B}} S(j, x, \mathcal{B}, \theta) P(j | x, \theta)}. \tag{11}$$

# 4 Generalized Extreme Value models

We consider here a random utility model where the utility associated to alternative $i$ is given by

$$U_i(x, \beta) = V_i(x, \beta) + \varepsilon_i, \tag{12}$$

where $\beta$ is a vector of unknown parameters to be estimated, and $\varepsilon_i$ is a random term (the fact that the model is individual-specific is ignored here for the sake of notational simplicity). Specific assumptions on the distributions of the random terms yield to operational choice models. Namely, Generalized Extreme Value models (McFadden, 1978) are based on the assumption that Cumulative Distribution Functions (CDF) of the joint distributions of the error terms are identically distributed across individuals, and that distributions are given by

$$F_\varepsilon = F_{\varepsilon_1,\ldots,\varepsilon_J}(V_1,\ldots,V_J) = e^{-G(e^{-V_1},\ldots,e^{-V_J};\gamma)}, \tag{13}$$

where $J$ is the number of alternatives in the choice set $\mathcal{C}$, $\gamma$ is a vector of parameters, and $G : \mathbb{R}^J \to \mathbb{R}$ is called a $\mu$-*GEV-generating* function, and has the following properties:

1. $G(y;\gamma) \geq 0$ for all $y \in \mathbb{R}^J_+$;
2. $G$ is homogeneous of degree $\mu > 0$, that is $G(\lambda y;\gamma) = \lambda^\mu G(y;\gamma)$, for $\lambda > 0$;
3. $\lim_{y_i \to +\infty} G(y_1,\ldots,y_i,\ldots,y_J;\gamma) = +\infty$, for each $i = 1,\ldots,J$;
4. The mixed partial derivatives of $G$ exist and are continuous. Moreover, the kth partial derivative with respect to $k$ distinct $y_i$ is non-negative if $k$ is odd and non-positive if $k$ is even that is, for any distinct indices $i_1,\ldots,i_k \in \{1,\ldots,J\}$, we have

$$(-1)^k \frac{\partial^k G}{\partial y_{i_1} \ldots \partial y_{i_k}}(y) \leq 0, \ \forall y \in \mathbb{R}^J_+. \tag{14}$$

Typical examples of $\mu$-GEV-generating functions $G$ are

$$G(y;\gamma = \{\mu\}) = \sum_{j=1}^{J} y_j^\mu, \tag{15}$$

generating a MNL model,

$$G(y;\gamma = \{\mu,\mu_1,\ldots,\mu_M\}) = \sum_{m=1}^{M} \left( \sum_{j\in\mathcal{C}_m} y_i^{\mu_m} \right)^{\frac{\mu}{\mu_m}}, \tag{16}$$

generating a nested logit model where the choice set $\mathcal{C}$ is partitioned into $M$ mutually exclusive nests $\mathcal{C}_1, \ldots, \mathcal{C}_M$, and

$$G(y; \gamma = \{\mu, \mu_1, \ldots, \mu_M, \alpha_{11}, \ldots, \alpha_{JM}) = \sum_{m=1}^{M} \left( \sum_{j \in \mathcal{C}} ({\alpha_{jm}}^{1/\mu} y_j)^{\mu_m} \right)^{\frac{\mu}{\mu_m}}, \quad (17)$$

generating a cross-nested logit (CNL) model where $\alpha_{jm} \geq 0$, $\forall j, m$, $\sum_{m=1}^{M} \alpha_{jm} > 0$, $\forall j$, $\mu > 0$, $\mu_m > 0$, $\forall m$, $\mu \leq \mu_m$, $\forall m$.

The choice model derived from the GEV assumption is given by

$$P(i|x, \theta) = \frac{\Lambda_i(x, \theta)}{\sum_{j \in \mathcal{C}} \Lambda_j(x, \theta)}, \quad (18)$$

where

$$\Lambda_i(x, \theta = (\beta; \gamma)) = e^{V_i(x,\beta) + \ln G_i(x,\beta,\gamma)}$$

and

$$G_i(x, \beta, \gamma) = \frac{\partial G}{\partial e^{V_i(x,\beta)}} \left( e^{V_1(x,\beta)}, \ldots, e^{V_J(x,\beta)}; \gamma \right). \quad (19)$$

In order to compute (11), we have

$$S(i, x, \mathcal{B}, \theta) P(i|x, \theta) = \frac{S(i, x, \mathcal{B}, \theta) \Lambda_i(x, \theta)}{\sum_{j \in \mathcal{C}} \Lambda_j(x, \theta)}.$$

