Transportation Research Part B 000 (2017) 1–26

Contents lists available at ScienceDirect

Transportation Research Part B

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

Random taste heterogeneity in discrete choice models: Flexible nonparametric finite mixture distributions

Akshay Vij[a,*], Rico Krueger[b]

[a] Institute for Choice, University of South Australia, Level 13, 140 Arthur Street, North Sydney, NSW, 2060, Australia
[b] Research Centre for Integrated Transport Innovation, School of Civil and Environmental Engineering, University of New South Wales, Sydney, NSW, 2052, Australia

ARTICLE INFO

Article history:
Received 4 March 2016
Revised 23 October 2017
Accepted 25 October 2017
Available online xxx

ABSTRACT

This study proposes a mixed logit model with multivariate nonparametric finite mixture distributions. The support of the distribution is specified as a high-dimensional grid over the coefficient space, with equal or unequal intervals between successive points along the same dimension; the location of each point on the grid and the probability mass at that point are model parameters that need to be estimated. The framework does not require the analyst to specify the shape of the distribution prior to model estimation, but can approximate any multivariate probability distribution function to any arbitrary degree of accuracy. The grid with unequal intervals, in particular, offers greater flexibility than existing multivariate nonparametric specifications, while requiring the estimation of a small number of additional parameters. An expectation maximization algorithm is developed for the estimation of these models. Multiple synthetic datasets and a case study on travel mode choice behavior are used to demonstrate the value of the model framework and estimation algorithm. Compared to extant models that incorporate random taste heterogeneity through continuous mixture distributions, the proposed model provides better out-of-sample predictive ability. Findings reveal significant differences in willingness to pay measures between the proposed model and extant specifications. The case study further demonstrates the ability of the proposed model to endogenously recover patterns of attribute non-attendance and choice set formation.

1. Introduction

The field of discrete choice analysis has long wrestled with the question of how best to represent heterogeneity in the decision-making process (for a recent review of the literature, the reader is referred to Yuan et al., 2015). In cases where tastes vary systematically with observable variables, heterogeneity may be captured through interactions between observable characteristics of the decision-maker and observable attributes of the alternatives. However, capturing heterogeneity systematically may be insufficient when tastes vary with unobservable variables or purely randomly, and can result in inconsistent parameter estimates (Chamberlain, 1980). In such cases, heterogeneity in the decision-making process may be captured through additional interactions between observable variables and the stochastic component.

* Corresponding author.
E-mail addresses: Akshay.Vij@unisa.edu.au (A. Vij), r.krueger@student.unsw.edu.au (R. Krueger).

https://doi.org/10.1016/j.trb.2017.10.013
0191-2615/© 2017 Published by Elsevier Ltd.

Mixed logit is by far the most popular discrete choice model for incorporating random taste heterogeneity. The model specifies choice probabilities as a mixture of multinomial logit probabilities. McFadden and Train (2000) show that mixed logit can approximate choice probabilities given by any discrete choice model derived from the theory of random utility maximization to any desired level of closeness. However, the "theorem is an existence proof only and does not provide guidance for finding the mixing distribution that attains an arbitrarily close approximation" (Train, 2008).

Most mixture distributions employed in practice can broadly be classified as one of two types: parametric or nonparametric. Parametric distributions have clearly defined functional forms with a fixed number of parameters, as in the case of a normal or lognormal distribution. Though parametric distributions often provide an excellent fit to the data, they are limited by their functional form in the shapes that they can assume. There is no prior best distribution; studies usually estimate models with different distributions, and the most appropriate distribution is determined through a comparison across goodness of fit measures and behavioral interpretation. Due to the many different distributions that can possibly be estimated, the process can be labor-intensive, and any search for the most appropriate distribution is necessarily ad hoc in that "not every possible distribution can be considered" (Keane and Wasi, 2013).

Nonparametric distributions do not have well-defined functional forms with a fixed number of parameters. Rather, the number of parameters increases and the functional form grows in complexity as more data become available, as in the case of latent class choice models (LCCMs). Seminonparametric distributions are closely related to nonparametric distributions, and may be thought of as nonparametric mixtures of parametric distributions, such as finite mixtures of normal distributions (used, for example, by Train, 2008 and Fosgerau and Hess, 2009). Nonparametric and seminonparametric distributions can asymptotically mimic any multivariate distribution, but computational constraints generally preclude the estimation of models with a high degree of complexity, and models estimated in practice tend to be simpler abstractions that are unable to leverage the flexibility offered by the general framework.

The objective of this study is to develop a fully flexible and computationally tractable way of incorporating random taste heterogeneity within existing discrete choice models. We propose using a mixed logit model with a nonparametric mixture distribution, where the support of the distribution is specified as a high-dimensional grid over the coefficient space, with equal or unequal intervals between successive points along the same dimension, and the location of each point on the grid and the probability mass at that point are model parameters that need to be estimated. First proposed by Dong and Koppelman (2014), the framework does not require the analyst to specify the shape of the distribution prior to model estimation, and the specification can approximate any multivariate probability distribution function to any arbitrary degree of accuracy. However, empirical applications in the literature have thus far been limited to distributions over low-dimensional coefficient spaces with a small number of mass points, likely due to the computational burden imposed by standard gradient-based maximum likelihood estimation routines employed by the study, and the value of the framework over other parametric and nonparametric distributions is unclear. We outline an expectation maximization (EM) algorithm for the estimation of the proposed model that is able to estimate behaviorally meaningful specifications over high-dimensional coefficient spaces with hundreds of mass points. The performance of the estimation algorithm is tested using multiple synthetic datasets, and the benefits of the model specification are evaluated using a case study on travel mode choice behavior.

Despite significant recent advances in discrete choice methods, the question of how best to incorporate random taste heterogeneity has remained an open line of enquiry. Since 2010 alone, we are aware of at least eleven new methods that have been proposed in the literature (c.f. Bastin et al., 2010; Bujosa et al., 2010; Fiebig et al., 2010; Bhat and Sidharthan, 2012; Bastani and Weeks, 2013; Greene and Hensher, 2013; Keane and Wasi, 2013; Dong and Koppelman, 2014; Train, 2016; Bansal et al., 2017; Bhat and Lavieri, 2017). However, for reasons expanded upon in the following section, each of these methods has proven inadequate in one way or another. This study contributes to the literature through the development of a model framework and estimation algorithm that can help overcome some of the constraints arising from these and other specifications. In particular, compared to extant multivariate nonparametric specifications, we show how our proposed approach of specifying the parameter space as a high-dimensional grid with unequal intervals can offer considerable improvements in flexibility, while requiring the estimation of a small number of additional parameters.

The remainder of the paper is structured as follows: Section 2 reviews the relevant literature on the use of mixture distributions for discrete choice analysis; Section 3 presents the methodological framework; Section 4 describes how the framework may be estimated in practice using the EM algorithm; Section 5 details findings from three Monte Carlo experiments evaluating the ability of the proposed framework to recover known parametric univariate, bivariate and high-dimensional multivariate distributions, respectively; Section 6 applies the framework to a study on travel mode choice behavior, and compares findings from the framework with extant approaches for incorporating random taste heterogeneity; Section 7 examines the robustness of the estimation method to different starting values for the model parameters; and Section 8 concludes the paper with a summary of key results and directions for future research.

2. Literature review

Section 2.1 surveys studies that have used parametric mixture distributions and Section 2.2 surveys studies that have used nonparametric and seminonparametric mixture distributions. Throughout, we identify advantages and disadvantages to each of the many different mixture distributions that have been used in the literature, and how they motivate the particular framework employed by this study.