Original Article

Comparison of continuous and discrete representations of unobserved heterogeneity in logit models

Received (in revised form): 19th April 2014

Xiaojing Dong

is Associate Professor of Marketing at the Leavey School of Business, Santa Clara University. In her research, using customer-level data, she develops empirical models to study how individual customers/consumers make decisions, and evaluate how marketing efforts influence those decisions. Her research has covered a selection of areas and industries, including pharmaceutical, financial and online social networks.

Frank S. Koppelman

is Professor Emeritus of Civil and Environmental Engineering at Northwestern University. He has more than 40 years of experience in the development and application of models of traveler behavior and traveler satisfaction with existing or proposed transportation and other public and private services. His work has been widely applied to both transportation and other services. He is currently Chairman of the Ridership Review Panel for the California High Speed Rail Administration.

Correspondence: Xiaojing Dong, Leavey School of Business, Santa Clara University, 500 El Camino Real, Lucas Hall, Santa clara, California 95053, USA
E-mail: Xdong1@scu.edu

ABSTRACT Representing unobserved heterogeneity or taste variations in Marketing Analytic behavioral-choice analysis is receiving increasing attention in the estimation of consumer-choice modeling. The mixed logit (MXL) model, which incorporates random coefficients into the multinomial logit model, has been widely adopted for this purpose. The most commonly adopted method in this context is to assume that the random coefficient follows a continuous, unimodal distribution, and the parameters of the distribution as well as the other parameters for the model can be obtained using maximum simulated likelihood estimation. In this article, we refer to this method as the continuous mixed logit (CMXL) model. This method requires the *a priori* assumption that the distribution of the random coefficient is continuous and, usually, unimodal. One way to relax this assumption is to estimate the distribution nonparametrically, by assuming a discrete distribution with finite support. We refer to this approach as the discrete mixed logit (DMXL) model. Based on the DMXL model, we propose the mass-point MXL model as one alternative to the continuous-distribution assumption and compare its performance with the latent class logit model (LCLM), also part of the DMXL family. Either model can be used to represent unobserved heterogeneity with a discrete distribution in the parameter space. In this article, we conduct empirical analyses and compare the continuous and discrete representations of unobserved heterogeneity in logit models using simulated data with known parameters and real data with discrete choices. Analysis with simulated data provides insights on the ability to distinguish between continuous and discrete parameter distributions and a better understanding of the goodness-of-fit measures used in evaluating model performance with real data. From the simulation study, we find that when the data is generated from a normal distribution, the CMXL model with the unimodal-distribution assumption is preferred to the DMXL mode.

From the real data analysis, we find that the CMXL model fails to recover heterogeneity that is identified by the DMXL model. In conclusion, we suggest that when estimating a random-coefficient MXL model, one should start with a CMXL model, but should not accept a 'no heterogeneity' conclusion without estimating a series of DMXL models using either the mass-point MXL model or the LCLM with different starting values.
Journal of Marketing Analytics (2014) **2,** 43–58. doi:10.1057/jma.2014.5

Keywords: Logit model; unobserved heterogeneity; maximum simulated likelihood estimation; discrete distribution

INTRODUCTION AND BACKGROUND

Representing unobserved heterogeneity in utility assessment has become an important consideration in the analysis of people's choice behavior in Marketing Analytics. Numerous papers have documented that failure to account for heterogeneity in logit models leads to biased parameter estimates (for example, Chamberlain, 1978; Heckman, 1981; Heckman and Singer, 1984; Crouchley, 1987; Reader, 1993) or biased probabilities (for example, Horowitz, 1981; Jones and Landwehr, 1988). Mixed logit models (MXL) have been widely adopted for the purpose of incorporating heterogeneity in logit choice analysis (McFadden and Train, 2000) using random coefficients. In most data describing transportation-mode choices, only one choice is recorded for each individual. As a result, we cannot obtain individual-specific parameters. However, it is still possible to capture the differences among people by assuming that the model parameters follow some distribution, instead of a point value as in the closed form generalized extreme value models (McFadden, 1978; Wen and Koppelman, 2001; Train, 2003), such as the multinomial logit (MNL) and nested logit (NL) models which have been widely used to describe choice behavior in a variety of domains during the last 20–30 years.

In most of the current MXL-model implementations, the heterogeneity in some parameters is assumed to follow a continuous distribution. We call this the continuous mixed logit (CMXL) model. Estimation is based on maximizing a simulated likelihood function, called the maximum simulated likelihood estimation (MSLE) method (McFadden and Train, 2000; Train, 2003). The analyst chooses the shape of the distribution and the estimation algorithm obtains the parameters of the distribution simultaneously with the other model parameters. The most commonly assumed distributions – normal, lognormal, triangular or uniform – can be represented with two parameters. An important limitation of this approach is that its estimation relies on an *a priori* assumption about the shape of the distribution, which might lead to biased estimates in cases where the underlying distribution has a substantially different shape from the assumed distribution (Heckman and Singer, 1984).

An alternative approach is to assume that the distribution is nonparametric in one or more dimensions. We call this the mass-point mixed logit (MPMXL) model, as the distribution of parameters is represented by a finite number of mass points. It is a discrete representation of heterogeneity, as the support of the distribution is discrete. This mass-point approach can be seen as closely related to the latent class model (LCLM), also called the Finite Mixture model, as it implies that the parameters influencing different individuals are associated with membership in a distinct class or group (Kamakura and Russell, 1989; Bhat, 1997). The differences between these models are discussed further in a later section. As they each represent heterogeneity using a discrete distribution, we call them discrete mixed logit (DMXL) models to distinguish them from the CMXL model.

In the following sections, we describe each of these heterogeneity representations (continuous and discrete), and we examine the performance of each method using both simulated and real data. In these analyses, we focus on the CMXL and MPMXL models; Kamakura and Russell (1989) present a detailed analysis of the LCLM. Further, we include a limited empirical comparison between the LCLM and MPMXL models using the real data.