We define

$$\begin{array}{rcl} \widehat{\Lambda}_i(x, \theta) & = & S(i, x, \mathcal{B}, \theta) \Lambda_i(x, \theta) \\ & = & e^{V_i(x,\beta) + \ln G_i(x,\beta,\gamma) + \ln S(i,x,\mathcal{B},\theta)}. \end{array}$$

Therefore, (11) writes

$$\begin{array}{l} \dfrac{S(i, x, \mathcal{B}, \theta) \Lambda_i(x, \theta)}{\sum_{k \in \mathcal{C}} \Lambda_k(x, \theta)} \Big/ \dfrac{\sum_{j \in \mathcal{B}} S(j, x, \mathcal{B}, \theta) \Lambda_j(x, \theta)}{\sum_{k \in \mathcal{C}} \Lambda_k(x, \theta)} \\ \\ \quad = \; S(i, x, \mathcal{B}, \theta) \Lambda_i(x, \theta) / \sum_{j \in \mathcal{B}} S(j, x, \mathcal{B}, \theta) \Lambda_j(x, \theta) \\ \quad = \; \widehat{\Lambda}_i(x, \theta) / \sum_{j \in \mathcal{B}} \widehat{\Lambda}_i(x, \theta), \end{array}$$

that is

$$\frac{S(i, x, \mathcal{B}, \theta) P(i|x, \theta)}{\sum_{j \in \mathcal{B}} S(j, x, \mathcal{B}, \theta) P(j|x, \theta)} = \frac{e^{V_i(\beta) + \ln G_i(x,\beta,\gamma) + \ln S(i,x,\mathcal{B},\theta)}}{\sum_{j \in \mathcal{B}} e^{V_j(\beta) + \ln G_j(x,\beta,\gamma) + \ln S(j,x,\mathcal{B},\theta)}}. \quad (20)$$

Note that for MNL, the terms $\ln G_i(x, \beta, \gamma)$ disappear and we obtain the well-known result that ignoring the sampling probabilities biases the constants while the other parameters are unbiased. Indeed,

$$\frac{S(i, x, \mathcal{B}, \theta)P(i|x, \theta)}{\sum_{j\in\mathcal{B}} S(j, x, \mathcal{B}, \theta)P(j|x, \theta)} = \frac{e^{V_i(\beta)+\ln S(i,x,\mathcal{B},\theta)}}{\sum_{j\in\mathcal{B}} e^{V_j(\beta)+\ln S(j,x,\mathcal{B},\theta)}}, \tag{21}$$

is a MNL model where the Alternative Specific Constants (ASC) are shifted by $\ln S(i, x, \mathcal{B}, \theta)$. Using (21) in (10) leads to a WESML estimator, readily available is estimation software packages.

Unfortunately, this nice property cannot be generalized as such to GEV models, as the probability (20) is **not** the probability of a genuine GEV model, although it looks very similar. Indeed, if the ASCs are shifted in the main expression, they are not shifted when used as arguments of $G_i(x, \beta, \gamma)$, defined by (19). Consequently, (W)ESML **cannot be used** as such for the estimation of non-MNL GEV models in the presence of selection bias, as it does not produce consistent estimates of the parameters. We illustrate in Section 7 that using ESML in this context indeed produces biased estimates of the parameters.

In the following, we present a special class of GEV models such that ESML can lead to consistent estimates of all parameters but the constants. Then, we present a new estimator for the general case of pure choice-base sampling.

## 5 Block additive GEV

We consider first the specific case where the choice set $\mathcal{C}$ is partitioned into $M$ mutually exclusive blocks

$$\mathcal{C} = \mathcal{C}_1 \cup \ldots \cup \mathcal{C}_M, \tag{22}$$

and the GEV-generating function has a block additive form,

$$G(y_1, \ldots, y_J; \gamma) = \sum_{m=1}^{M} G^m(y_{\mathcal{C}_m}; \gamma), \tag{23}$$

where $y_{\mathcal{C}_m}$ denotes the subvector of $y$ with components in $\mathcal{C}_m$, and each $G^m$ is a $\mu$-GEV generating function (note that $\mu$ can always be normalized to 1). We emphasize immediately that the nested logit model defined by (16) does not belong to this category, except when it collapses to a MNL model. This is why we refer to *blocks* and not *nests*, in order to avoid any confusion. This can be viewed as a Network GEV model (see Daly and Bierlaire, 2006) where the nodes just below the root share the same homogeneity parameter.

