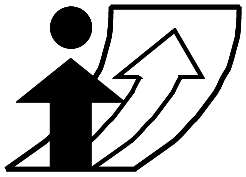
Comparison of Methods Representing Heterogeneity in Logit Models

Xiaojing Dong, Northwestern University
Frank S. Koppelman, Northwestern University

Conference paper

Session: Development of Theory



Moving through nets:
The physical and social dimensions of travel
10th International Conference on Travel Behaviour Research
Lucerne, 10-15. August 2003

Comparison of Methods Representing Heterogeneity in Logit Models

Xiaojing Dong

Frank S. Koppelman

Department of Civil Engineering, Northwestern University, Evanston, IL 60208

Phone: 1-847-467-1852
Fax: 1-847-491-4011
eMail: xjdong@northwestern.edu

Abstract

Representing heterogeneity or taste variations in behavioural choice analysis becomes more and more an important issue in the area of transportation mode choices. Mixed logit (MXL) models have been widely adopted for this purpose. In this paper, using the data about commuting trip mode choice in the morning peak, we compare different methods that are employed to represent heterogeneity using MXL models. These methods differ in the distribution assumptions including continuous distribution, discrete mass point and a mixture of multiple continuous distributions. We review and compare the effectiveness (goodness of fit), interpretation and computational difficulty of alternative methods of representing the distribution of unobserved heterogeneity.

Keywords

Mixed Logit Models, Heterogeneity, International Conference on Travel Behaviour Research, IATBR

Preferred citation

Dong, Xiaojing, Frank Koppelman, (2003), Comparison of Methods Representing Heterogeneity in Logit Models, paper presented at the 10th International Conference on Travel Behaviour Research, Lucerne, August 2003.

1. Introduction and background

Representing the unobserved heterogeneity in the utility assessment has become an important consideration while analysing people's mode choice behaviour. Mixed Logit models (MXL) have been widely adapted for this purpose (McFadden and Train, 2000). Among most of the available data describing transportation mode choices, only one choice of one individual is recorded. Theoretically, with this type of cross-sectional data, it is not possible to distinguish between randomness and taste variations. As a result, we cannot obtain individual-specific parameters. However, it is still feasible to capture the differences among people by assuming that the model parameters follow some distribution, instead of a fixed parameter as constrained in the closed form GEV 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 behaviour in a variety of domains during the last twenty to thirty years.

In most of the current implementations of MXL models, the heterogeneity in some parameters is assumed to follow a continuous distribution. We call this the Continuous Mixed Logit Model (CMXL). Estimation is based on maximization of a simulated likelihood function, called the Maximum Simulated Likelihood method. The analyst chooses the shape of the distribution and the estimation procedure obtains the parameters of the distribution. The most commonly assumed distributions; the normal, lognormal, triangular and uniform distributions; can be represented by two parameters. In addition, all of these distributions allow the introduction of correlations among parameters. An important limitation of this approach is that its estimation relies on the *a priori* assumption about the shape of the distribution, which might lead to an inferior solution 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 non-parametric in one or multiple 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. This mass point approach can be seen as closely related to the Latent Class model (LCM, 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 and Bhat, 1997). The differences between these models are discussed further in the next two sections.

In the MPMXL models, each mass point represents a group of people or a market segment, assuming they are homogeneous in their taste preferences within the segment. This assumption can be relaxed by substituting the one point estimation for the same segment with a continuous distribution, to allow heterogeneity even within the segment. As a result, the

variation across people's tastes in the population is described by a mixture of multiple continuous distributions. When the within-segment heterogeneity is represented by a Normal distribution, the Mixed Logit model is therefore called Mixture-Normal Mixed Logit (MNMXL) model.

In the following sections, we describe each of these three types of mixed logit models, CMXL, MPMXL (and LCM) and MNMXL, followed by a comparison of implementing these models using a mode choice data set, based on which we conclude our findings.

2. Different methods of representing heterogeneity in Mixed Logit models

The general form of a MXL can be represented by

$$P_n(i) = \int \frac{\exp(X_i\beta)}{\sum_{j\in C_n}\exp(X_j\beta)} dF(\beta) \qquad (2.1)$$

In this formulation, the index *n* denotes an individual, the index *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 β are parameters to be estimates. The marginal probability of individual *n* choosing alternative *i* is computed by an integral over the distribution of parameter β, with the integrant the same as for an MNL model (Ben-Akiva, M. and S. Lerman, 1985). It is also called a Logit-Kernel model (Ben-Akiva and Bolduc, 1996). Then the question is what should be the formulation for $F(\beta)$? In the following section, we describe three possible formulations for $F(\beta)$, including a continuous distribution, a mass point distribution with finite number of supports, and a mixture of multiple continuous distributions. We also compare the computational efficiency of each model.

