Sampling of alternatives for spatial choice modelling

Andrew Daly*
Institute for Transport Studies
University of Leeds, Leeds, UK
Thijs Dekker, Stephane Hess
Institute for Transport Studies
University of Leeds, Leeds, UK
* E-mail: daly@rand.org

Abstract:

In various empirical applications it is necessary to estimate choice models with substantial numbers of alternatives. In mode and destination choice models, for example, individuals face a wide range of spatially distributed destinations and a set of modes by which they can travel to the respective destination. Since the calculation of choice probabilities requires consideration of all the alternatives in the choice set, tens of thousands or more, such models can make heavy demand on computer resources, particularly run time, but also potentially storage requirement. This problem increases substantially when making use of more advanced types of models, such as random coefficients models, or systems of models, such as in activity based modelling. Independently of the specific model structure used, an option to reduce these demands is to look for some way of reducing the number of alternatives actually used in the estimation by sampling methods.

By definition, sampling of alternatives leads to increased error in parameter estimation, of which the magnitude is not completely straightforward to calculate. McFadden [2] set out the Positive Conditioning (PC) property under which consistent estimates of a Multinomial Logit (MNL) can be obtained using sampling of alternatives. Estimation under sampling of alternatives with PC sampling procedures requires maximisation of a modified likelihood function with an added correction term in the utility function. More recently, Guevara and Ben-Akiva [1] extend the work of McFadden [2] to the GEV framework, such that consistent estimates can be obtained for nested logit models, being either tree-nested or cross-nested. Since the denominator of the logit formula and the additional GEV term both comprise the logsum, sampling needs to be done at two points in the model. It is important to note that the sampling procedure used to estimate the logsum need not be the same for both logsums. In this paper, we further explore the research initialised by Guevera and Ben-Akiva [1], but providing a framework to provide consistent parameter estimates under random coefficients models falls beyond the scope of this paper.

Even though PC sampling procedures may provide consistent estimates, the increase in error introduced by these procedures is not known at all well. Nerella and Bhat [3] give an indication of the magnitude of the error, both for MNL and for more complicated models, such as mixed logit. For

MNL, they give some guidelines on minimum sample size to achieve stability, but that is with a simple sampling strategy in simulated data and so not likely to be transferable to real data and more sophisticated sampling procedures. In particular, the efficiency of sampling can vary substantially between contexts and sampling procedures. For mixed logit models, Nerella and Bhat (2004) show that the results are less accurate, but in this case the estimates are not necessarily (proven to be) consistent and would not usually be used in practice. Summarising the current state of the literature, we have to conclude that sampling alternatives in MNL causes noise, but we would not be able to state in advance what that noise would be in a specific situation; sampling alternatives in mixed logit causes more noise and bias may also be present, but we are not able to say how much of either.

In the paper, we extend the current state of knowledge on the impact of alternative PC sampling procedures and the resulting sampling error in MNL and GEV models. We particularly focus on travel demand modelling, based on mode and destination choice models. Typically, in these types of models individuals are faced with various modes and destination choices of which the choice probability is heavily affected by the travel accessibility from a specific origin. We investigate the way in which different distributions of choice probability over the alternatives, as would occur with variations in mode choice and trip length for different travel purposes, affects the effectiveness of different PC sampling schemes and estimation procedures. Like Nerella and Bhat [3] we use simulated data, but with a clear focus on applicability. Hence, the aim is to provide more transferable results on the arising sampling error.

Our focus differs from Guevara and Ben-Akiva [1], who apply their framework in a residential location choice model. The properties of mode and destination choice models are different from residential choices, because the choice probability is heavily affected by the travel accessibility from a specific origin. This is in clear contrast to residential choice models, where accessibility is of less importance as individuals move on an infrequent basis. A clear aim of the paper is therefore to test the impact of alternative PC sampling in such a setting on the resulting sampling error in the Guevara and Ben-Akiva [1] framework. In short, using simulated data we evaluate the efficiency and effectiveness of the Guevara-Ben-Akiva approach by exploring various PC sampling schemes and sample sizes in an attempt to minimise the estimation error for a given computational burden. The alternative sampling strategies are applied at the two places where sampling is required in the GEV model. Furthermore, we test sensitivities using the two approaches recommended by Guevara and Ben-Akiva, being (i) separate sampling procedures and (ii) using the same samples for both logsums. In the former case, the sampled approximation of the logsum does not depend on the chosen alternative. In the latter the sample is constructed around the chosen alternative. Hence, an iterative estimation procedure is required due to dependence of the expansion factors on the model parameters.

Both approaches seem to work well in the presented results by Guevara and Ben-Akiva, but additional insights in the resulting sampling error are vital information in this relatively new field of research. The current paper is therefore of interest to travel demand modellers in general, but in particular those concerned with mode and destination choice or activity-based modelling, in which very large numbers of alternatives typically occur.

References:

[1] Guevara, C. and Ben-Akiva, M., “Sampling of alternatives in multivariate extreme value (MEV) models”, Word Conference on Transport Research, Lisbon, 2010.

[2] McFadden, D.L., “Modelling the choice of residential location”, in Karlqvist, A., Lundqvist, L., Snickars, F. and Weibull, J., Spatial interaction theory and residential location, North-Holland, pp. 75-96, 1978.

[3] Nerella, S. and Bhat, C., “A numerical analysis of the effect of sampling of alternatives in discrete choice models”, Tranportation Research Record: journal of the transportation research board, 1894, 11-19, 2004.