With respect to sampling, we assume that the following conditions hold:

1. $\pi(\mathcal{B}|i,x)$ is positive only if $\mathcal{B}$ is a union of blocks from $\{\mathcal{C}_1,\ldots,\mathcal{C}_M\}$; i.e., alternative selection respects the block structure of the GEV model.

2. The sampling terms $S(i,x,\mathcal{B},\theta)$ are uniform for $i$ in a block $\mathcal{C}_m$; i.e., any variation in sampling rates across responses within a block is handled by the weights Q in (10). With a slight abuse of notation, let $S(m,x,\mathcal{B},\theta)$ denote those common sampling terms within block $m$.

As $G^m$ is 1-homogenous, $G_i^m$ is 0-homogenous and shifting the V's in (19) does not affect $G_i(x,\beta,\gamma)$. In (11), we have

$$\begin{aligned} S(i,x,\mathcal{B},\theta)P(i|x,\theta)D &= S(i,x,\mathcal{B},\theta)e^{V_{in}(x,\beta)+\ln G_i^m(x,\beta,\gamma)} \\ &= e^{V_{in}(x,\beta)+\ln S(m,x,\mathcal{B},\theta)+\ln G_i^m(S(m,x,\mathcal{B},\theta)x,\theta)} \end{aligned} \tag{24}$$

where the denominator $D = \sum_{j=1}^{J} e^{V_{jn}(x,\beta)+\ln G_j(x,\beta,\gamma)}$ cancels out in (11), which writes

$$\frac{e^{V_{in}(x,\beta)+\ln S(m,x,\mathcal{B},\theta)+\ln G_i^m(S(m,x,\mathcal{B},\theta)x,\theta)}}{\sum_{k=1}^{M}\sum_{j\in\mathcal{B}\cap\mathcal{C}_k} e^{V_{jn}(x,\beta)+\ln S(m,x,\mathcal{B},\theta)+\ln G_j^k(S(m,x,\mathcal{B},\theta)x,\theta)}}. \tag{25}$$

This is a genuine GEV model, where the V's have been shifted. Note that the two assumptions made above are critical for the sum at the denominator to correspond to a GEV model

Consequently, we obtain the same property as for MNL, that the sampling effects S are absorbed by the alternative specific constants, and yield to a transformed GEV model. A common (W)ESML procedure yields consistent estimates of all parameters except the constant.

The result above is certainly special, but it is applicable when joining several complex models together. For instance, a route choice model for public transportation and a car type choice model for private transportation can each be analyzed separately by a GEV model of any complexity. Then, the mode choice dimension can be added in the model using the block additive form described above. In this case, if public transportation is over- or under- sampled, the bias will be absorbed in the constants, similarly to the MNL case.

# 6 GEV and choice-based sampling

We assume here that the term $S$ does not depend on $x$, that is

$$S(i, x, \mathcal{B}, \theta) = S(i, \mathcal{B}, \theta).$$

This assumption is fairly general, as selection based on $x$ can be captured by $Q$ in (9). In this case, the simple form of (20) suggests to use another estimator which explicitly estimates $\ln S(i, \mathcal{B}, \theta)$ from the data. It is obtained by solving

$$\max_{\beta,\gamma,\omega} \sum_{n=1}^{N} Q(i_n, x_n, \mathcal{B}_n)^{-1} \ln \frac{e^{V_{i_n}(\beta)+\ln G_{i_n}(x,\beta,\gamma)+\omega_{i_n}}}{\sum_{j\in\mathcal{C}} e^{V_j(\beta)+\ln G_j(x,\beta,\gamma)+\omega_j}}, \qquad (26)$$

where $\omega_i$ is a parameter designed to directly estimate $\ln S(i, \mathcal{B}, \theta)$.

We illustrate below that these parameters can be identified for alternatives $i$ such that the GEV term $\ln G_i(x, \beta, \gamma)$ is not zero. The estimator defined by (26) can easily be implemented, as its formulation is very similar to a real GEV model. It is actually available with the Biogeme software package (Bierlaire, 2003, Bierlaire, 2005, `biogeme.epfl.ch`).

It is very important to emphasize that, if the structure of the GEV model is non trivial, that is $\ln G_i(x, \beta, \gamma)$ is not zero, *both the ASC and the sampling probability are identified.* For alternatives such that $\ln G_i(x, \beta, \gamma)$ $= 0$, only the sum is identified. As illustrated below, the ASC and the sampling probability are identified in a nested model if the nest parameter is significantly different from 1. If it is 1, or close to 1, only the sum is identified.