2.1 Continuous Mixed Logit model

When $F(\beta)$ is assumed to be a continuous distribution with probability density function $f(\beta)$, then equation (2.1) can be rewritten as

$$P_n(i) = \int \frac{\exp(X_i\beta)}{\sum_{j\in C_n}\exp(X_j\beta)} f(\beta)\, d\beta \qquad (2.2)$$

As the integral doesn't have a closed form solution, a simulation approach is employed to compute the probability. Then equation (2.2) is replaced by its simulated formulation

$$\widetilde{P}_n(i) = \frac{1}{NR}\sum_{r=1}^{NR}\frac{\exp\left(X_i\beta^r\right)}{\sum_{j\in C_n}\exp\left(X_j\beta^r\right)} \tag{2.3}$$

β^r is the r-th random draw from its distribution $f(\beta)$, and *NR* is the total number of random draws. The simulated probability is computed by first making a random draw of β^r from the assumed probability density distribution $f(\beta)$, then computing the Logit probability conditional on this parameter value. This computation is repeated *NR* times, and the mean of the *NR* computed MNL probabilities is the simulated value of the marginal probability. Then using Maximum Simulated Likelihood estimation (MSLE), the parameters for distribution $f(\beta)$ can be estimated, as well as the other parameters that are assumed to be homogeneous across the population.

In implementation, in order to obtain better coverage in the support of $f(\beta)$, Halton draws (Train, 1999 and Bhat, 2001) are used, and when the dimension of $f(\beta)$ is big, Scrambled Halton draws should be used to avoid high correlations across dimensions.

There are two identified shortcomings of the CMXL while implementing it. First, as simulation is required in the estimation procedure, the computational time is significantly higher compared to those closed form GEV models. Halton sequences are developed to provide a better coverage with fewer random numbers. Also, as a result of the use of simulation, simulation bias or even biased estimators are produced. The solution is to increase 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*, or denoted as *O(NR)*.

Second, it requires *a priori* assumption of the distribution formulation of $f(\beta)$ and the estimation results are dependent on this assumption. In most cases, a single modal distribution with two parameters is employed, which will fail the estimation if the true distribution form is a bi-modal distribution. In order to relax the distribution assumption in the estimation process, a non-parametric distribution is assumed, which is called the Mass-Point distribution, and the Mixed logit model with a Mass-Point distribution assumption on some of its parameters is called Mass-Point Mixed Logit model.

2.2 Mass-Point Mixed Logit model

Mass point distribution has a finite number of supports, and the location and probability weight of each mass point can be estimated using the Maximum Likelihood estimation

process. By replacing the probability distribution $f(\beta)$ in equation (2.2) with a mass point distribution whose probability weight at the m-th mass point is λ^m, and by replacing the integration in equation (2.2) with a sum over a fixed number of mass points *M*, the formulation for the MPMXL model can be expressed as:

$$P_n(i) = \sum_{m=1}^{M} \frac{\exp\left(X_i \beta^m\right)}{\sum_{j \in C_n} \exp\left(X_j \beta^m\right)} \lambda^m \tag{2.4}$$

In MPMXL, the probability of individual *n* choosing alternative *i* can be regarded as the weighted average of *M* Logit probabilities; where the Logit probabilities are computed at each possible value of β, and the weights are the probability of β to be at each value β^m. Therefore, it has a closed form representation for the choice probability, and the estimation procedure doesn't require simulation. As such, the computational complexity of this model is in the order of the number of total classes *M*, denoted as *O(M)*. As M is usually much smaller than the number of random draws NR used in estimating the CMXL model, MPMXL model requires much less computational resource than the CMXL model.

As mentioned above, this model is actually similar to the Latent Class model, except that in mass point distribution, we focus on the distribution of some parameters, while in LCM, we are more interested in the segments of the population. While estimating these two models, for example, if we have all the other parameters fixed across the population, except two parameters. Then in MPMXL, we assume there are two mass points for each of these two parameters, this results in estimating four classes four distinct parameters. In a LCM with four classes for two parameters, we need to estimate eight distinct parameters. However, if we fix some of the eight parameters to be identical, we could end up with the same model as the MPMXL model. Detailed discussion about the difference between these two models is presented in the next section with the empirical analysis.

