Competing methods for representing random taste heterogeneity in discrete choice models

Author(s):
Hess, Stephane; Fosgerau, Mogens

Publication Date:
2006-07

Permanent Link:
https://doi.org/10.3929/ethz-a-005240054

Rights / License:
In Copyright - Non-Commercial Use Permitted
Competing methods for representing random taste heterogeneity in discrete choice models

Stephane Hess * Mogens Fosgerau †

July 12, 2006

Abstract

This paper reports the findings of a systematic study using Monte Carlo experiments aimed at comparing the performance of various methods in retrieving random taste heterogeneity in a discrete choice context. Specifically, the analysis compares the performance of the four most commonly used continuous distribution functions to that of two more advanced approaches discussed in this paper. The first of these two approaches improves on the flexibility of a base distribution by adding in Legendre polynomials. The second approach uses a discrete mixture of multiple continuous distributions, where, in the present study, the base distributions are all Normal. The analysis shows very constant performance for the two approaches across a host of different scenarios, with varying levels of gains in performance when compared to the more basic distributions. Finally, and rather worryingly, the analysis shows that, across the various case studies, the poorest overall performance is obtained by the commonly used Normal distribution, which is in fact even outperformed by the Uniform distribution.

Keywords: random taste heterogeneity, mixed logit, Legendre polynomials, mixtures of distributions

*Institute for Transport Planning and Systems, ETH Zürich, stephane.hess@ivt.baug.ethz.ch
†Danish Transport Research Institute, Copenhagen, mf@dtf.dk
1 Introduction

In this paper, we describe the results of a systematic Monte Carlo analysis aimed at comparing a set of competing methods for representing random taste heterogeneity in discrete choice structures such as the Mixed Multinomial Logit (MMNL) model (cf. Revelt and Train, 1998; Train, 1998; McFadden and Train, 2000; Hensher and Greene, 2003; Train, 2003). In such models, it is necessary to specify a distribution for each random taste coefficient prior to model estimation. This is a non-trivial problem, in the almost complete absence of information on the shape of the true distribution, and in the knowledge that a bad choice of mixture distribution can seriously bias results (cf. Hess et al., 2005; Fosgerau, 2006).

Various authors have estimated a range of parametric distributions, aiming to gauge the advantages of distributions with a high degree of flexibility (see for example Hensher and Greene, 2003; Train and Sonnier, 2005; Hess et al., 2006a). Although different distributions have different properties, flexibility is generally determined by the number of parameters for the distributions. A two-parameter distribution corresponds to just a two-dimensional subset of some space of distributions. As such, while it may be possible to find a low-parameter parametric distribution that fits well in a specific situation, it will not be more flexible than other parametric distributions with the same number of parameters. This acts as the main motivation for exploring alternative ways of representing random taste heterogeneity.

Another shortcoming of the majority of parametric distributions used in mixed logit models is that they only allow for a single mode in the distribution. This is a major assumption, especially in the case where the population is segmented into mutually exclusive groups along some socio-demographic dimension that the modeller has no information on. The main exception to this rule is the Johnson S$_B$ distribution, which allows for two modes. However, in practice, this distribution is often difficult to estimate, as for example observed by Hess and Rose (2006). Both approaches discussed in this paper allow for multiple modes in the distribution.

The method of sieves is a natural choice for generating flexible distributions. Consider some model containing an unknown function to be estimated, where, in the present case, the unknown function is the unknown density of a taste coefficient $\beta$. The unknown function can be thought of as a point in an infinite-dimensional parameter space. Rather than trying to estimate a point in an infinite-dimensional space, one estimates over an approximating finite-dimensional parameter space. As the dimension of the approximating space grows, the resulting estimate approaches the true unknown function in many circumstances (Chen, 2006). Additionally, the dimension of the approximating
space can increase with the size of the dataset such that better approximations to
the true function are obtained for larger datasets. In econometrics, the resulting
estimators are known as semi-nonparametric (Gallant and Nychka, 1987).