# 7 Illustrations

So the sake of simplicity of the following analysis, we do not consider weighting, so that Q is always 1 in (10) and (26), and we do not sample the alternatives, so that $\pi(\mathcal{B}|i, x) = 1$ if $\mathcal{B} = \mathcal{C}$, and 0 otherwise. We first illustrate the new estimator on pseudo synthetic data. Then we apply it to real data sets and compare the results with the ESML estimator.

## 7.1 Synthetic data: mode choice in Switzerland

The first set of data has been generated based on a real stated preferences data set collected for the analysis of a future high speed train in Switzerland (Bierlaire et al., 2001). The alternatives are

1. Regular train (TRAIN),
2. Swissmetro (SM), the future high speed train,
3. Driving a car (CAR).

Each of the 6768 observations in the sample has been used to generate 75 synthetic observations, to obtain a population of 507600 individuals. The value of each attribute in the population has been generated from a normal distribution $N(\mu, \sigma^2)$, where $\mu$ is the value of the corresponding attribute in the original database, and $\sigma = 0.05\mu$. A choice has been associated with each observation in the population using a nested logit model with the following specification table:

| | | Alternatives | | |
|---:|---:|---:|---:|---:|
| Param. | Value | TRAIN | SM | CAR |
| ASC_CAR | -0.1880 | 0 | 0 | 1 |
| ASC_SM | 0.1470 | 0 | 1 | 0 |
| B_TRAIN_TIME | -0.0107 | travel time | 0 | 0 |
| B_SM_TIME | -0.0081 | 0 | travel time | 0 |
| B_CAR_TIME | -0.0071 | 0 | 0 | travel time |
| B_COST | -0.0083 | travel cost | travel cost | travel cost |

The nesting structure is defined by

|  | $\mu_m$ | TRAIN | SM | CAR |
|---|---|---|---|---|
| NESTA | 2.27 | 1 | 0 | 1 |
| NESTB | 1.0 | 0 | 1 | 0 |

where $\mu_m$ is the nest parameter $\mu_m$ in (16), the scale $\mu$ being set to one. The "true" value of the parameters were obtained from the estimation of the model on the real data set. Due to this nest structure, only one $\omega$ parameter in (26) is identified, in this case the one associated with alternative CAR. Indeed, the parameter $\omega_{\mathtt{SM}}$ is confounded with the ASC as this is the sole alternative in the second nest. In the first nest, similarly to the ASCs, one of the two $\omega$s is constrained to 0.

We have extracted 100 samples from the population using a choice-based sampling strategy defined as follows:

| Strata | $W_g N_P$ | $W_g$ | $H_g$ | $H_g N_s$ | $R_g$ |
|---|---|---|---|---|---|
| TRAIN | 67938 | 13.4% | 60% | 3000 | 4.42E-02 |
| SM | 306279 | 60.3% | 20% | 1000 | 3.26E-03 |
| CAR | 133383 | 26.3% | 20% | 1000 | 7.50E-03 |
| Total | 507600 | 1 | 1 | 5000 |  |

A model has been estimated with each of these 100 samples, once with the ESML estimator, and once with the new estimator. Table 1 reports, for each parameter, its "true" value, the mean and standard deviation of the 100 estimated values, and the associated t-test, that is $(\widehat{\theta}_k - \theta^*_k)/\widehat{\sigma}_k$, the ratio of the difference between the mean estimated value and the true value, and the estimated standard deviation.

Most parameters estimated with the ESML are significantly different from the true value, with a t-test larger than 1.96. This illustrates that ESML may produce biased estimates with nested logit models.

The new estimator produces estimates which are not significantly different from the true value for all parameters except ASC_SM. The true values of the $\omega$ parameters are $\omega_{\mathtt{TRAIN}} = \ln$ 4.42e-02 = -3.1200, $\omega_{\mathtt{SM}} = \ln$ 3.26e-03 = -5.7245, $\omega_{\mathtt{CAR}} = \ln$ 7.50e-03 = -4.8932. Similarly to the ASCs, these parameters are identified up to a constant. In order to reflect the fact that we have constrained $\omega_{\mathtt{TRAIN}} = 0$ in the model, we report $\omega_i + 3.12$ in the table.

We note that the value of ASC_CAR and the (shifted) value of $\omega_{\texttt{CAR}} = \ln R_{\texttt{CAR}}$ are both correctly estimated. Moreover, the value of $\omega_{\texttt{SM}}+$ ASC_SM is also correctly estimated, like it would be in a MNL model, as illustrated in the last row of Table 1.