Compared to the CMXL model, the MPMXL model reduces the computational time significantly and is free from simulation biases as it has a closed form formulation. It also relaxes the *a priori* assumption on the distributions of the parameters. However, it tries to represent the complicated heterogeneity in the population simply by a finite number of points, in most cases no more than 5 (Heckman and Singer, 1984). In other words, the population is represented by at most five different market segments, and within each segment, all the people are considered to have the exactly same tastes in their utility assessment. This may not fully describe the complicated taste variations across the population.

If the within-segment taste variations are assumed to follow a continuous distribution, such as normal distribution, the overall population heterogeneity is represented by a mixture of normal distribution, and this is what we call Mixture Normal Mixed Logit model (MNMXL).

2.3 Mixture Normal Mixed Logit model

In this type of Mixed Logit model, the MNL probability part in equation (2.4) is replaced by the marginal probability of the Continuous Mixed Logit model, as described in equation (2.2).

$$P_n(i) = \sum_{m=1}^{M} \left[\int \frac{\exp\left(X_i \beta^m\right)}{\sum_{j \in C_n} \exp\left(X_j \beta^m\right)} f_m(\beta)\, d\beta \right] \lambda^m \qquad (2.5)$$

This way, the heterogeneity within each market segment is represented by the integral term inside the brackets in equation (2.5), and the difference across market segments is represented by the weighted sum outside the brackets, which is the same as the outer part in the MPMXL model as described in equation (2.4). As such, this model can be regarded as a combination of both the Continuous MXL model and the Mass-Point MXL model. It inherits the properties of both models: first, it enhances the CMXL model by being able to capture the parameter distributions with multiple modes; second, it is superior to the MPMXL in that it relaxes the homogeneity assumption within each class. However, because of the continuous assumption within each class, simulation is necessary to compute the probability in equation (2.5), which leads to longer computational time and possible simulation bias. Actually, the computational complexity of this model is in the order of the products of number of random draws NR and the total number of classes M, denoted as $O(NR \times M)$.

In this section, we focus our attention on the differences among the CMXL , MPMXL and MNMXL models in terms of the model formulations, behavioural implications and computational complexities. In the following section, all models are applied to a mode choice analysis, and their performances are compared further based on the empirical results.

3. Empirical Comparison of the three methods representing heterogeneity in logit models

This analysis considers the work trip mode choice in the New York metropolitan area[1]. It contains work trip information for 6844 individuals who traveled to work during morning

[1] Data Provided by the New York Metropolitan Transportation Council (NYMTC).

peak hours. The information includes the chosen mode, trip origin and destination, travel service and cost characteristics for each available mode and the socio-economic characteristics of the travelers and their households. The observed mode shares for the five modes considered are shown in Table 1.

Table 1 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%

This analysis estimates several models by imposing the three different distribution assumptions, as discussed in the previous section, on two parameters, which are alternative specific constant for drive alone (CNST_DA) and the alternative specific variable "vehicles per worker" for drive alone (VPW_DA). We started by estimating an MNL model as reference, followed by a Continuous MXL model with a binomial normal distribution assumption on these two parameters. Then using mass point distribution to substitute the continuous distribution assumption, we estimate a Mass-Point MXL model. To compare the performance of MPMXL model with Latent Class model, we also estimate a LCM. Then by further assuming heterogeneity within each class, we estimate two Mixture Normal MXL models, one contains three classes, based on the estimates from the Mass-Point MXL model, and the other one has two classes, based on the LCM. For the purpose of comparison, the results of these six models are listed two tables, those for the first four models (MNL, CMXL, MPMXL and LCM) are in Table 2 and those for the last two Mixture Normal MXL models are listed in Table 6.

Table 2 Estimates for the MNL, the Continuous MXL, the Mass-Point MXL and the Latent Class models