CONTINUOUS AND DISCRETE UNOBSERVED HETEROGENEITY IN MIXED LOGIT MODELS

The general form of a MXL is

$$P_n(i) = \int \frac{\exp(X_i\beta)}{\sum_{j\in C_n} \exp\left(X_j\beta\right)} \mathrm{d}F\left(\frac{\beta}{\theta}\right) \qquad (1)$$

In this formulation, the index n denotes an individual, i and j denote alternatives, C_n denotes the set of available alternatives for the individual n, X are the explanatory variables included in the utility formulations, and the βs and θs are the coefficients to be estimated. The βs represent utility-function parameters and the θs represent the parameters of the mixture distribution(s) for some or all of the βs. The marginal probability of individual n choosing alternative i is computed by an integral over the distribution of all the possible values of the random coefficient β conditional on the distribution parameterized by θ, with the integrand the same as for a MNL model (Ben-Akiva and Lerman, 1985). This model is sometimes called a Logit-Kernel model, highlighting that the MNL is the integrand of the model (Ben-Akiva and Bolduc, 1996). A fundamental issue is the selection of the distribution, $F(\beta/\theta)$. In the following section, we describe two possible formulations for $F(\beta/\theta)$; a continuous distribution and a mass-point distribution with discrete support points. We also compare the models' computational efficiencies.

Continuous mixed logit model

When $F(\beta/\theta)$ is assumed to be a continuous distribution with probability density function $f(\beta)$, equation (1) can be rewritten as

$$P_n(i) = \int \frac{\exp(X_i\beta)}{\sum_{j\in C_n} \exp\left(X_j\beta\right)} f(\beta)\mathrm{d}\beta \qquad (2)$$

As the integral usually doesn't have a closed-form solution, a simulation approach is employed to numerically approximate the probability. Thus, equation (2) is replaced by its simulated formulation

$$P_n(i) = \frac{1}{NR}\sum_{r=1}^{NR} \frac{\exp(X_i\beta^r)}{\sum_{j\in C_n} \exp\left(X_j\beta^r\right)}$$

where β^r is the rth random draw from its distribution, $f(\beta)$, and NR is the total number of random draws. The simulated probability is computed by making a random draw of β^r from the assumed probability density distribution $f(\beta)$, and computing the logit probability conditional on this parameter value. This computation is repeated NR times, and the arithmetic mean of the NR computed MNL probabilities is the simulated value of the marginal probability, conditional on the values of θ. Using MSLE, we can obtain θ, which are the parameters for $f(\beta)$, as well as other coefficients in the logit model assumed to be homogeneous across the population.

There are two identified shortcomings of the CMXL model. First, as simulation is required in the estimation procedure, the computational time is significantly greater than that for closed-form models. Halton sequences (Train, 2000; Bhat, 2001) and nets (Niederreiter, 1992; Sándor and Train, 2002) have been developed to provide better coverage of the distribution with fewer random draws. In addition, simulation biases for parameter estimates are produced through the use of simulation. The biases are reduced by increasing the number of random draws, *NR*, which will at the same time increase the computation time. Actually, as using MSLE requires evaluating the same probability and likelihood function *NR* times, the computational complexity of this model is in the order of *NR*, denoted $O(NR)$. Given the rapid increase in available computation power, this is not a critical issue. In practice, even though long computation times are required to obtain a sufficiently large number of random draws, it is usually acceptable.

Second, the estimation of a CMXL model requires an *a priori* assumption of the distribution function, $f(\beta)$, and the estimation results are dependent on this assumption. In most cases, a unimodal distribution with two parameters is employed. Such a distribution may lead to biased estimates if it differs from the true distribution. This limitation of the CMXL model cannot be resolved by increasing the number of draws. Therefore, we need to find a way to relax the assumption of a unimodal distribution. A natural solution is to adopt a nonparametric approach in which the model parameters are assumed to be distributed with a finite number of supports, the location and probability weight of each support being identified empirically. This type of distribution is called the mass-point distribution, and the MXL model with this distribution is called the Mass-Point Mixed Logit model (MPMXL).

Mass-point mixed logit model

MPMXL models represent unobserved heterogeneity using a discrete distribution with finite support. Therefore, they belong to the family of DMXL models, as does the LCLM (Kamakura and Russell, 1989).

By replacing the probability distribution $f(\beta)$ in equation (2) with a mass-point distribution whose probability weight at the *m*th mass point is λ^m, and by replacing the integration in equation (3) with a sum over a finite number of mass points, *M*, we arrive at the formulation for the MPMXL model:

$$P_n(i) = \sum_{m=1}^{M} \frac{\exp(X_i\beta^m)}{\sum_{j\in C_n} \exp(X_j\beta^m)} \lambda^m \qquad (3)$$

In the MPMXL model, the probability of individual n choosing alternative i can be regarded as the weighted average of *M* logit probabilities computed at each possible value of β; the weights are the probability of coefficient β being equal to each value of β^m: $\lambda^m = \Pr\{\beta = \beta^m\}$. Therefore, it has a closed-form representation for the choice probability. As a result, the estimation of the model, including the parameters for the locations and probability weights can be obtained using maximum likelihood estimation, without simulation (Chintagunta *et al*, 1991). As such, the computational complexity of this model is in the order of the number of mass points, *M*, denoted $O(M)$. As *M* is usually much smaller than the number of random draws, *NR*, used in estimating the CMXL model with the simulated likelihood approach, the MPMXL model is expected to require much less computational effort than the CMXL model.

As mentioned earlier, this model is similar to the LCLM (Kamakura and Russell, 1989): in the context of the logit model, they both represent the unobserved heterogeneity in the model parameters using distributions with finite support. The difference between them lies in their degrees of freedom (or number of model parameters to estimate) required to define the same number of groups. In other words, they identify the same number of groups using different numbers of parameters. Consider an example where we try to estimate a model with two betas in the utility function of a logit model, each of which is represented by a discrete distribution defined by two location parameters. In the MPMXL model, we can think of locating the two betas on a two-dimensional grid, in each dimension there are two mass points. With this grid system, the MPMXL model requires (at most) four mass points on the grid. The LCLM requires two location parameters for each location point. Without the grid system used in the MPMXL model, the LCLM can identify only a number of location points equal to the number of parameters. In other words, the MPMXL model allows each model parameter in the logit model to have multiple discrete supports, and the total number of different supports is defined by the sum of all the discrete supports identified for each logit model parameter. In the LCLM, however, the total number of different supports (or the number of latent classes in the logit model parameters) is defined by the