We have performed a similar analysis for a cross-nested logit (CNL) model on the same population, and the same 100 samples, where new choices have been generated with a "true" CNL model. The specification of the utility functions is the same as before, and the nesting structure is defined by

| | $\mu_m$ | TRAIN | SM | CAR |
|---|---|---|---|---|
| NESTA | 4.0 | 0.9 | 0.5 | 0.1 |
| NESTB | 2.0 | 0.1 | 0.5 | 0.9 |

where the entries in the table are the values of parameters $\mu_m$ and $\alpha_{jm}$ in (17), the scale $\mu$ being set to one. The results reported in Table 2 are consistent with the findings obtained with the NL model: ESML produces biased values, while the new estimator produces values which are not significantly different from the true values. Moreover, because this GEV structure does not collapse to MNL, as both nest parameters are different from one and no alternative is alone in a nest, all ASCs and $\omega$s are separately identified (except, of course, for the base alternative). Finally, we observe that the complexity of the CNL model produces relatively large standard deviations for the constants, as well as for the nest parameters. This illustrates that estimating a complex GEV model together with the sampling probabilities comes with a cost, and that larger samples are required to obtain accurate estimates.

## 7.2 Synthetic data: choice of energy in Québec

We perform a similar exercise based on synthetic data generated from a real revealed preferences data set, capturing the choice of energy in Québec (Bernard et al., 1996), where the choice of energy for house and water heating is among the following alternatives:

1. Dual energy for house, and electricity for water (DE)

|  |  | ESML |  |  | New estimator |  |  |
|---|---|---|---|---|---|---|---|
|  | True | Mean | t-test | Std. dev. | Mean | t-test | Std. dev. |
| ASC_SM | 0.1470 | -2.2479 | -25.4771 | 0.0940 | -2.4900 | -23.9809 | 0.1100 |
| ASC_CAR | -0.1880 | -0.8328 | -7.3876 | 0.0873 | -0.1676 | 0.1581 | 0.1292 |
| BCOST | -0.0083 | -0.0066 | 2.6470 | 0.0007 | -0.0083 | 0.0638 | 0.0008 |
| BTIME_TRAIN | -0.0107 | -0.0094 | 1.4290 | 0.0009 | -0.0109 | -0.1774 | 0.0009 |
| BTIME_SM | -0.0081 | -0.0042 | 3.1046 | 0.0013 | -0.0080 | 0.0446 | 0.0014 |
| BTIME_CAR | -0.0071 | -0.0065 | 0.9895 | 0.0007 | -0.0074 | -0.3255 | 0.0007 |
| NestParam | 2.2700 | 2.7432 | 1.7665 | 0.2679 | 2.2576 | -0.0609 | 0.2043 |
| S_TRAIN | -3.1200 |  |  |  |  |  |  |
| S_SM | -5.7245 |  |  |  |  |  |  |
| S_CAR | -4.8932 |  |  |  |  |  |  |
| S_SM_Shifted | -2.6045 |  |  |  |  |  |  |
| S_CAR_Shifted | -1.7732 |  |  |  | -1.7877 | -0.0546 | 0.2651 |
| ASC_SM+S_SM | -2.4575 |  |  |  | -2.4900 | -0.2958 | 0.1100 |

Table 1: Nested logit model on the Swissmetro synthetic data

|  |  | ESML |  |  | New estimator |  |  |
|---|---|---|---|---|---|---|---|
|  | True | Base avg | t-test | Base stderr | Corr avg | t-test | Corr stdev |
| ASC_SM | 0.4520 | -1.0249 | -11.9786 | 0.1233 | 0.8321 | 0.1139 | 3.3367 |
| ASC_CAR | 0.1650 | -0.7719 | -10.2298 | 0.0916 | 0.4092 | 0.0677 | 3.6051 |
| BCOST | -0.0049 | -0.0058 | -1.8222 | 0.0005 | -0.0044 | 0.3793 | 0.0012 |
| BTIME_TRAIN | -0.0048 | -0.0087 | -6.5725 | 0.0006 | -0.0045 | 0.2715 | 0.0012 |
| BTIME_SM | -0.0040 | -0.0064 | -3.1970 | 0.0007 | -0.0037 | 0.2426 | 0.0011 |
| BTIME_CAR | -0.0049 | -0.0061 | -1.9366 | 0.0006 | -0.0045 | 0.2802 | 0.0013 |
| NESTA | 4.0000 | 2.9003 | -2.0751 | 0.5299 | 4.8414 | 0.4034 | 2.0854 |
| NESTB | 2.0000 | 1.4935 | -3.4632 | 0.1462 | 2.5172 | 0.4697 | 1.1011 |
| S_TRAIN | -3.3323 |  |  |  |  |  |  |
| S_SM | -5.7410 |  |  |  |  |  |  |
| S_CAR | -4.4326 |  |  |  |  |  |  |
| S_SM_Shifted | -2.4087 |  |  |  | -3.6570 | -0.1114 | 11.2056 |
| S_CAR_Shifted | -1.1003 |  |  |  | -2.1203 | -0.0897 | 11.3681 |