	MNL		CMXL		MPMXL		LCM (2 Classes)	
	Mean	Std. Err	Mean	Std. Err	Mean	Std. Err	Mean	Std. Err
Constant for Drive alone	1.1211	(0.0911)	0.7373	(0.1491)	-1.0004	-------	-0.9319	-------
STD_DA	-------	-------	-0.1136	(0.1014)	4.0422	-------	4.1174	-------
Vehicle per worker (VPW) for DA	0.5684	(0.0718)	1.0486	(0.1755)	3.4104	-------	2.9623	-------
STD_VWDA	-------	-------	0.6116	(0.1343)	5.0780	-------	4.9394	-------
Corr_DAVW	-------	-------	0.1490	------	-0.8914	-------	-------	-------
In-vehicle travel time (IVTT)	-0.0163	(0.0027)	-0.0169	(0.0028)	-0.0180	(0.0031)	-0.0175	(0.0027)
Out of vehicle travel time (OVTT) for auto and taxi	-0.2502	(0.0211)	-0.2502	(0.0217)	-0.2516	(0.0237)	-0.2474	(0.0213)
OVTT for Transit and Commuter rail	-0.0442	(0.0144)	-0.0441	(0.0147)	-0.0454	(0.0159)	-0.0446	(0.0147)
Travel cost	-0.1243	(0.0142)	-0.1283	(0.0150)	-0.1359	(0.0160)	-0.1317	(0.0145)
Parking cost for DA and SR	-0.0239	(0.0046)	-0.0266	(0.0050)	-0.0352	(0.0074)	-0.0268	(0.0058)
VPW for Taxi, Transit and Commuter rail	-0.7580	(0.1028)	-0.6978	(0.1037)	-0.7833	(0.1090)	-0.7168	(0.0973)
Destination is Manhattan, for Taxi	4.0585	(0.3302)	4.1529	(0.3341)	4.3917	(0.3440)	4.2395	(0.3333)
Destination is Manhattan, for Transit/Commuter rail	1.3897	(0.1510)	1.4737	(0.1596)	1.6790	(0.1839)	1.5694	(0.1611)
Travel distance for Transit and Commuter rail	0.0420	(0.0062)	0.0447	(0.0065)	0.0494	(0.0077)	0.0474	(0.0069)
Constant for Taxi	-1.7481	(0.3344)	-1.7887	(0.3361)	-1.7311	(0.3389)	-1.7746	(0.3317)
Constant for Transit	-0.8073	(0.1843)	-0.8655	(0.1873)	-0.8374	(0.1974)	-0.8504	(0.1796)
Constant for Commuter rail	-0.7314	(0.2407)	-0.8043	(0.2471)	-0.7975	(0.2742)	-0.8200	(0.2399)
Likelihood	-3898.3		-3889.0		-3848.1		-3852.5	
Number of parameters (k)	14		17		18		17	
BIC=-2*LL+ln(NOBS)*k	7920.32		7928.2		7855.11		7855.15	

In Table 2 there are four groups of rows. The first group shows the estimates of the two parameters CNST_DA and VPW_DA, for which we are exploring the performance of different methods to represent heterogeneity. The second and third groups display the estimated results for those variables that are assumed to be homogeneous over the whole population, including some explanatory variables (such as in-vehicle travel time, out-of-vehicle travel time, travel cost, parking cost, travel distance, a dummy variable to indicate whether the destination is Manhattan, etc.), and alternative specific constants for modes other than Drive Alone. The last group contains the measures of goodness-of-fit, including Log-Likelihood values at convergence, Bayesian Information Criteria (BIC) values and the number of parameters in each model, which is used to calculate BIC. BIC is computed as a function of log-likelihood value at convergence, with a penalty on the number of parameters (Schwartz, 1978). In model selection based on BIC, we should select the model minimizing the BIC value. An introduction to the BIC is provided in the Appendix.

Based on the second and third groups of rows, the estimated results for the fixed parameters are consistent across the different models. This indicates that the differences in the fit of the models result mainly from the different ways of incorporating taste variations into the models. In the following sections, we will discuss the differences among these models based on these empirical results, in terms of both the model estimates for the two parameters (CNST_DA and VPW_DA) and model fits.

3.1 MNL Model and Continuous MXL model

The comparison is started by studying the first two models, MNL and Continuous MXL models. The MNL model is estimated using Maximum Likelihood method and the Continuous MXL model is estimated using Maximum Simulated Likelihood with a binomial normal distribution assumption on the two parameters. The MNL model provides only point estimate for each of the two parameters. With about 7000 observations in this dataset, the standard errors for these point estimates are pretty low. But the estimated values are both different from the means obtained by the CMXL model. As for the estimated results for the CMXL model, standard deviation in for drive alone is low (-0.1136) and not statistically significant, which indicates that the heterogeneity assumption on this constant may not be necessary. The standard deviation estimate for VPW_DA is 0.6116, and statistically significant. The correlation for this joint normal distribution is only 0.15. In terms of the comparison based on the goodness-of-fit measures, the Log-Likelihood value of CMXL model is better than for the MNL model, but the BIC value favors the MNL model. The increase in the Log-Likelihood in CMXL model is a result from having three more parameters than the MNL model, including two standard deviations and one covariance. Based on the

BIC value, the MNL model should be preferred to CMXL model, which means that no heterogeneity needs to be considered in these two parameters. A similar result is also found while we try to impose a lognormal distribution on the parameters for in-vehicle travel time and travel cost.