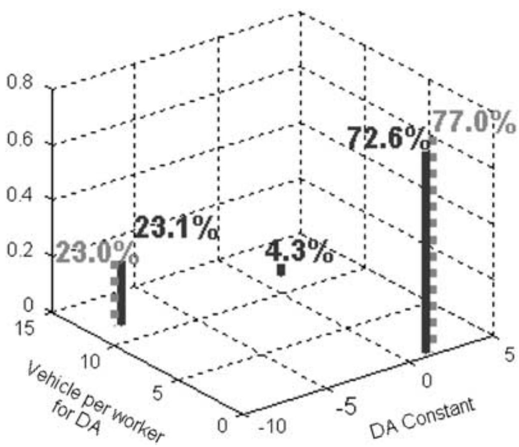

Figure 1: Estimates for CNST_DA and VPW_DA by the MPMXL model and the LCLM.

total number of discrete supports for the combination of all the model parameters. This is illustrated by the plot in Figure 1. In this plot, the MPMXL identifies three discrete supports (maximum possible four) for the logit model's two parameters, as represented by the three solid blue lines. The LCLM identifies two segments, represented by the red dotted lines, on the two model parameters. In the LCLM, the segments are identified on the multiple dimensions defined by all the logit model parameters together, so they are restricted to the number of total segments. In contrast, each model parameter in the MPMXL model is allowed to follow a marginal discrete distribution; therefore, it potentially could allow more discrete support for the model parameters, which could identify more segments than logit model parameters. The difference between the MPMXL model and the LCLM is similar to the difference between the Latent Class Cluster model and the Latent Class Factor model (Magidson and Vermunt, 2001).

When only one random coefficient is estimated with a discrete distribution, these two models are identical. Detailed discussion about the difference between these two models is presented later with the empirical analysis of the real data.

The estimation of the MPMXL model differs from that of the CMXL model in two main ways. First, it reduces the computational time significantly and is free from simulation biases as it has a closed-form formulation. With modern computational power, this may not be vital; however, as sample size or model complexity increases, the differences may become important. In practice, although the MPMXL model converges much faster than the CMXL, estimating the CMXL model with sufficient random draws is still acceptable. Second, the MPMXL model relaxes the need for an *a priori* assumption on the distribution of the parameters.[1] This added flexibility leads to a more interesting issue, which is the focus of this article.

The marketing literature has documented some comparisons between continuous and discrete representations of heterogeneity either by insights and qualitative discussions (Wedel *et al*, 1999) or based on simulated data (Andrews *et al*, 2002). According to Wedel *et al* (1999), 'To a large extent, the issue of a continuous versus discrete distribution of heterogeneity is an empirical one'. Andrews *et al* (2002) conducted extensive experiments with simulated data, in which they generated multiple observations for each household. In their paper, they manipulate different factors defining the true heterogeneity distributions as well as the number of observations for each household. Generating 360 different data sets and comparing the performance of the continuous and discrete MXL models in terms of both model fit and recovery of the true individual-specific parameters under various situations, they find that continuous and discrete heterogeneity perform equally well in terms of recovering household-specific parameters, except that when the number of observations for each household is low, the performance of the continuous distribution is very poor. These conclusions are quite interesting and relevant in the marketing literature, where most data sets contain multiple observations for each household or individual. This is different in transportation analysis where the majority of the choice data sets contain only one observation for each case (individual or household). We cannot distinguish randomness in individual behavior from differences across individuals with these data. We compare the performance of the continuous and discrete heterogeneity for cross-sectional data using a smaller (compared with the experiments conducted by Andrews *et al*, 2002) set of Monte Carlo experiments using simulated data. These simulation analyses are complemented by empirical analysis with real data on commuting behavior in the New York metropolitan area. The combination of these two analyses provides us with a full picture of the performance of continuous and discrete MXL models.

SIMULATION ANALYSIS

We generated two data sets with continuous and discrete distributions of one beta and estimate each using both CMXL and DMXL models. In the generated data sets, we include 20 000 observations, assumed to be large enough to identify a distribution of one parameter. Each observation selects one of the two alternatives with the largest indirect utility, defined as

$$U_{na} = \beta_1 x_{na,1} + \beta_2 x_{na,2} + \varepsilon_{na}$$
$$U_{nb} = \beta_1 x_{nb,1} + \beta_2 x_{nb,2} + \beta_0 + \varepsilon_{nb}$$

The two alternatives are denoted a and b, and individuals are indexed by n. The model includes two independent variables, $x_1$and x_2, for both alternatives with coefficients β_1 and β_2. As in most of the choice models with a common choice set among the decision makers (Guadagni and Little, 1983; Kamakura and Russell, 1989; Chintagunta *et al*, 1991, to name a few seminal papers), we add an intercept β_0 to the utility of one of the alternatives, alternative b. In each utility function, there is an additive random error ε_{na} for alternative a and ε_{nb} for alternative b, which are generated as IID extreme values, with parameters (0, 1). For simplicity and illustrative purposes, only β_1 is treated as random coefficient, and β_2 is fixed. β_1 is generated from a continuous distribution in one data set and a two-point discrete distribution in the other. The other coefficients are the same in both data sets: $\beta_2 = -6$, and $\beta_0 = 1$. Each generated data set is estimated with both continuous and discrete distributions for β_1. To estimate the parameters with the continuous distribution, 200 Halton draws are used to evaluate the simulated probability and likelihood. The parameters of the distribution, as well as the other fixed coefficients, including β_2 and β_0 are obtained using MSLE. In estimating the parameters with the discrete distribution,[2] the locations and probability weights for each mass point are estimated. In addition, the mean and variance for each mass-point distribution are computed, and their standard errors are obtained using the Delta method (Greene, 2000). These are presented together with the model estimates.

Data generated with a continuous random coefficient

In the first data set, the choice outcome for each observation is generated by assuming β_1 is distributed normally, with mean −6 and variance 2: $\beta_1 \sim N(-6, 2)$. The estimation results using both continuous and discrete distributions are presented next.[3]

We estimate an MNL model for the purpose of comparison, and the mixed logit model under three distinct *a priori* assumptions; that the distribution is normal, lognormal and discrete with two or three support points. The results for the MNL model and the CMXL models with normal and lognormal assumptions are reported in Table 1, together with the true values of the parameters with standard errors of the estimates in parentheses.