Table 2: Cross-Nested logit model on the Swissmetro synthetic data

2. Electricity for both house and water (EE)

3. Wood for the house and electricity for water (WE).

The synthetic population is composed of 284800 individuals. A choice has been associated with each observation in the population using a nested logit model with the following specification table:

| | | Alternatives | | |
|---|---|---|---|---|
| Param. | Value | DE | EE | WE |
| ASC_DE | -2.0 | 1 | 0 | 0 |
| ASC_WE | -3.0 | 0 | 0 | 1 |
| B_COST_DE_RURAL | -15.0 | cost × rural | 0 | 0 |
| B_COST_DE_URBAN | -13.0 | cost × urban | 0 | 0 |
| B_COST_EE_RURAL | -16.0 | 0 | cost × rural | 0 |
| B_COST_EE_URBAN | -12.0 | 0 | cost × urban | 0 |
| B_COST_WE_RURAL | -26.0 | 0 | 0 | cost × rural |
| B_COST_WE_URBAN | -20.0 | 0 | 0 | cost × urban |
| B_FIXED_COST | -0.6 | fixed cost | fixed cost | fixed cost |

The nesting structure is defined by

| | $\mu_m$ | DE | EE | WE |
|---|---|---|---|---|
| NESTA | 3.0 | 1 | 1 | 0 |
| NESTB | 1.0 | 0 | 0 | 1 |

where $\mu_m$ is the nest parameter $\mu_m$ in (16), the scale $\mu$ being set to one. Due to this nest structure, only one $\omega$ parameter in (26) is identified, in this case the one associated with alternative DE. Indeed, the parameter $\omega_{\texttt{WE}}$ is confounded with the ASC as this is the sole alternative in the second nest. In the first nest, similarly to the ASCs, one of the two $\omega$s is constrained to 0.

We have extracted 200 samples from the population using a choice-based sampling strategy defined as follows:

|  | $W_g N_P$ | $W_g$ | $H_g$ | $H_g N_s$ | $R_g$ |
|---|---|---|---|---|---|
| DE | 57510 | 20.2% | 60% | 5127 | 8.91E-02 |
| EE | 162040 | 56.9% | 10% | 854 | 5.27E-03 |
| WE | 65250 | 22.9% | 30% | 2563 | 3.93E-02 |
| Total | 284800 |  |  | 8544 |  |

As for the previous example, a model has been estimated with each of these 200 samples, once with the ESML estimator, and once with the new estimator. Table 3 reports, for each parameter, its “true” value, the mean and standard deviation of the 200 estimated values, and the associated t-test against the true value.

The general analysis of the results is consistent with the previous section, as we observe a significant improvement of the quality of the estimates of all parameters, including the constant of the alternative which is not alone in a nest. We note that the ESML results seem less biased in this example than in the previous one, except for the nest parameter, which is significantly different from the true value.

### 7.3 Real data

When estimating a nested logit model on the real Swissmetro data set, composed of 6768 observations, we obtain significantly different results using ESML and the new estimator, as presented in Table 4. The final loglikelihood with the new estimator is significantly better than with ESML. Also, the parameter S_CAR, capturing the selection bias, is significantly different from 0. It is interesting to note that, except for the constants, the standard error of the parameters are similar with both estimators.

We have estimated a nested logit model with 9 alternatives on the real Québec data, and have reached the exact same conclusions. The results are described in Table 5.

