Journal of Environmental Economics and Management 90 (2018) 101–118

Contents lists available at ScienceDirect

Journal of Environmental Economics and Management

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

# Estimation and welfare analysis from mixed logit models with large choice sets

Roger H. von Haefen [a], Adam Domanski [b, *]

[a] *Department of Agricultural and Resource Economics and Center for Environmental and Resource Economic Policy, North Carolina State University, NCSU Box 8109, Raleigh, NC, 27695-8109, USA*
[b] *ECONorthwest, 1200 Sixth Ave, Suite 615, Seattle, WA, 98101, USA*

A R T I C L E I N F O

*Article history:*
Received 23 October 2017
Received in revised form 6 April 2018
Accepted 2 May 2018
Available online 21 May 2018

*JELcodes:*
C25
Q51

*Keywords:*
Discrete choice
Sampling of alternatives
Welfare measurement
Recreation

A B S T R A C T

We show how McFadden's sampling of alternatives approach and the expectation-maximization (EM) algorithm can be used to consistently estimate latent-class, mixed logit models in applications with large choice sets. We present Monte Carlo evidence confirming our approach works well in small samples, apply the method to a dataset of Wisconsin angler site destination choices, and report welfare estimates for several policy scenarios. Of interest to applied researchers, our results quantify the tradeoffs between model run-time, accuracy, and precision of welfare estimates associated with samples of different sizes. Moreover, although our results confirm that larger efficiency losses arise with smaller samples as theory would predict, they also suggest that depending on researcher needs, random samples as small as 28 alternatives (5% of the full set of alternatives in our application) can produce relatively accurate welfare estimates that are useful for exploratory modeling, sensitivity analysis, and policy purposes.


## 1. Introduction

Environmental economists frequently use discrete choice models to analyze household decisions. These models are particularly useful when the choice set households face is large and substitution is important. However, an implementation problem arises when the choice set is very large (on the order of hundreds or thousands of alternatives) as computational limitations can make estimation difficult, if not intractable. These complications arise, for example, when modeling recreational site-choice behavior (Deepwater Horizon NRDA Trustees, 2016), automobile choice (Bento et al., 2009) and residential demand (Bayer et al., 2007).[1] McFadden (1978) suggested that using a sample of alternatives in estimation can obviate these difficulties and produce consistent parameter estimates. His approach has been widely used in environmental applications (e.g., Klaiber and Phaneuf, 2010; Tra, 2010; Sinha et al., 2017). When implementing the sampling of alternatives approach, researchers typically assume that unobserved utility is independently and identically distributed type I extreme value.

* Corresponding author.
*E-mail addresses:* roger_von_haefen@ncsu.edu (R.H. von Haefen), domanski@econw.com (A. Domanski).

[1] In particular, the Deepwater Horizon NRDA Trustees' (2016) data set included roughly 41,000 observations with coastal recreation trip data for over 2000 miles of Gulf coastline; Bayer et al. (2007) San Francisco Census data set included precise housing locational choices for 650,000 Bay area residents, and Bento et al. (2009) had disaggregate make, model and year car holdings data (over 2000 alternatives) for roughly 26,000 households. To different degrees, the choice alternatives were aggregated to alleviate computational concerns - 83 coastal sites in the Deepwater context, almost 39,500 Census block "neighborhoods" in the San Francisco housing choice context, and roughly 250 car types in the automobile choice context.

https://doi.org/10.1016/j.jeem.2018.05.002
0095-0696/

Independence implies that the odds ratio for any two alternatives does not change with the addition of a third alternative. This property, known as the independence of irrelevant alternatives (IIA), is necessary for consistent estimation under the sampling of alternatives approach, but is often a restrictive and implausible characterization of choice.