In the table, the estimate of the fixed parameter, β_2, for the MNL model is about three standard errors away from its true value, and the estimate for random parameter β_1 is over four standard errors away from its true mean value. The maximum likelihood estimation is asymptotically normal, based on which we can conduct an informal statistical test. The chance that $\hat{\beta}_2$ (the estimate of β_2) is equal to or less than the true value of β_2 is 0.3 per cent, computed with $\Phi((\beta_2 - \hat{\beta}_2)/(\hat{\sigma}_2))$, where $\hat{\sigma}_2$ is the asymptotic standard error of the estimate $\hat{\beta}_2$, and $\Phi()$ is the cumulative distribution function of the standard normal distribution. Using the same method, the probability that $\hat{\beta}_1$ (the estimate of β_1) is equal or less than the true value of β_1 is 0.02 per cent. In both cases, the null hypothesis that the estimated parameter equals the true value cannot be accepted at 95 per cent. This result suggests that ignoring continuous heterogeneity leads to biased estimates. This is also verified by estimating the real data as presented in the next section.

The estimation results for the CMXL model with normal distribution show that it recovers all parameters, including the distribution of β_1 and the two fixed coefficients, very closely. This is expected; the assumed distribution is the same shape as that used to generate the data. Further, another CMXL model estimation assuming the negative of a lognormal distribution, which closely approximates the normal distribution also recovers the properties of the

Table 1: Estimation results for MNL and CMXL with data generated with continuous distribution

	True values	*MNL model*		*With normal distribution*		*With log-normal distribution*	
Fixed parameters							
Constant	1	0.8977	(0.0246)	0.9327	(0.0279)	0.9328	(0.0278)
Fixed parameter β_2	−6	−5.7486	(0.0910)	−5.9776	(0.1195)	−5.9792	(0.1195)
Distribution of β_1							
Mean	−6	−5.5977	(0.0893)	−5.9457	(0.1452)	−5.9656	(0.1546)
Standard deviation	1.414	—	—	1.4730	(0.2493)	1.5877	(0.3078)
Goodness-of-fit measures							
Log-likelihood at convergence		−6489.9		−6484.0		−6484.0	
Adjusted ρ^2		0.5316		0.5320		0.5320	
BIC value		13 009.5		13 007.6		13 007.6	

Abbreviations: BIC, Bayesian information criterion; CMXL, continuous mixed logit; MNL, multinomial logit.

distribution very closely (mean and standard deviation[4]). The goodness-of-fit measures for both CMXL models are essentially identical. These results indicate that the model estimation is robust, in terms of continuous distribution selected *a priori*, if a reasonable distribution is applied.

Comparing these two CMXL models with the MNL model, the Bayesian information criterion (BIC) is used (Schwarz, 1978). Now, we would like to digress to a brief review of the BIC, as it will be used extensively in comparing different models in this article. The BIC is a function of the log-likelihood at convergence, *LL*, the number of parameters in the model, *K*, and the total number of observations *NOBS*:

$$BIC = -2 \times LL + K \times \ln(NOBS)$$

From Bayes' point of view, this is an asymptotic approximation to the marginal density of the data. From the likelihood point of view, this is a negative function of the log-likelihood at convergence (*LL*), with a penalty of the number of parameters as a function of the sample size. Consistent with the concept that a larger log-likelihood value is better, a smaller BIC value is preferred. This measure is commonly used in comparing unnested models. The model with the lowest BIC would be selected as the best.

Now let's go back to Table 1. In fact, when comparing the CMXL model with the MNL model, the likelihood ratio test is applicable. Based on the results in Table 1, both of the CMXL models reject the MNL model with a 95 per cent confidence level using the likelihood ratio test. The BIC also prefers the CMXL models to the MNL model. The Bayesian information criterion (BIC) is presented here to be consistent with later analysis, where the BIC is repeatedly used in model comparison.

Using the same data, we estimate a model assuming β_1 follows a discrete distribution. The estimation results for both two and three mass points are presented in Table 2 including the BIC values for each model.

In estimating the MPMXL model, the number of mass points is decided based on the BIC values. As the number of parameters increases, the BIC value decreases to a minimum and then starts to increase. In practice, we start by estimating a model with the lowest possible number of mass points (two in this case) and then add one mass point with each estimation, until the BIC value starts to increase (Kamakura and Russell, 1989). In our estimation, we stop at three

Table 2: Estimation results for two and three mass points for data generated with a continuous distribution

	True values	*Two mass points*		*Three mass points*	
Fixed parameters					
Constant	1	0.93	(0.03)	0.94	(0.02)
β_2	−6	−5.93	(0.12)	−6.03	(0.11)
Distribution of β_1					
Mean	−6	−5.83	(1.69)	−6.07	(0.27)
Standard deviation	1.414	1.07	(24.44)	2.05	(4.27)
Mass-point estimation					
Mass point location 1	—	−6.13	(0.31)	−5.83	(0.45)
Mass point location 2	—	−2.04	(1.57)	−14.27	(0.44)
Mass point location 3	—	—	—	−0.96	(1.94)
Mass point probability 1	—	0.93	(0.16)	0.92	(0.02)
Mass point probability 2	—	0.07	(0.16)	0.05	(0.02)
Mass point probability 3	—	—	—	0.03	(0.03)
Goodness-of-fit measures					
Log-likelihood at convergence		−6484.3		−6481.9	
Adjusted ρ^2		0.53		0.53	
BIC		13 018.1		13 033.2	

mass points, as the BIC value is higher than that for the model with two mass points. Based on the BIC values, the model with two mass points is preferred.