## 8 Conclusion

We have shown both theoretically and empirically that ESML applied to the estimation of GEV models on choice-based samples does not yield to

|  |  | ESML |  |  | New estimator |  |  |
|---|---|---|---|---|---|---|---|
|  | True | Base avg | t-test | Base stderr | Corr avg | t-test | Corr stdev |
| ASC_DE | -2 | -0.5280 | 6.9173 | 0.2128 | -2.0455 | -0.2230 | 0.2039 |
| ASC_WE | -3 | -1.8771 | 5.2530 | 0.2138 | -0.9472 | 9.1075 | 0.2254 |
| B_COST_DE_RURAL | -15 | -15.7350 | -0.9641 | 0.7624 | -14.7913 | 0.2906 | 0.7181 |
| B_COST_DE_URBAN | -13 | -13.8040 | -1.5306 | 0.5253 | -12.8853 | 0.2208 | 0.5193 |
| B_COST_EE_RURAL | -16 | -16.6651 | -0.8404 | 0.7914 | -15.8285 | 0.2293 | 0.7483 |
| B_COST_EE_URBAN | -12 | -12.6509 | -1.3388 | 0.4862 | -11.9306 | 0.1438 | 0.4829 |
| B_COST_WE_RURAL | -26 | -27.3967 | -1.0455 | 1.3358 | -25.7547 | 0.1960 | 1.2512 |
| B_COST_WE_URBAN | -20 | -20.9689 | -1.1814 | 0.8201 | -19.8855 | 0.1407 | 0.8141 |
| B_FIXED_COST | -0.6 | -0.6551 | -0.6333 | 0.0871 | -0.6029 | -0.0407 | 0.0709 |
| NESTA | 3 | 2.1629 | -4.5576 | 0.1837 | 3.0068 | 0.0250 | 0.2699 |
| SB_DE | 2.8282 |  |  |  | 2.8324 | 0.0178 | 0.2323 |
| ASC_WE + SB_WE | -0.99138 |  |  |  | -0.9472 | 0.1959 | 0.2254 |

Table 3: Nested logit model on the Québec synthetic data

| | ESML | | | | New estimator | | | |
|---|---|---|---|---|---|---|---|---|
| Parameters | 7 | | | | 8 | | | |
| $\mathcal{L}(0)$ | -6964.7 | | | | -6964.7 | | | |
| $\mathcal{L}(\theta^*)$ | -5203.9 | | | | -5160.3 | | | |
| | Param. | Std. Err | t-test (0) | t-test (1) | Param. | Std. Err | t-test (0) | t-test (1) |
| ASC_CAR | -0.1884 | 0.0754 | -2.4970 | | 5.4856 | 2.1496 | 2.5519 | |
| ASC_SM | 0.1475 | 0.1005 | 1.4669 | | -0.3880 | 0.1098 | -3.5335 | |
| B_CAR_TIME | -0.0071 | 0.0012 | -6.0234 | | -0.0097 | 0.0012 | -8.2135 | |
| B_COST | -0.0083 | 0.0006 | -14.4558 | | -0.0109 | 0.0007 | -16.6062 | |
| B_SM_TIME | -0.0081 | 0.0017 | -4.7251 | | -0.0114 | 0.0018 | -6.3579 | |
| B_TRAIN_TIME | -0.0108 | 0.0011 | -9.6022 | | -0.0131 | 0.0011 | -12.1740 | |
| NEST | 2.2626 | 0.1864 | 12.1362 | 6.7724 | 1.2361 | 0.0826 | 14.9735 | 2.8602 |
| S_CAR | | | | | -6.4116 | 2.1132 | -3.0341 | |