There are various ways of approximating an infinite-dimensional space of dis-
tributions by finite-dimensional spaces. In this paper, we shall confine attention
to just two convenient possibilities and we shall fix the number of parameters to
be estimated, corresponding to the dimension of the approximating space, at low
values. What we obtain is thus just some very flexible distributions with more
parameters than desired in a very straightforward way, as discussed in Section 2.2.

The first approach we consider is that described by Fosgerau and Bierlaire
(2005). The main feature of this approach is that it can use any continuous dis-
tribution as its base. This is then extended by means of Legendre polynomials
such that any continuous distribution can be approximated at the limit, provid-
ing it has support within the support of the base distribution. The number of
parameters can be increased one by one by increasing the number of Legendre
polynomials used. Fosgerau and Bierlaire (2005) present the technique as a test
of the appropriateness of the base distribution, used by testing the model with
additional terms against the base model. Here, we simply use the resulting model
as a flexible means of retrieving random taste heterogeneity.

The other approach that we consider employs a mixture of distributions
(MOD) estimator. Specifically, we make use of a discrete mixture of Normal
distributions with different means and variances that are to be estimated, where
such a mixture of Normals can approximate any continuous distribution. In
existing work, Coppejans (2001) considers the MOD estimator for the case of
cross-sectional binary choice data, deterministic taste coefficients but randomly
distributed error terms, parallelling the estimator of Klein and Spady (1993). As
such, our use of the idea of a finite mixture of Normals is somewhat different.
Another discussion on mixtures of Normal distribution is given by Geweke and

In this paper, we will not provide theoretical results concerning consistency
and asymptotic properties of the estimators of the distribution of $\beta$ that we
employ. Fosgerau and Nielsen (2005) proves consistency of an estimator of $\beta$ in a
case when the distribution of the unobserved component of utility $\varepsilon$ is unknown.
It seems feasible to extend this result to the case of a MMNL model with an
unknown mixing distribution.

The remainder of this paper is organised as follows. The following section
presents the empirical framework used in the analysis. This is followed in Section
3 by a discussion of the results. Finally, Section 4 presents the conclusions of the
analysis.

3
2 Empirical framework

This section presents the empirical framework used for our Monte Carlo analysis. We first discuss the generation of the various datasets (Section 2.1), before describing the various methods that are used for the representation of random taste heterogeneity in our study (Section 2.2).

2.1 Generation of data

The setup used in this analysis makes use of binary choice data in a panel context. The use of panel data is crucial, since otherwise it becomes hard to distinguish the distribution of $\beta$ from the distribution of $\epsilon$. The conditional indirect utility function for the first alternative is set to zero, while, in choice situation $t$ for respondent $n$, the utility of the second alternative is given by:

$$U_{n,t} = \alpha_n + v_{n,t} + \frac{1}{\mu} \varepsilon_{n,t},$$

where $\varepsilon$ follows a logistic distribution, $v_{n,t}$ is an observed quantity, and $\alpha_n$ is an individual-specific latent random variable. This is the simplest possible setup that allows us to identify the distribution of an unobserved random parameter. This simplicity is a virtue, since we can then focus on the issue at hand, namely the ability of different estimators to recover a true distribution.

We simulate datasets of a size that is realistic in applied situations, containing 1,000 “individuals” making 8 “choices” each. We generate data for seven different choices of true distribution for $\alpha_i$, with details given below. The observed variable $v$ is drawn from a standard normal distribution, while the scale parameter $\mu$ is fixed at a value of 2.

It is important to realise that results from a single experiment can be influenced by randomness, such that it is impossible to reach general conclusions. To this extent, we generate 50 datasets for each distributionootnote{This is clearly only possible in a Monte Carlo study, and is not applicable when using real-world data. However, it is in that context possible to use bootstrap methods to generate confidence intervals around the estimated distribution. These confidence intervals can then be used to learn how much is determined from the data about the estimated distribution.}. Estimating the models many times for each true distribution of $\alpha$ allows us to take into account the fact that the estimates are random variables obtained as functions of random data. Altogether, we generate 50 datasets for each of the seven true distributions, leading to 350 datasets.