In Table 1, we list the results for the two models with two or three mass points. In both estimations, the parameters related to the fixed model parameters (including the intercept β_0 and the fixed parameter β_2) are recovered very well, as demonstrated in the first two rows in the table. In the next two rows of the table, the mean and standard deviation of the β_1 are derived based on the estimated mass point distribution. From these results, it seems that the model with three mass points is closer to the distribution from which the data were generated, in comparison to the model with two mass points. However, the BIC values, a common measure used in model selection in this context, suggest the opposite. This demonstrates the limitation of selecting the number of mass points based on the BIC values. In fact, as stated in Weakliem (2004), 'BIC will favor smaller models – that is, models with fewer parameters' (p. 169). The features and limitations have been well studied in the literature (for example, McQuarrie and Tsai, 1998; Burnham and Anderson, 2004; Kuha, 2004; Bollen *et al*, 2012). Based on these studies, as well as our estimation results, it is important to recognize that the BIC is an asymptotically consistent measure for model fit. In practice, it may be biased towards models with too few parameters.

To summarize our findings in this simulation study, we learn four points: First, there exists multiple optima for likelihood functions of these models; therefore, it is necessary to estimate a discrete distribution logit model with multiple starting values to test the convergence and its sensitivity to the starting value.

Second, the BIC does not necessarily work well in selecting the best model among MPMXL models. It tends to favor the model with fewer supports, as discussed in Weakliem (2004). This reflects the point raised in the book by Rossi *et al* (2005), where they pointed out that BIC is such an approximation method, although it is convenient to use, 'it can be extremely inaccurate and should be avoided whenever possible' (p. 165).

Third, when estimating a DMXL model using a mass point structure, due to the possibility of multiple solutions, it is suggested that at least two estimations with different starting values be conducted: one with the starting values focused on one mass point, and the other with the probability mass evenly distributed across all the mass points.

Finally, it could be misleading to estimate the model assuming discrete distribution if the true distribution is continuous. As a result, it is necessary to estimate the model with both the discrete and the continuous distributions before selecting the better model. In fact, comparing the best continuous model from Table 1 with the best discrete model from Table 2 indicates a strong preference for the continuous estimation results: The BIC value of the continuous model is approximately 13 007 and approximately 13 018 for the best discrete model. Thus, if the true distribution is continuous, it is best approximated by a similarly shaped continuous distribution.

Data generated with a discrete distribution

The second simulated data set is based on β_1 generated from a two-point discrete distribution, with values at −8 and −4 and probability masses 40 and 60 per cent, respectively. The estimation procedures are similar to those presented in the preceding section. The results of all the model estimates, including MNL, CMXL and DMXL models, are reported in Table 3.

Similar to the results listed in Table 1, the estimates for MNL model have smaller standard errors than the corresponding parameters for all the other models in Table 3. The parameter

Table 3: Estimation results for data generated with discrete distribution

	True values	*MNL model*		*Estimation with a continuous distribution*		*Estimation with two mass points*		*Estimation results with three mass points (different starting values)*			
Fixed parameters											
Constant	1	0.88	(0.02)	0.93	(0.02)	0.93	(0.02)	0.95	(0.03)	0.93	(0.03)
β_2	–6	–5.51	(0.09)	–5.84	(0.11)	–5.84	(0.11)	–5.92	(0.12)	–5.84	(0.11)
Distribution of β_1											
Mean	–5.60	–4.93	(0.09)	–5.41	(0.13)	–5.43	(0.58)	–5.70	(0.92)	–5.43	(0.59)
Standard deviation	1.96	—	—	1.78	(0.21)	1.85	(9.31)	3.19	(17.26)	1.85	(9.78)
Mass point estimation											
Mass point location 1	–8	—	—	—	—	–7.54	(0.53)	–6.88	(0.72)	–7.54	(0.87)
Mass point location 2	–4	—	—	—	—	–3.81	(0.20)	–3.59	(0.47)	–3.81	(0.45)
Mass point location 3	—	—	—	—	—	—	—	–25.18	(7.54)	–3.81	(6.17)
Mass point probability 1	0.4	—	—	—	—	0.44	(0.13)	0.47	(0.21)	0.44	(0.04)
Mass point probability 2	0.6	—	—	—	—	0.56	—	0.51	(0.20)	0.54	(0.04)
Mass point probability 3	—	—	—	—	—	—	—	0.02	—	0.02	—
Goodness-of-fit measures											
Log-likelihood at convergence		–6873.0		–6859.3		–6858.1		–6855.5		–6858.2	
Adjusted ρ^2		0.50		0.50		0.50		0.51		0.51	
BIC value		13 775.7		13 758.2		13 765.7		13 780.4		13 785.5	

estimates for the fixed coefficient, β_2, and random coefficient, β_1, are both biased, based on the results listed in the first two rows of both Table 2 and Table 3. Combined with the biased estimators for the MNL model in Table 2, this suggests that ignoring heterogeneity (either continuous or discrete in nature) leads to biased model estimates.

Based on the estimation results presented in Table 2, the two DMXL models (two mass points and three mass points) are both estimated with a large number of different starting values. It is interesting to find that with two mass points, the estimation always converges at the same results. This indicates that the mass points' distribution can be recovered fairly well, and the estimation results are robust, when the distribution assumption is accurate.

The DMXL model with three mass points is also estimated using a large number of starting values. The estimation results all converge to one of two possible results, which are both listed in Table 3. In the first estimation, it identifies three mass points with one of them located far away at −25.18 from the other two, with low probability weight at about 2 per cent. As expected with a low probability weight, the standard error for this location estimate is very high. The other two locations and probability estimates are close to the true mass-point distribution, with much larger standard errors on the estimates for probability weights.

The second estimation results are very different. In this result, the distribution is identified by one mass point at −7.54 with probability 0.44, which is close to the first mass point in the true distribution: with probability 0.4 at location −8; and the estimated second and third points are almost at the same location −3.81 with the total probability of 0.56 (=0.54+0.02). The combination of these two points is close to the second mass point in the true mass point distribution, which is located at −4 with probability 0.6. In short, with one redundant point, the second estimation for the three mass-points model recovers the true two mass point distribution better than the first estimation result.

The bad news is that the BIC value of this model is slightly worse than the first estimation, which could be misleading. The good news is that when the BIC value is used to decide the number of mass points, we conclude at the right result, with the two-mass-point distribution, even using different starting values.

Among the four models with random coefficients in Table 3, the estimation results for the fixed parameters, β_0 and β_2, and their standard errors are similar. The mean and the standard deviations of β_1 are obtained directly for the CMXL model, and are computed based on the model outputs for the DMXL models. The values are similar across all four models with random coefficients, but the standard errors for both mean and standard deviations are much smaller in the CMXL estimation than in the three results from the DMXL models.

