Sampling of alternatives using multidimensional analysis

M. Bierlaire[*] A. Lucadamo[†]

May 22, 2007

Report TRANSP-OR 070522
Transport and Mobility Laboratory
School of Architecture, Civil and Environmental Engineering
Ecole Polytechnique Fédérale de Lausanne
transp-or.epfl.ch

[*] École Polytechnique Fédérale de Lausanne, Transport and Mobility Laboratory, Station 18, CH-1015 Lausanne, Switzerland. michel.bierlaire@epfl.ch

[†] Department of "Matematica e Statistica", University of Naples "Federico II". alucadam@unina.it

Abstract

Discrete choice models in general and random utility models in particular may be intractable when the number of alternatives is large. In the transportation context, it typically happens for route choice and destination choice models. In the specific case of the widely used multinomial logit model, it has been shown that the model could be estimated as if the choice was made among a subset of the alternatives. In this paper, we propose to design the sampling of alternatives based on a Principal Component Analysis and a Cluster Analysis of the actual data set, in order to increase the efficiency of the estimates. We present a case study of a destination choice model to empirically illustrate the added value of our approach.

1 The Multinomial logit

The multinomial logit is the simplest model in discrete choice analysis when more than two alternatives are in a choice set. It is derived from utility-maximizing theory. The consumer chooses the alternative which maximizes this utility (McFadden 1978). Obviously not all the attributes of the alternatives will be observed. The utility is divided into two parts, V_{in} which is the systematic part, and ε_{in} which summarizes the contribution of unobserved variables. The probability to select an alternative i from the choice set C_n is then:

$$P\left(i|C_n\right) = \Pr(V_{in} + \varepsilon_{in} \geq V_{jn} + \varepsilon_{jn} \quad \forall j \in C_n\ )$$

If we assume that the disturbances are independent and identically extreme value distributed we obtain a Multinomial Logit model. The probability that the alternative i will be chosen is:

$$P_n\left(i\right) = \frac{e^{\mu V_{in}}}{\sum_{j \in C_n} e^{\mu V_{jn}}}$$

The term μ is a scale parameter, generally normalized to 1. The model is described in various textbooks, such as Ben-Akiva & Lerman (1985).

2 Sampling of alternatives

When there are many alternatives in C_n, as in destination choice models and in route choice models, there is a computational burden for the estimation. In this case, utilizing the independence from irrelevant alternatives property (IIA) of the logit model, it's possible to estimate the parameters with a subset of alternatives. Clearly, in this case, it would be only possible to maximize a conditional likelihood function rather than the true likelihood. A procedure for sampling the alternatives assigns to observation n a subset of alternatives D that must include the chosen alternative. The conditional probability of alternative i being chosen, given a sample of alternatives D, is

$$\pi_n(i|D) = \frac{\pi_n(D|i)\, P_n(i)}{\sum_{j \in D} \pi_n(D|j)\, P_n(j)}$$

where $\pi_n(D|i)P_n(i)$ is the joint probability of drawing a chosen alternative i and a subset of alternatives D.

The conditional probability $\pi_n(i|D)$ exists if

$$\pi_n(i|D) > 0 \forall j \in D$$

This is condition is called *positive conditioning property*, and is necessary for the derivation of a consistent estimator for the multinomial logit model (McFadden 1978), or the GEV model (Bierlaire, Bolduc & McFadden 2006), with samples of alternatives.

The simplest approach to sample design is to draw a simple random sample of J alternatives and to add the chosen alternative if it is not otherwise included. To prevent the possibility of samples with different choice set sizes, it is possible to draw randomly J alternatives from all the available alternatives, except for the chosen alternative, that is added afterward. Other methods are the “Importance Sampling of Alternatives” and “Stratified Importance Sampling” (Ben-Akiva & Lerman 1985).