3.2 Mass-Point MXL model and Latent Class Model

Then, using the Maximum Likelihood method, we estimate a MXL model assuming the two parameters following a binomial distribution with finite number of supports. This is what we called the Mass-Point MXL model. The location of each mass point for this joint distribution is defined jointly by the locations for each of the two parameters, and is called "joint mass point" in this analysis. Two-dimensional vectors are used to denote the locations for these mass points, with the first dimension for DA constant and the second dimension for "vehicle per worker for DA". While estimating this model, we actually started with a larger number of mass points for each of the two parameters, and we estimated both the mass point locations and the probability weights at the same time. Then based on the model estimates, we adjust the number of mass points for each parameter, and the number of the joint mass points for the combination of the two parameters. If the distance between any two masses for that parameter is too low (less than 0.001), the number of masses on that dimension are reduced by 1; if the probability weight of some joint mass point is too low (less than 0.1%), we restrain that joint mass point to have zero probability weight. After refining several times, the final model has two mass points for each parameter and three joint mass points. This requires four location estimates and three probability estimates; the results are listed in Table 3 and Table 4 respectively, with the standard errors on the estimated parameters[2].

[2] As the probability weights need to be constrained between 0 and 1, and the sum of these three probability masses has to be 1, we estimate two parameters α_1 and α_2, and the probabilities are computed as $p_1 = \frac{\exp(\alpha_1)}{\exp(\alpha_1)+\exp(\alpha_2)+1}$, $p_2 = \frac{\exp(\alpha_2)}{\exp(\alpha_1)+\exp(\alpha_2)+1}$ and $p_3 = 1 - p_1 - p_2$. From Maximum Likelihood Estimation, we can only obtain the asymptotic standard errors for the parameters α_1 and α_2, the asymptotic standard errors are obtained using Delta method.

Table 3 Estimated Locations for MPMXL

	Locations	Standard Errors
DA Constant	-8.3767	1.9210
	1.2147	1.8252
VPW_DA	0.2887	0.0927
	11.6706	0.1566

Table 4 Joint Mass Point probability from MPMXL model

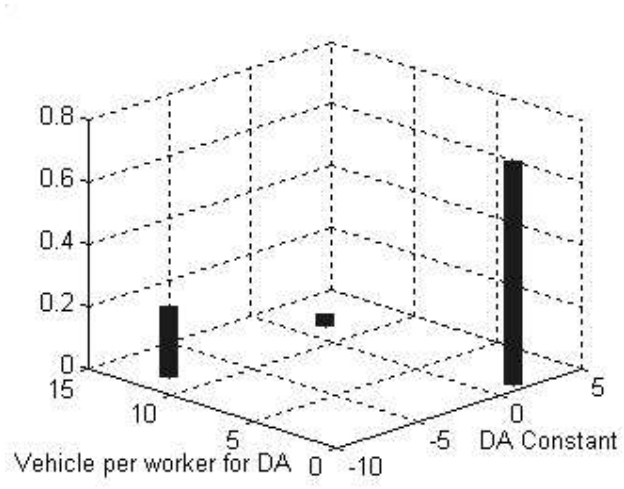
DA Constant	VPW_DA	Joint probability	Standard Errors
1.2147	0.2887	0.7257	0.1483
-8.3767	11.6706	0.2309	0.1394
1.2147	11.6706	0.0433	0.0170

The location estimates listed in Table 3 show that the two mass points for each parameter are pretty far apart, and they all are statistically significant. And the probability weights estimates in Table 4 indicate that there exist three joint mass points, or there are three groups in the population while considering these two components in evaluating their utilities. The first group is the majority, accounting for about 73% of the population. And they have both the two parameters estimates close to those from the MNL model, where both are positive. The second group consists of about 23% of the population, who have driving as their last choice of commuting mode, if no impact from other information. But the parameter estimates for VPW_DA is very as high as 11.6706, which indicates that these people are very sensitive to the auto-ownership compared to the number of workers in their households. This indicates that 23% of the population make their decision about whether driving to work primarily based on whether they have a car available for them to drive. These people could be those who consider privacy and flexibility more than others, which can be offered only by driving alone, but these considerations cannot be included in the explanatory variables. So, if they have enough cars compared to number of commuters in their households, they would be happy to choose driving, otherwise, they will have to find an alternative mode. The third group is the smallest group, consisting of only 4% of the population. Their preferences for DA mode are similar to the vast majority, but they are much more sensitive to “vehicles per worker” than others. Imagine if a household has fewer cars than commuters, then the member in this household who belongs to the 4% population, would be the person who would get to drive if only one more car is purchased in the household, and would also be the first to give up his/her

car if the number of cars in the household is reduced by one. These taste variations across the population was not identified using CMXL model. And the distribution is plotted in Figure 1.