Comparing the model fits between the CMXL model and the best DMXL model, we observe that the continuous model has a better (smaller) BIC value than the two discrete models. This contradicts what we would expect, as the true distribution is discrete with two mass points. There are two reasons for this result. First, the BIC measure penalizes model with more parameters. Second, the CMXL model has fewer parameters by nature, because it relies on the *a priori* assumption of the distribution form. In contrast, the DMXL model relies on the parameters to identify the distribution of the random betas. That is, the BIC value favors the parametric (continuous estimation with an assumed distribution shape) over the nonparametric (discrete estimation without any assumption on distribution formulations) model estimations. When the data were generated based on a discrete distribution, we expect the discrete model to have a better value of model fit measure than the continuous model. The fact that this does not occur suggests the need for further investigation about the adoption of a goodness-of-fit measure to fairly compare models using *a priori* assumptions with models estimated nonparametrically. This also suggests that, in practice, if the BIC value favors a DMXL model over a CMXL, we can conclude with confidence that the DMXL model fits the data better than the CMXL model.

In this section, we discussed and compared the performance of both continuous and discrete representations of unobserved heterogeneity in logit model using two generated data sets. The advantage of employing the simulated data sets is that we know the 'truth', so that we can evaluate and compare the performances of different heterogeneity assumptions for estimations. From this analysis, we also obtain some cautions and tricks useful for analyzing real data. The disadvantage of using simulated data is that it is not real. In the next section, we implement the two heterogeneity representations on a real data set, based on the cautions and tricks and comparisons of their performances are discussed.

EMPIRICAL COMPARISON OF THE TWO MODELS

This analysis considers the work on trip-mode choice in the New York metropolitan area.[5] It contains work-trip information for 6844 individuals who traveled to work during the peak morning-traffic period. The information includes the chosen mode, trip origin, destination, travel service and cost characteristics for each available mode and the socioeconomic characteristics of the travelers and their households. The observed mode shares for the five modes considered are shown in Table 4.

The objective in this empirical analysis is to apply both the continuous and discrete methods of representing the unobserved heterogeneity of this data set and compare

Table 4: Mode shares in the data

Travel mode for work trip	*Mode share (%)*
Drive alone	63.6
Shared ride	13.3
Taxi	1.2
Transit (bus and subway)	17.8
Commuter rail	4.1

both the models' goodness-of-fit measures and the implications for the travelers' behavior. We started by looking at those variables considered most likely to be valued differently in the population: travel time and cost. However, the parameters of these two variables show no heterogeneity using any of the above-mentioned methods. This may result from the commonality of the population under study; morning-peak work commuters in the New York metropolitan area.

Potentially, all the parameters in the model could be assumed to be heterogeneous. In this analysis, for the purpose of demonstration, we estimate the models with both continuous- and discrete-distribution assumptions for only two parameters, the alternative specific intercept for drive alone (CNST_DA) and the alternative specific variable 'vehicles per worker' for drive alone (VPW_DA), to identify unobserved heterogeneity in people's preference for the 'drive alone' mode. We started by estimating an MNL model as a point of reference, followed by a CMXL model with a binomial normal distribution assumption. With the discrete assumption, we estimate two models: the MPMXL model and the LCLM. The estimates for these four models (MNL, CMXL, MPMXL and LCLM) are reported in Table 5.

The first group of parameters in Table 5 represents the mean, standard deviation and correlation for the drive-alone constant and 'vehicles per worker for drive alone' for each model. The second group represents variables constrained to be homogeneous over the population, including some explanatory variables. Finally, the third group represents alternative specific constants. The last section of the table contains the goodness-of-fit measures, including log-likelihood at convergence, number of parameters in each model, adjusted rho squared and BIC.

The estimation results for the fixed parameters (in the second and third groups) are consistent across these different models, suggesting that the differences in the models' fits result mainly from the different ways of incorporating taste variations into the models. In the following sections, we discuss the comparisons among these models based on the empirical results in terms of both the model estimates for CNST_DA and VPW_DA and the model fits.

The multinomial logit and continuous mixed logit models

The MNL model is estimated using the maximum likelihood method and the CMXL model is estimated using MSLE with a binomial normal distribution assumption on the two parameters. There are three points that are of particular interest in comparing these two models. First, the CMXL model obtains a nonsignificant estimate for the standard deviation for the DA constant, which suggests it is homogeneous for this population. Second, the CMXL model gets a significant estimate for the standard deviation of the VPW_DA variable and a very different mean estimate from that in the MNL model, which indicates that ignoring heterogeneity results in biased estimates for the parameter mean. This aligns well with what we have seen in the preceding analysis with simulated data. Finally, the nested-likelihood ratio test selects the CMXL model over the MNL model, but the BIC value favors the MNL model. Similar results are also found when we impose a normal or lognormal distribution on the parameters for in-vehicle travel time and travel cost. If the analysis was stopped here, based on the BIC value, the conclusion would be that there is no need to consider heterogeneity in this model for this data set.

Mass-point mixed logit and latent class models

We test the discrete-distribution assumption on the two parameters' heterogeneity by