3 "PCA Cluster Sampling (PCACS)"

Our proposal is to generate stratified sampling based on a Principal Component Analysis (PCA) and a Cluster Analysis. The central idea of the Principal Component Analysis is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in data set (Jolliffe 2002). This is achieved by transforming it into a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables. To obtain the components we must find the eigenvalues and the eigenvectors of the following matrix:

$$MX'WXM$$

where M is the metric matrix, X is the data matrix and W is the matrix of the weights. The goal is then to maximize the following expressions:

$$u'MX'WXMu$$

with the constraint $u_1^{'}Mu_1 = 1$. We can consider the maximization of the Lagrange multiplier: $L = u_1^{'}Au_1 - \lambda_1(u_1^{'}Mu_1 - 1) = \max$ and considering the partial derivative we obtain the first eigenvalue and the first eigenvector. The first component will be $c_1 = XMu_1$. To obtain the other components we must simply introduce some orthogonality constraints, that, i.e. for the second component, will be $u_1^{'}u_2 = 0$.

With the Principal Component Analysis we obtain components that are uncorrelated and we can proceed with the second step of the analysis. We introduce a Cluster Analysis, a method for grouping objects of similar kind into respective categories. There are different algorithms to obtain this goal, we used a hierarchical tree. This algorithm begins with each object in a class by itself. In every step the two more similar objects, according to some distance measures, are joined together. The most straightforward way of computing distances between objects in a multi-dimensional space is to compute Euclidean distances, but also other measures could be used.

When the algorithm stops we can cut the tree according to some optimality measures and we obtain a certain number of clusters.

They will have different sizes and therefore, for the sampling, we must assign different selection probabilities in different strata, while maintaining uniform selection probabilities within strata. We can proceed to the sampling in the following way:

1. Let k be the number of clusters we obtain from PCA and Cluster Analysis;

2. Let define by J the number of alternatives in the full choice set;

3. Let R_i be the number of alternatives in every cluster where $i = 1, ..., k$;

4. Let $J_i^{'}$ be the size of the sub-set we defined, $i = 1, ..., k$;

5. Let define with $R_i^{'}$ the number of alternatives we have to draw from every cluster, where $i = 1, ..., k$;

then the following equality must hold: $\frac{R_i^{'}}{J_i^{'}} = \frac{R_i}{J}$ and then: $R_i^{'} = \frac{R_i}{J} J_i^{'}$.

In this way we obtain a number of alternatives from every cluster that is proportional to the size of it. The probability to be selected for every alternative is the same, but in classical random sampling we do not know what kind of alternatives we select, so it is possible to obtain all the alternatives with similar characteristics and so there could be some problems with the estimation. With the Cluster Sampling instead we obtain a choice set which reflects the full one better.

To illustrate the advantages of this technique we applied it to a destination choice model.

4 Results

Our analysis concerns a household survey conducted in 2005 in the Greater Zürich area. The data-set includes about 700 alternatives and more than 50 observed variables (Burgle 2006). The first step was the building of a model for the full choice set. We used a multinomial logit with only linear-in-parameter utilities, we used BIOGEME (Bierlaire 2003) to estimate the values and we obtained 7 significant variables. The second step of the analysis was the building of data sets of different size (12-15-20-40 alternatives) with the random sampling and the PCACS. The sampling procedure was repeated 5 times for the two techniques. In this way we could compute the variance due to the sampling of alternatives. The last step of the analysis was the estimation of the parameters on the reduced choice-sets and then the comparisons between the two techniques of sampling. The measures we considered for the evaluation of the differences between the two techniques are the ability to recover model parameters, to replicate the choice probability of the chosen alternative for each observation and to estimate the overall log-likelihood function accurately (Nerella & Bhat 2004). For each of the criteria identified above, the evaluation of proximity was based on three properties:

1. The bias, or the difference between the mean of estimates for each sample size of alternatives across the 5 runs and the true values;

2. The variance in the relevant parameters across the 5 runs for each sample size of alternatives;

3. The total error, or the difference between the estimated and the true values across all 5 runs for each sample size of alternatives.