Also, in terms of goodness-of-fit, the BIC value of the Mass-Point MXL model is improved dramatically from both MNL and CMXL models, which implies that the MPMXL model should be preferred. We postulate that the failure of Continuous MXL model to discover the taste variations among the population while evaluating their utilities might be a result from the normal distribution assumption, which is restricted to have only one mode. This is the major shortcoming attributed to the CMXL model; specifically, a distribution formulation assumption has to be imposed *a priori*, which, in this case, results in inferior results.

Figure 1 Joint mass probability for DA constant and DA vehicle per worker



Based on the MPMXL model, we also estimate a Latent Class model. In MPMXL model, we can identify three groups in the population. Then we estimate a Latent Class model with three classes. The estimated locations and probability weights for the three classes are listed in Table 5, with the standard errors listed in the parentheses. As mentioned in the previous section, both MPMXL model and LCM assume heterogeneity can be represented by a finite number of mass points. The difference is that the MPMXL model focuses on the distribution of the random parameters, while the number of joint mass points is defined by both the differences in the locations along each dimension (marginal distribution), and the probability weights for the joint mass points (joint distribution). On the contrary, as LCM focuses on classes of the population, it is not concerned with the different values for any single parameters that are in different classes. It takes all the parameters assumed with heterogeneity as a whole, and attempts to find the number of distinct classes for the whole set of parameters

that gives the best goodness-of-fit for the model. The difference in implementing these two models with the data we present here is that in the MPMXL model, as it is interested in both the marginal and the joint mass point distribution for each of the two parameters, for different joint mass points, one component in the two dimensions could be the same. For example, in Table 4 the DA constant takes the same value in the first and third groups, both are 1.2147. In general, the LCM doesn't impose constraints like this, and therefore when the MPMXL model and the LCM are assumed to have the same number of classes, the MPMXL model usually has fewer parameters to estimate. In this analysis, the MPMXL model needs four parameters to identify the three classes (Table 3), but the LCM needs 6 parameters (Table 5). This is the reason that when the probability of the third group is as low as 4%, the estimates for that group have very large standard errors, as listed in the last row of Table 5. This happens even though all the other parameter estimates are similar to those from MPMXL model, including both the locations and the probability weights. Actually, the estimates for the third group in MPMXL model are not freely estimated, as they are constrained to be the same as one component in each of the other two groups.

Table 5 Location and probability weights estimates from LCM with three classes

DA Constant		VPW_DA		Probability Weights	
Mean	Std. Err.	Mean	Std. Err.	Mean	Std. Err.
1.2091	(0.1193)	0.2896	(0.0833)	0.7234	(0.0904)
-8.0422	(1.9123)	11.3340	(2.1693)	0.2343	(0.0849)
5.4152	(12.7916)	6.0500	(20.3117)	0.0424	(0.0112)

Realizing the difference in the number of parameters estimated by these two models while having the same number of classes, we estimate another LCM model with only two classes, so that the total number of parameters is same as the MPMXL model. In this case, the parameter estimates are almost identical between these two models, as well as the goodness-of-fit measure. The probability weights are 23% and 77% for the two classes in the LCM model. Further, the BIC value is better than that obtained with three classes, therefore, between the two LCM models, we select the one with two classes, and listed its results in Table 2. Comparing the MPMXL model and the LCM with two classes, the only difference lies in the joint distribution for DA constant and VPW_DA. The LCM identifies the two points with bigger weights in Figure 1, but not the smallest point.

3.3 Mixture Normal MXL Model

Both the MPMXL model and the LCM use finite number of supports to represent heterogeneity, assuming people within each group people are homogeneous in evaluating their utility assessment. The Mixture Normal MXL model further relaxes the homogeneity assumption, by imposing a normal distribution assumption with-in each class or group.

In this analysis, we estimate two Mixture Normal MXL models based on the outputs from the MPMXL model and the LCM respectively.

According to the MPMXL model the locations of each joint mass point has to satisfy some constraints, therefore in the first Mixture Normal MXL model, we fix the mean of each normal component at those joint mass points estimated from the Mass-Point MXL model, only the standard deviation and covariance are estimated. And we assume there are three normal components, and this model is referred to as Three-Class MNMXL (or 3C-MNMXL) model. Based on the LCM estimates, we assume that the two parameters follow a mixture of two binomial normal distributions, and we estimate the means, standard deviations and the correlations for each of these two distributions. And this model is referred to as Two-Class MNMXL (or 2C-MNMXL) model in the following discussion.