Table 5: Estimates for the MNL, the CMXL, the MPMXL and the LCLM

	MNL		*CMXL*		*MPMXL*		*LCLM (2 classes)*	
Random parameters								
Mean of constant, DA	1.1211	(0.0911)	0.7373	(0.1491)	−1.0004	(1.4282)	−0.9319	(0.5375)
Std. Dev., DA	N/A	N/A	0.1136	(0.1014)	4.0422	(9.6321)	4.1174	(7.0391)
VPW_DA	0.5684	(0.0718)	1.0486	(0.1755)	3.4104	(1.6337)	2.9623	(0.6280)
Std. Dev., VPW_DA	N/A	N/A	0.6183	(0.0177)	5.0780	(12.6181)	4.9394	(9.5778)
Correlation of DA and VPW_DA	N/A	N/A	−0.1473	(0.1413)	−0.8914	(3.1030)	N/A	N/A
Fixed parameters								
IVTT (min)	−0.0163	(0.0027)	−0.0169	(0.0028)	−0.0180	(0.0031)	−0.0175	(0.0027)
OVTT, auto and taxi (min)	−0.2502	(0.0211)	−0.2502	(0.0217)	−0.2516	(0.0237)	−0.2474	(0.0213)
OVTT, TCR (min)	−0.0442	(0.0144)	−0.0441	(0.0147)	−0.0454	(0.0159)	−0.0446	(0.0147)
Travel cost (dollar)	−0.1243	(0.0142)	−0.1283	(0.0150)	−0.1359	(0.0160)	−0.1317	(0.0145)
Parking cost, DA and SR (dollar)	−0.0239	(0.0046)	−0.0266	(0.0050)	−0.0352	(0.0074)	−0.0268	(0.0058)
VPW for taxi, TCR	−0.7580	(0.1028)	−0.6978	(0.1037)	−0.7833	(0.1090)	−0.7168	(0.0973)
Destination is Manhattan, taxi	4.0585	(0.3302)	4.1529	(0.3341)	4.3917	(0.3440)	4.2395	(0.3333)
Destination is Manhattan, TCR	1.3897	(0.1510)	1.4737	(0.1596)	1.6790	(0.1839)	1.5694	(0.1611)
Travel distance, TCR	0.0420	(0.0062)	0.0447	(0.0065)	0.0494	(0.0077)	0.0474	(0.0069)
Constant, taxi	−1.7481	(0.3344)	−1.789	(0.3361)	−1.7311	(0.3389)	−1.7746	(0.3317)
Constant, transit	−0.8073	(0.1843)	−0.866	(0.1873)	−0.8374	(0.1974)	−0.8504	(0.1796)
Constant, commuter rail	−0.7314	(0.2407)	−0.804	(0.2471)	−0.7975	(0.2742)	−0.8200	(0.2399)
Goodness-of-fit measures								
Log-likelihood at convergence	−3898.34		−3889.04		−3848.08		−3852.51	
Number of parameters (*K*)	14		17		18		17	
Adjusted ρ^2	0.5482		0.5490		0.5536		0.5532	
BIC	7920.32		7928.20		7855.11		7855.15	

Note: DA = drive alone; IVTT = in-vehicle travel time; OVTT = out-of-vehicle travel time; TCR = transit and commuter rail; VPW = vehicle per worker; VPW_DA = vehicle per worker for drive alone.

estimating the MPMXL model and the LCLM. For the MPMXL model, we assume a joint discrete distribution of the two parameters (CNST_DA and VPW_DA) located on a two-dimensional grid to define the joint location for the two parameters at four joint mass points. Estimation of this model was initiated with a larger set of discrete points in each dimension that were reduced in cases when the distance between any two masses along either parameter dimension is very small (<0.001) or the probability weight of a joint mass point is less than 0.1 per cent and not statistically significant. The total number of joint mass points is also decided based on the BIC. After several refining steps, the final MPMXL model has three joint mass points. This requires three location-parameter estimates and two probability estimates as reported in Table 6. For the LCLM estimation, the number of classes is defined based on the BIC value as discussed before, and the best one is found to be with two classes. The estimation results for the LCLM are reported in the lower part of Table 6.

Table 6 shows that the MPMXL model identifies three mass points for the joint distribution of these two parameters. The first one represents 73 per cent of the population and has location parameters close to those estimates from the MNL model listed in Table 5. The second group represents the 23 per cent of the population whose choice to drive alone is very sensitive to car availability in the household. The third group consists of only 4 per cent of the population.

The LCLM identifies only two classes, effectively combining the first and third groups from the MPMXL model. The discrete distributions identified by these two models are plotted in Figure 1, where the solid lines indicate the mass points for the MPMXL model estimates and the dotted lines indicate results from the LCLM. From this plot, we can see that the two points identified by LCLM are similar to the first two mass points from the MPMXL model. The only difference is that LCLM does not identify the smallest mass point found by the MPMXL model. The reason the MPMXL model can discover the small group missed by the LCLM is that, in the MPMXL model, by locating the joint distribution of these two parameters on a grid system, the third group is actually not identified freely from the other two mass locations. Even with this difference, the goodness-of-fit measures for these two models are almost identical; this can be seen from the BIC values listed in the last column of Table 6.

Comparing these two DMXL models with the CMXL and MNL models shown in Table 5, we notice that the likelihood ratio test between the CMXL and the MNL strongly rejects the MNL model and the BIC values strongly favor the MPMXL and LCLM models over both the continuous models. These results are obtained despite the apparent bias of the BIC measure towards the CMXL demonstrated in previous section using simulated data. The log-likelihood, adjusted rho-square and BIC values all strongly suggest that the discrete representation of unobserved heterogeneity

Table 6: Estimation results for MPMXL and two-class LCLM

	DA constant		*VPW_DA*		*Joint probability*		*BIC*
MPMXL	1.2147	(1.8252)	0.2887	(0.0927)	0.7257	(0.1483)	7855.11
	−8.3767	(1.9210)	11.6706	(0.1566)	0.2309	(0.1394)	
	1.2147	(1.8252)	11.6706	(0.1566)	0.0433	(0.0170)	
LCLM	1.3206	(0.1216)	0.2601	(0.0838)	0.7696	(0.0263)	7855.15
	−8.4581	(1.9930)	11.991	(2.2775)	0.2303	(0.0263)	

fits the data better than the continuous representation. We postulate that the failure of the CMXL model in this real data analysis indicates that the true distribution is more complicated than can be approximated by a unimodal (normal) distribution. It could be either discrete in nature, or bimodal or multimodal (Dong and Koppelman, 2003).

CONCLUSION

We compare two different classes of models, continuous and discrete, for incorporating unobserved heterogeneity into the estimation of mode-choice models using both simulated data and a real data set describing mode-choice behavior. The analysis using simulated data indicates that the CMXL model with an *a priori* assumption of a unimodal distribution performs better (based on the conventional goodness-of-fit measures) than estimation using a DMXL model, whether the true distribution is continuous or discrete. This is apparently due to the differences in the number of model parameters required for estimations. Besides, it shows that a DMXL model estimates could be misleading if the true distribution is indeed continuous. Therefore, it is suggested to estimate a model by starting with a CMXL model and also to use different starting values for the probability masses in the DMXL model.