Before computing all the mentioned performance measures we can have some preliminary information from the data simply by considering the significance and the signs of the parameters estimated on the different sub-sets. We will show here the results we obtained with data sets composed of 20 alternatives, but they are similar also for the other sizes. We can see from the first two tables that for all 5 samples obtained by the two different techniques, the signs of the coefficients are the same as the full choice set. This is the first thing we must look at to judge the accuracy of the new estimations. There are anyway some differences in the values of the Robust t-test. In fact we can note that in table 1, relative to the random sampling, there are two samples in which a parameter, the density of children, has a low value for the Robust t-test. For PCACS (table 2) this does not happen.

At this point we can consider the different measures we underlined previously. In table 3 there are the differences between the mean, across the 5 runs, of the parameters and the values estimated on the full choice set. We can see that with the PCACS the sum of the differences between the parameters is inferior to the Random Sampling, so we have a lower bias. Table 4 summarizes the variance of the parameters across the 5 runs. The last row shows that there is a little improvement with the PCA Cluster Sampling. In table 5 there are the differences between the true values and all the estimated values. We do not insert all the differences, but we can see directly the sum of them and we can note how the PCACS shows once again the lowest value.

The second useful indicator to compare the techniques is the ability to replicate the choice probability of the chosen alternative for each observation. Also in this case we can compute the bias, the total error and the variance across the 5 samples (table 6). In 7 instead there are the indicators related to the ability to recover the true log-likelihood function. In both the cases the values are better for the PCACS. Obviously, as with any numerical exercise, the usual cautions for generalizing the results, also apply to this paper. There is a need for more computational and empirical research on the topic of sampling of alternatives to draw more definitive conclusions. However, we think that when the full choice set is too big to be used, the PCACS could be a useful technique to use to obtain good estimation of the parameters, in fact we can obtain a choice set which reflects the full one better than other techniques.

5 Acknowledgments

The presented paper is made possible by the support of other people. We would like to thank Michaela Burgle, Michael Loechl and Kay W. Axhausen of the ETH Zürich for the data-set and the helpful advises.

References

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

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

Bierlaire, M., Bolduc, D. & McFadden, D. (2006). The estimation of generalized extreme value models from choice-based samples, *Technical Report TRANSP-OR 060810*, Transport and Mobility Laboratory, ENAC, EPFL, Lausanne, Switzerland. Download from transp-or.epfl.ch.

Burgle, M. (2006). Residential location choice model for the greater zurich area, *Proceedings of the 6th Swiss Transport Research Conference*, Ascona, Switzerland.

Jolliffe, I. T. (2002). *Principal Component Analysis*, Springer, New York.

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.

Nerella, S. & Bhat, C. R. (2004). Numerical analysis of the effect of sampling of alternatives in discrete choice models, *Transportation Research Record* **1894**: 11–19.

	Full choice set		**Random 1**		**Random 2**		**Random 3**		**Random 4**		**Random 5**	
Parameters	*Val.*	*t-test*	*Val.*	*t-test*	*Val.*	*t-test*	*Val.*	*t-test*	*Val.*	*t-test*	*Val.*	*t-test*
access	0.51	6.33	0.76	5.84	0.79	5.97	0.71	5.32	0.30	4.39	0.82	5.78
childdensity	-0.05	-2.18	-0.04	-2.48	-0.03	-2.25	-0.02	-1.82	-0.04	-2.85	-0.01	-0.94
distwork	-0.14	-16.85	-0.08	-12.6	-0.08	-11.9	-0.05	-11.2	-0.09	-12.8	-0.05	-11.0
popyoung	0.02	10.32	0.01	9.4	0.01	9.64	0.01	7.88	0.02	11.37	0.01	4.41
rentratio	1.23	-4.86	-0.89	-4.31	-0.87	-4.01	-0.67	-3.32	-0.93	-4.55	-0.93	-3.14
taxindex	-0.02	-4.09	-0.02	-5.67	-0.02	-5.78	-0.02	6.48	-0.02	-4.47	0.02	-5.85
timetoplatz	0.07	11.58	0.06	10.62	0.06	10.59	0.04	9.04	0.06	9.93	0.05	8.66