Table 4: Estimation of a NL model on real Swissmetro data

|  | ESML |  |  |  | New estimator |  |  |  |
|---|---|---|---|---|---|---|---|---|
| Parameters | 18 |  |  |  | 21 |  |  |  |
| $\mathcal{L}(0)$ | -6159.6 |  |  |  | -6159.6 |  |  |  |
| $\mathcal{L}(\theta^*)$ | -1897.8 |  |  |  | -1882.6 |  |  |  |
|  | Param. | Std. Err. | t-test (0) | t-test (1) | Param. | Std. Err. | t-test (0) | t-test (1) |
| ASC2 | 3.57313 | 0.78605 | 4.54566 |  | 0.13363 | 0.96975 | 0.13779 |  |
| ASC3 | 2.10748 | 0.80494 | 2.61816 |  | 1.9256 | 0.79239 | 2.43012 |  |
| ASC4 | 4.50485 | 0.71839 | 6.27074 |  | 0.88464 | 0.77121 | 1.14708 |  |
| ASC5 | 2.52055 | 0.96238 | 2.61907 |  | 2.29622 | 0.95361 | 2.40793 |  |
| ASC6 | 5.22892 | 0.74329 | 7.03482 |  | 2.90643 | 0.72116 | 4.03018 |  |
| ASC7 | 5.25963 | 0.73004 | 7.20461 |  | 5.03169 | 0.71813 | 7.00667 |  |
| ASC8 | 1.12052 | 0.74219 | 1.50976 |  | 1.12956 | 0.73606 | 1.53462 |  |
| ASC9 | 1.88937 | 0.75008 | 2.51889 |  | 1.7756 | 0.73786 | 2.4064 |  |
| b1coutm | -5.6492 | 0.58594 | -9.6413 |  | -5.3703 | 0.60043 | -8.9441 |  |
| b2rev | -0.487 | 0.15873 | -3.0682 |  | -0.4665 | 0.15095 | -3.0904 |  |
| b3rev | -0.1939 | 0.13441 | -1.4423 |  | -0.1904 | 0.13263 | -1.4357 |  |
| b4rev | -0.298 | 0.12617 | -2.3623 |  | -0.2855 | 0.124 | -2.3026 |  |
| b5rev | -0.3383 | 0.18538 | -1.8246 |  | -0.3392 | 0.18488 | -1.8347 |  |
| b6rev | -0.2833 | 0.13098 | -2.1631 |  | -0.2724 | 0.12822 | -2.1242 |  |
| b7rev | -0.2824 | 0.12513 | -2.257 |  | -0.2717 | 0.12307 | -2.2079 |  |
| b8rev | -0.7555 | 0.13664 | -5.5289 |  | -0.7481 | 0.13468 | -5.5548 |  |
| b9rev | -0.464 | 0.13508 | -3.4351 |  | -0.4598 | 0.13322 | -3.4515 |  |
| NEST | 3.03155 | 0.3361 | 9.0197 | 6.04443 | 3.2335 | 0.38858 | 8.32125 | 5.7478 |
| SB_2 |  |  |  |  | 10.8209 | 0.95508 | 11.3298 |  |
| SB_4 |  |  |  |  | 11.0945 | 0.24404 | 45.4613 |  |
| SB_6 |  |  |  |  | 6.85779 | 0.28067 | 24.4339 |  |

Table 5: Estimation of a NL model on real Québec data

consistent estimates of the parameters in the general case. However, we have shown that the form of the GEV probability allows for the derivation of a simple estimator. We have illustrated the good quality of this estimator both on synthetic and real data.

The authors would like to thank Moshe Ben-Akiva and Kenneth Train for useful discussions and comments about earlier versions of the paper.

# References

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

Bernard, J.-T., Bolduc, D. and Bélanger, D. (1996). The estimation of electricity demand in Québec using micro data, *Canadian Journal of Economics* **XXIX**(1): 92–113.

Bierlaire, M. (2003). BIOGEME: a free package for the estimation of discrete choice models, *Proceedings of the 3rd Swiss Transportation Research Conference*, Ascona, Switzerland. www.strc.ch.

Bierlaire, M. (2005). An introduction to BIOGEME version 1.4. biogeme.epfl.ch.

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

Cosslett, S. (1981). Efficient estimation of discrete choice models, *in* C.Manski and D. McFadden (eds), *Structural analysis of discrete data with econometric applications*, MIT Press, Cambridge, Ma.

Daly, A. and Bierlaire, M. (2006). A general and operational representation of generalised extreme value models, *Transportation Research Part B* **40**(4): 285–305.

Imbens, G. (1992). An efficient method of moments estimator for discrete choice models with choice-based sampling, *Econometrica* **60**(5): 1187–1214.

Manski, C. (1999). Nonparametric identification under response-based sampling, *in* C. Hsiao, K. Morimune and J. Powell (eds), *Nonlinear Statistical Inference: Essays in Honor of Takeshi Amemiya*, Cambridge University Press, New York.

Manski, C. and Lerman, S. (1977). The estimation of choice probabilities from choice-based samples, *Econometrica* **45**: 1977–1988.

Manski, C. and McFadden, D. (1981). Alternative estimators and sample designs for discrete choice analysis, *in* C. Manski and D. McFadden (eds), *Structural analysis of discrete data with econometric application*, MIT Press, Cambridge, Mass.

McFadden, D. (1978). Modelling the choice of residential location, *in* A. Karlquist *et al.* (ed.), *Spatial interaction theory and residential location*, North-Holland, Amsterdam, pp. 75–96.

Morgenthaler, S. and Vardi, Y. (1986). Choice-based samples. A nonparametric approach, *Journal of econometrics* **32**: 109–125.