While estimating the 3C-MNMXL model based on the results from Mass-Point MXL model, the standard deviations and correlations for the second and third groups (with 23% and 4% probability weights) are very close to zero, and not statistically different from zero, therefore we constrain them to be zero and estimate the distribution for only the first group with the largest probability weight (73%). And the estimated results are displayed in Table 6[3]. The numbers inside "[ ]" are those with fixed values, and the numbers inside "( )" are standard errors for corresponding estimates. Comparing this model with the MPMXL model, we can notice that the probability weights estimates for each group remains almost the same, and the standard deviation estimates for the binomial normal distribution for the first (the largest) group are both statistically significant. And the correlation estimate is negative one[4]. We would expect the correlation between these two estimates, because they both are alternative specific for DA mode, and they both describe an individual's preference bias over DA mode without consideration about any specific trip. This estimation results can be represented by a

3 Both these two models have all the other estimates for the fixed parameters very close to those listed in Table 2, therefore, in Table 6, we eliminated those outputs.

4 We didn't compute the asymptotic standard error for this parameter. Using Delta method we could derive the asymptotic standard error for the covariance. The covariance based on the model estimates is –0.4513, and the derived asymptotic standard error is 0.2539.

similar plot as in Figure 1, by substituting the highest bar with a binomial normal distribution. The overall joint distribution: with one binomial and two mass-point distributions, indicates that there exists some continuous heterogeneity among the majority of the population, while there are two smaller groups that evaluate these two parameters very differently. In terms of the marginal distributions for these two parameters, they both follow a distribution as a combination of one normal distribution for most of the individuals and one mass points for a smaller group. This type of distribution would not be able to recover by the Continuous MXL model, and therefore the reason that we cannot identify heterogeneity with CMXL model. Comparing these results with those from MNL model, we notice that the MNL Model estimates for these two parameters are very close to the means for the largest group. And for the mean estimates from the Continuous MXL model, the one for DA constant is lower than the point estimate in the MNL model, as it might detect some variation caused by the second group identified from the MPMXL model and the 3C-MNMXL model; similarly, the mean estimates for VPW_DA by the CMXL model is higher than that from the MNL Model estimates, because of the mass points with 11.6706 identified by the MPMXL model and the 3C-MNMXL. And we should notice that the reason to discover this type of complicated distributions is a joint effort by both the Mass-Point MXL model, who defines the locations of the joint mass points and the means of the joint normal distributions, and the MNMXL model. In terms of the goodness-of-fit, by imposing the binomial normal distribution assumption on the joint mass point with the largest probability weight, the 3C-MNMXL model obtains a better likelihood value, with fewer parameters, which also results in a better BIC value.

Then we estimate the 2C-MNMXL model based on the LCM, and the estimates are also listed in Table 6, for the purpose of comparison. In this model, only two classes are considered, and the locations of the mass points and the means of the normal distribution are all estimated in this model. Similar to the 3C-MNMXL model, we assume a normal distribution in the class with the largest probability weight. And the results look very similar to those obtained from the 3C-MNMXL model. They are different slightly in the probability weights, especially for the one with the largest weight. It is 72% in the three-class model, and 79% in the two-class model. This indicates that with after eliminating one class, the larger group actually picks up the third group identified from the MPMXL model. And because of the same reason, the standard deviations for the binomial normal distribution in the largest group are higher for the two-class estimates than the three-class estimate. Based on the three-class estimates, the third class is located at (1.2147, 11.6706), which is far away from the mean of the normal distribution. The two-class MNMXL model cannot identify it, just treats it as an outlier of the normal distribution. In both models, the derived correlation is –1. In terms of the goodness-of-fit measure, the two-class model gets a slightly better log-likelihood at convergence than the

other one, but have three more parameters, and the BIC value is not as good as that for the three-class model.

4. Conclusion

In this analysis, we compare three different ways to incorporate heterogeneity consideration into mode choice study using five different models. By comparing the results from these different models, we can see that simply imposing a continuous distribution assumption with only one mode could lead to inferior results. Relaxing the distribution assumption by only mass point distribution might result in a too simple representation for the complicated taste variations across the population. The best way might be to start with the mass point distribution assumption to identify the different groups, and then further relax the homogeneity assumption within each group using a continuous distribution. It increases the required computational efforts, but it can identify more complicated patterns. And we also conduct the comparison between the Mass-Point MXL model and the LCM, from which we conclude that the LCM might miss some group with smaller probability weight that could be identified by the Mass-Point MXL model.