The seven true distributions were chosen with the aim of representing a wide array of possibilities that challenge our ability to estimate them.
point here is to select the distributions such that they lie well within the support of \(v_{n,t}\) which is standard normal. To this extent, we have selected the distributions to lie mostly within the interval [-2,2].

Specifically, we use the following seven datasets:

**DM(2) data:** discrete mixture with two support points, \(\alpha = -1\) with probability \(\pi_1 = 0.5\), and \(\alpha = 1\) with probability \(\pi_2 = 0.5\)

**DM(3) data:** discrete mixture with three support points, \(\alpha = -1\), \(\alpha = 0\) and \(\alpha = 1\), with equal mass of \(\pi_1 = \pi_2 = \pi_3 = \frac{1}{3}\)

**LN data:** Lognormal shifted to the left, generated by \(\alpha = \exp(u)/2 - 1\), where \(u \sim N(0,1)\)

**N data:** Standard normal, \(\alpha \sim N(0,1)\)

**NM data:** Normal with point mass at zero. With probability \(\pi_1 = 0.8\), \(\alpha \sim N(-1,1)\), and with probability \(\pi_2 = 0.2\), \(\alpha = 0\)

**2N data:** Mixture of two Normals, with \(\pi_1 = 0.5\), \(\alpha \sim N(-1,0.5)\), and with \(\pi_2 = 0.5\), \(\alpha \sim N(1,0.5)\)

**U data:** Uniform distribution, \(\alpha \sim U[-1,1]\)

### 2.2 Estimated distributions

In this section, we discuss the two main method compared in this analysis, with the Fosgerau-Bierlaire approach described in Section 2.2.1, and the MOD approach described in Section 2.2.2. This is followed in Section 2.2.3 by a brief description of various continuous distributions estimated in addition to the two main methods.

#### 2.2.1 Fosgerau & Bierlaire approach

Let \(\Phi\) be the standard normal cumulative distribution function with density \(\phi\) and let \(G\) be an absolute continuous distribution with density \(g\). We take \(\Phi\) as the base distribution with which we seek to estimate the true distribution \(G\).

Since both \(\Phi\) and \(G\) are increasing, it is possible to define \(Q(x) = G(\Phi^{-1}(x))\) such that \(Q(\Phi(\beta)) = G(\beta)\). Furthermore, \(Q\) is monotonically increasing and ranges from 0 to 1 on the unit interval. Thus, \(Q\) is a cumulative distribution function for a random variable on the unit interval. Denote by \(q\) the density of this variable, which exists since \(G\) is absolute continuous. Then we can express the true density as \(g = q(\Phi)\phi\).
Consider now a model $P(y|v, \alpha)$ conditional on the random parameter $\alpha$ which we assume to have the true distribution $G$. Then the unconditional model is

$$P(y|v) = \int_\alpha P(y|v, \alpha)g(\alpha)d\alpha$$

$$= \int_x P(y|v, \Phi^{-1}(x))q(x)dx$$

Thus the problem of finding the unknown density $g$ is reduced slightly to that of finding $q$, an unknown density on the unit interval. The terms $\Phi^{-1}(x)$ are just standard normal draws used in numerical simulation of the likelihood (cf. Train, 2003).

Now, let $L_k$ be the $k^{th}$ Legendre polynomial on the unit interval (cf. Bierens, 2006; Fosgerau and Bierlaire, 2005). These functions constitute an orthonormal base for functions on the unit interval$^2$ such that $\int L_k L_{k'}$ is equal to 1 when $k = k'$ and zero otherwise. We can then write

$$q(x) = \frac{(1 + \sum_{k \geq 1} \gamma_k L_k)^2}{1 + \sum_k \gamma_k^2}.$$  

(3)

Squaring the numerator ensures positivity, while the normalisation in the denominator ensures that $q(x)$ integrates to 1. Thus this expression is in fact a density. Bierens (2006) proves that any density on the unit interval can be written in this way.