In the analysis with real choice data, the CMXL model fails to identify heterogeneity that can be discovered by the DMXL model. This suggests that the CMXL model, even though commonly used in this context, sometimes leads to inferior results. Therefore, it is suggested that when analyzing heterogeneity in logit models, a CMXL model estimation should be followed by a DMXL estimation to avoid missing unobserved heterogeneity.

NOTES

1. However, it limits the distribution to a small number of discrete points rather than a continuous distribution.
2. When estimating with only one parameter following a discrete distribution, the MPMXL model and LCLM are identical.
3. Another data set is generated with larger variance, $\beta_1 \sim N(-6, 25)$, and the results are similar, except that the LCLM has three mass points. As it doesn't provide more instructive conclusions than what is presented in the article, we omit it.
4. If μ and σ^2 denote the mean and variance of the underlying normal distribution, the mean and variance of the lognormal distribution are $\exp(\mu+\sigma^2/2)$ and $\exp(2\mu+2\sigma^2)-\exp(2\mu+\sigma^2)$.
5. Data provided by the New York Metropolitan Transportation Council (NYMTC).

REFERENCES

Andrews, R.L., Ainslie, A. and Currim, I.S. (2002) An empirical comparison of logit choice models with discrete versus continuous representations of heterogeneity. *Journal of Marketing Research* 39(November): 479–487.

Ben-Akiva, M. and Bolduc, D. (1996) Multinomial Probit with a Logit Kernel and a General Parametric Specification of the Covariance Structure. Working paper, Département d'Économique, Université Laval, Quebec, Canada.

Ben-Akiva, M. and Lerman, S.R. (1985) *Discrete Choice Analysis*, Cambridge, MA: MIT Press.

Bhat, C.R. (1997) An endogenous segmentation mode choice model with an application to intercity travel. *Transportation Science* 31(1): 34–48.

Bhat, C.R. (2001) Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. *Transportation Research* 35B(7): 677–693.

Bollen, K.A., Ray, S., Zavisca, J. and Harden, J.J. (2012) A comparison of Bayes factor approximation methods including two new methods. *Sociological Methods and Research* 41(2): 294–324.

Burnham, K.P. and Anderson, D.R. (2004) Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research* 33(2): 261–304.

Chamberlain, G. (1978) Analysis of covariance with qualitative data. *Review of Economic Studies* 47(1): 225–238.

Chintagunta, P., Jain, D.C. and Vilcassim, N.J. (1991) Investigating heterogeneity in brand preferences in logit models for panel data. *Journal of Marketing Research* 28(4): 417–428.

Crouchley, R. (1987) *Longitudinal Data Analysis (Collection of Articles)*. Farnham, UK: Gower.

Dong, X. and Koppelman, F.S. (2003) Comparison of methods representing heterogeneity in logit models. Presented at 10th International Conference of Travel Behavior Research, Lucerne, Switzerland.

Greene, W.H. (2000) *Econometric Analysis*. Upper Saddle River, NJ: Prentice-Hall.

Guadagni, P.M. and Little, J.D.C. (1983) A logit model of brand choice calibrated on scanner data. *Marketing Science* 2(3): 203–238.

Heckman, J.J. (1981) Heterogeneity and state dependence. In: S. Rosen (ed.) *Studies in Labor Markets*. Chicago, IL: University of Chicago Press, pp. 91–139.

Heckman, J. and Singer, B. (1984) A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* 52(2): 271–320.

Horowitz, J. (1981) Sampling specification and data errors in probabilistic discrete-choice models. In: D.A. Hensher and L.W. Johnson (eds.) *Applied Discrete-Choice Modeling*. Appendix C. New York: Halsted Press, pp. 217–236.

Jones, J.M. and Landwehr, J.T. (1988) Removing heterogeneity bias from logit model estimation. *Marketing Science* 7(1): 41–59.

Kamakura, W.A. and Russell, G.J. (1989) A probabilistic choice model for market segmentation and elasticity structure. *Journal of Marketing Research* 26(4): 379–390.

Kuha, J. (2004) AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods and Research* 33(2): 188–229.

Magidson, J. and Vermunt, J.K. (2001) Latent class factor and cluster models, bi-plots and related graphical displays. *Sociological Methodology* 31(1): 223–264.

McFadden, D. (1978) Modeling the choice of residential location. In: A. Karlquist, F. Snickars and J. Weibull (eds.) *Spatial Interaction Theory and Residential Location*, Amsterdam, the Netherlands: North-Holland, pp. 75–96.

McFadden, D. and Train, K. (2000) Mixed MNL models for discrete response. *Journal of Applied Econometrics* 15(5): 447–470.

McQuarrie, A.D.R. and Tsai, C.-L. (1998) *Regression and Time Series Model Selection*. Singapore: World Scientific.

Niederreiter, H. (1992) *Random Number Generation and Quasi-Monte Carlo Methods*. SIAM CBMS-NSF Regional Conference Series in Applied Mathematics, Philadelphia, PA: SIAM.

Reader, S. (1993) Unobserved heterogeneity in dynamic discrete choice models. *Environment and Planning A* 25(4): 495–519.

Rossi, P.E., Allenby, G.M. and McCulloch, R. (2005) *Bayesian Statistics and Marketing*. New York: Wiley.

Sándor, Z. and Train, K. (2002) Quasi-random Simulation of Discrete Choice Models. Working Paper, Department of Economics, University of California, Berkeley.

Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics* 6(2): 461–464.

Train, K. (2000) Halton Sequences for Mixed Logit. Working paper, Department of Economics, University of California, Berkeley.

Train, K. (2003) *Discrete Choice Methods with Simulation*. Cambridge, UK: Cambridge University Press.

Weakliem, D.L. (2004) Introduction to the special issue on model selection. *Sociological Methods and Research* 33(2): 167–187.

Wedel, M. *et al.* (1999) Discrete and continuous representations of unobserved heterogeneity in choice modeling. *Marketing Letters* 10(3): 219–232.

Wen, C.-H. and Koppelman, F.S. (2001) The generalized nested logit model. *Transportation Research B* 35(7): 627–641.