Table 1: Parameters estimated with the random sampling (size=20)

	Full choice set		**PCA Cl. 1**		**PCA Cl. 2**		**PCA Cl. 3**		**PCA Cl. 4**		**PCA Cl. 5**	
Parameters	*Val.*	*t-test*	*Val.*	*t-test*	*Val.*	*t-test*	*Val.*	*t-test*	*Val.*	*t-test*	*Val.*	*t-test*
access	0.51	6.33	0.21	3.38	0.31	4.38	0.71	5.32	0.30	4.39	0.82	5.78
childdensity	-0.05	-2.18	-0.04	-3.10	-0.04	-2.38	-0.05	-2.79	-0.04	-2.92	-0.04	-2.75
distwork	-0.14	16.84	-0.08	-13.6	-0.09	-12.4	-0.09	-11.9	-0.07	-13.2	-0.09	-11.5
popyoung	0.02	10.32	0.01	9.08	0.02	10.45	0.01	8.41	0.01	8.65	0.01	9.12
rentratio	1.23	-4.86	-0.99	-4.26	-0.81	-3.95	-0.93	-4.12	-0.98	-4.59	-0.93	-4.32
taxindex	-0.02	-4.09	-0.01	-3.17	-0.01	-2.80	-0.01	-3.28	-0.01	-3.14	-0.01	-4.19
timetoplatz	0.07	11.58	0.04	7.63	0.05	7.17	0.05	8.47	0.03	7.01	0.06	8.98

Table 2: Parameters estimated with the PCA Cluster Sampling (size=20)

	True	Random Sampling		PCA Cluster Sampling	
		Mean	*Diff. abs.*	*Mean*	*Diff. abs.*
access	0.518	0.680	0.162	0.292	0.226
childdensity	-0.052	-0.033	0.019	-0.046	0.006
distwork	-0.142	-0.075	0.067	-0.089	0.053
popyoung	0.018	0.014	0.004	0.016	0.002
rentratio	-1.227	-0.859	0.368	-0.988	0.239
taxindex	-0.015	-0.015	0	-0.016	0.001
timetoplatz	0.073	0.052	0.021	0.053	0.020
Total			**0.641**		**0.549**

Table 3: Differences between the mean of the parameters calculated for the reduced choice sets and the true values (size=20)

Parameters	Random Sampling	PCA Cluster Sampling
Access	0.04500	0.05000
Childdensity	0.00000	0.00000
Distwork	0.00000	0.00000
Popyoung	0.00000	0.00000
Rentratio	0.01100	0.00500
Taxindex	0.00000	0.00000
Timetoplatz	0.00000	0.00000
TOTAL	**0.05600**	**0.05500**

Table 4: Variance of parameters across the 5 runs (size=20)

Parameters	**Random Sampling**	**PCA Cluster Sampling**
Access	1.2311	1.0538
Childdensity	0.0957	0.0449
Distwork	0.3384	0.2793
Popyoung	0.0217	0.0172
Rentratio	1.8357	1.4844
Taxindex	0.0052	0.0169
Timetoplatz	0.1035	0.1299
TOTAL	**3.6313**	**3.0264**

Table 5: Total difference between true values and all the parameters computed for the reduced choice-set (size=20)

	Random Sampling	**PCA Cluster Sampling**
Bias	0.47782	0.36800
Total Error	2.91202	1.84553
Variance	0.01496	0.01384

Table 6: Ability to replicate the choice probability

	Random Sampling	**PCA Cluster Sampling**
Bias	1460.96	1016.37
Total Error	7304.824	5081.857
Variance	141286.41	6948.96

Table 7: Ability to estimate the overall log-likelihood function