Table 6 Estimates for Mixture Normal MXL models based on results from Mass-Point MXL model and LCM

Based on the results from MPMXL (3 classes, with fixed mean)						Based on the results from LCM (2 classes, no fixed means)					
DA Constant		VPW_DA		Correlation	Probability	DA Constant		VPW_DA		Correlation	Probability
Mean	Std. Dev.	Mean	Std. Dev.		Weights	Mean	Std. Dev.	Mean	Std. Dev.		Weights
[1.2147]	0.9490 (0.2784)	[0.2887]	0.4756 (0.1356)	-1	0.7199 (0.0291)	1.3258 (0.1538)	1.6321 (0.2969)	0.3653 (0.1080)	0.7959 (0.2798)	-1	0.7904 (0.0789)
[-8.3767]	[0.0]	[11.6706]	[0.0]	[0.0]	0.2373 (0.0282)	-8.4365 (2.0241)	[0.0]	12.3426 (2.3302)	[0.0]	[0.0]	0.2096 (0.0789)
[1.2147]	[0.0]	[11.6706]	[0.0]	[0.0]	0.0428 (0.0075)	-------	-------	-------	-------	-------	-------
Log-Likelihood				-3845.4						-3844.6	
Number of parameters				17						20	
BIC=-2*LL+NPARMS*ln(NOBS)				7840.9						7865.9	

5. Reference:

Ben-Akiva, M. and Steven R. Lerman (1985) Discrete Choice Analysis, Cambridge Mass, MIT Press

Ben-Akiva M. and D. Bolduc (1996), "Multinomial Probit with a Logit Kernel and a General Parametric Specification of the Covariance Structure" working paper, Department d'Economique, Universite Laval, Quebec, Canada

Bhat, C.R. (1997), An Endogenous Segmentation Mode Choice Model with an Application to Intercity Travel, Transportation Science, Vol. **31**, No. 1, pp. 34-48.

Bhat, C.R. (2001), "Quasi-Random Maximum Simulated Likelihood Estimation of the Mixed Multinomial Logit Model", Transportation Research, Vol. **35B**, pages 677-693

Heckman, J., and B. Singer (1984), "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data", *Econometrica*, Vol. **52**, No. **2**., pages 271-320.

Kamakura, Wagner A. and Gary J. Russell (1989), "A Probabilistic Choice Model for Market Segmentation and Elasticity Structure," *Journal of Marketing Research*, **26** (November), 379-390

McFadden, D. (1978), Modeling the choice of residential location, Spatial interaction theory and residential location, A. Karlquist et al. (ed.), North-Holland, Amsterdam, pages 75-96

McFadden, D. and K. Train (2000), Mixed MNL Models for Discrete Response, *Journal of Applied Econometrics*, Vol. **15**, No. 5, pp. 447-470.

Schwarz, G. (1978) "Estimating the Dimension of a Model", The Annals of Statistics, Vol. **6**

Train, K. (1999) "Halton Sequences for Mixed Logit", Working paper, Department of Economics, University of California, Berkeley.

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

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

Appendix

Bayesian Information Criteria (BIC) is computed as a function of log-likelihood value at convergence, with a penalty on the number of parameters (Schwartz, 1978). If *LL* denotes the log-likelihood at convergence, *K* denotes the number of parameters estimated in the model and *NOBS* denotes the total number of observation used to estimate the model, then

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

BIC is derived based on the posterior model probability in Bayesian statistics. Let y denotes the data, M denotes the model, then the posterior probability of M conditional on the observed data is

$$p(M \mid y) = \frac{p(y \mid M)\, p(M)}{p(y)}$$

As it is conditional on y, *p(y)* is a constant, the above equation can be simplified as

$$p(M \mid y) \propto p(y \mid M)\, p(M)$$

Among this, on the left hand, $p(M \mid y)$ is the posterior probability for model M given data *y*, *p(M)* is the prior for model M, and $p(y \mid M)$ is the likelihood. If θ denotes the parameters, then we have

$$p(y \mid M) = \int p(y \mid M, \theta)\, p(\theta \mid M)\, d\theta$$

which is also called the integrated likelihood. BIC is an approximation of –2 log (integrated likelihood). Minimizing BIC is equivalent to maximizing the integrated likelihood, which is equivalent to maximizing the posterior probability of a model, when the priors are all equal. However, there are critiques about BIC, mainly about the role that the priors play. Even that, as BIC is easy to compute, and doesn't have to be used for nested models, it is widely used in selecting statistical models.