In recent years, applied researchers have developed several innovative models that relax IIA. Perhaps the most notable and widely used is the mixed logit model (Train, 1998) which introduces unobserved preference heterogeneity through model parameters. This variation allows for richer and more plausible substitution patterns and thus makes the mixed logit model an attractive tool for discrete choice modeling. However, there is limited evidence that McFadden's sampling of alternatives approach can be used in this context. Guevara and Ben-Akiva (2013) prove that the sampling approach generates consistent parameter estimates when the size of the sampled choice set asymptotically approaches the full choice set. Whether this asymptotic result is useful when the choice set used in estimation is small – as is often necessary for computational reasons – is unclear.[2]

In this article, we demonstrate how McFadden's sampling of alternatives approach and the expectation-maximization (EM) algorithm can be used to estimate latent class (or finite mixture) mixed logit models with large choice set applications. The latent class framework probabilistically assigns individuals to classes, where preferences are heterogeneous across – but homogeneous within – classes. This approach allows the researcher to recover separate preference parameters for each consumer type[3] without assuming a more parametric mixing distribution. As first demonstrated by Swait (1994), latent class models can be conveniently estimated with the recursive EM algorithm. Doing so transforms estimation of the non-IIA mixed logit model from a one-step computationally intensive estimation into recursive estimation of IIA conditional logit models. By reintroducing the IIA property at each maximization step of the recursion, sampling of alternatives generates asymptotically consistent parameter estimates.

Moreover, we report results from a detailed Monte Carlo simulation that suggests our approach also works well in finite samples. Using the simulation results as guidance, we then empirically evaluate the welfare implications of our strategy using a recreational dataset of Wisconsin anglers. The Wisconsin dataset is attractive for this purpose because it includes a large number of recreational destination alternatives (569 in total) that allows us to test the sampling of alternatives approach against estimation using the full choice set. By comparing estimates generated with the full choice set to estimates generated with samples of alternatives of different sizes,[4] we can compare the benefits and costs of sampling of alternatives in terms of estimation run time, small sample bias, and efficiency loss. Our strategy involves repeatedly running latent class models on random samples of alternatives of different sizes. In particular, we examine the effects of using sample sizes of 285 (50%), 142 (25%), 71 (12.5%), 28 (5%), 11 (2%), and 6 (1%) alternatives. Our results suggest that for our preferred latent class specification, using a 71-alternative sample size will generate on average a 75% time savings and 51% increase in the 95% confidence intervals for the five willingness-to-pay measures we construct. We also find in our application that models with sample sizes as small as 28 alternatives may be sufficiently informative for policy purposes, but that smaller sample sizes often generate point estimates with relatively large confidence intervals.

These results provide useful guidance for researchers, policymakers, and other practitioners interested in estimating models with large choice sets. The overall time saved during estimation can allow researchers to model a broader and more complex set of specifications. Additionally, time-saving techniques enable practitioners to explore alternative specifications before dedicating limited resources to estimating final models. This flexibility can be especially useful in the early stages of data analysis when the researcher's goal is to quickly identify promising specifications deserving further study. And while processor speed is constantly improving, this method will always be available to estimate models at or beyond the frontier of computing power.

## 2. Discrete choice models

This section reviews the conditional logit model, the IIA assumption, and the mixed logit model with continuous and discrete mixing distributions. We begin by briefly discussing the generic structure of random utility maximization (RUM) discrete choice models in the context of recreation demand modeling.

The central building block of discrete choice models is the conditional indirect utility function, $U_{ni}$, where $n$ indexes individuals and $i$ indexes alternatives. A common assumption in empirical work is that $U_{ni}$ can be decomposed into two additive components, $V_{ni}$ and $\varepsilon_{ni}$. $V_{ni}$ embodies the determinants of choice such as travel cost, site characteristics, and demographic/site characteristic interactions that the econometrician observes as well as preference parameters. In most empirical applications, a linear functional form is assumed, i.e., $V_{ni} = \beta_n \boldsymbol{x}_{ni}$ where $\boldsymbol{x}_{ni}$ are observable determinants of choice and $\beta_n$ are preference parameters that may vary across individuals. $\varepsilon_{ni}$ captures those factors that are unobserved and idiosyncratic from the analyst's perspective. Under the RUM hypothesis, individual $n$ selects recreation site $i$ if it generates the highest utility

[2] Guerava and Ben-Akiva (2013) also show how sampling of alternatives can be used with nested logit models, which relax IIA across nests but maintain IIA within nests. Although it is well known that these models can be estimated sequentially using samples of alternatives at each stage (e.g., Ben-Akiva and Lerman, 1985), Guerava and Ben-Akiva demonstrate how they can be estimated more efficiently in a one-step, full-information framework.

[3] Consumer types can be driven by any combination of attitudinal, spatial, demographic, or other variation in the population.

[4] Throughout the paper unless otherwise stated, "sample size" will refer to the sample of alternatives, not the sample of observations.