The choice of Legendre polynomials is not a necessity. There are numerous other bases for functions on the unit interval that could have been used. Legendre polynomials are convenient because they have a recursive definition that is easily implemented on a computer.

To define the estimator that we use in this paper, we simply select a cut-off $K$ for $k$, such that we only use the first $K$ terms of (3). Thus we have a representation of a flexible $q_K$ with $K$ parameters. This is inserted into equation (2) to enable estimation by maximum likelihood. For more details on this approach, see Fosgerau and Bierlaire (2005).

2.2.2 Mixtures of distributions approach

In our MOD approach, we combine a standard continuous mixture approach with a discrete mixture approach, as described for example by Hess et al. (2006b). Specifically, the target distribution is a discrete mixture of several independently

$^2$See Bierens (2006) for a precise definition of this and following statements in this paragraph.
distributed Normal distributions. As such, we define a set of mean parameters, \( \mu_k \) and a corresponding set of standard deviations, \( \sigma_k \), with \( k = 1, \ldots, K \). For each pair \((\mu_k, \sigma_k)\), we then define a probability \( \pi_k \), where \( 0 \leq \pi_k \leq 1 \), \( \forall k \), and where \( \sum_{k=1}^{K} \pi_k = 1 \). A draw from the mixture distribution is then produced on the basis of two uniform draws \( u_1 \) and \( u_2 \) contained between 0 and 1, where we get:

\[
\alpha = \Phi^{-1}_{\mu_1, \sigma_1}(\mu_1), \quad \text{if } u_2 < \pi_1 \\
\alpha = \Phi^{-1}_{\mu_k, \sigma_k}(\mu_1), \quad \text{if } \sum_{l=1}^{k-1} \pi_l \leq u_2 < \sum_{l=1}^{k} \pi_l \text{ with } 1 < k \leq K - 1 \\
\alpha = \Phi^{-1}_{\mu_K, \sigma_K}(\mu_1), \quad \text{if } \sum_{l=1}^{K-1} \pi_l \leq u_2,
\]

(4)

where \( \Phi^{-1}_{\mu_k, \sigma_k} \) is the inverse cumulative distribution of a Normal with mean \( \mu_k \) and standard deviation \( \sigma_k \). With \( k \) normal terms, the resulting distribution allows for \( k \) separate modes, where the different modes can differ in mass. However, the flexibility of this approach is not limited to allowing for multiple modes. Additionally, the method can allow for saddles in a distribution, and can also allow for a point-mass at a specific value, in which case the associated standard deviation parameter becomes 0. In general, it is difficult to make a case for mass-points in a continuous distribution. However, there is one exception, namely a heightened mass at zero. This is useful in the representation of taste heterogeneity for attributes that some individuals are indifferent to, a concept discussed for example in the context of the valuation of travel time savings (VTTS) by Cirillo and Axhausen (2006). It can also be useful in the context of attribute processing strategies in SP data, with some respondent ignoring certain attributes, such that they obtain a zero coefficient (cf. Hensher, 2006).

An illustration of the flexibility of this approach is given in Figure 1, which shows cumulative distribution functions (CDF) for various examples of a mixture of two Normal distributions. In the first example, the only parameter that changes is \( \pi_1 \) (and hence by extension also \( \pi_2 \)), where, with \( \pi_1 = 1 \), we have a standard Normal distribution, with the shape gradually changing as we increase the mass for the second Normal, \( \pi_2 \). The second example illustrates the potential of the method to retrieve a point mass at a given value. Here, the standard deviation for the second support point, \( \sigma_2 \) is gradually decreased, where, with \( \sigma_2 = 0 \), we get a point mass of 50% at a value of 0 (\( \mu_2 = 0 \)), with the CDF turning into a step function at a value of 0. In the third example, the two support points have mean values at \(-2\) and 2, and share a common standard deviation, while \( \pi_1 = \pi_2 = 0.5 \). As we gradually increase the standard deviations, we move from a distribution
with two separate peaks (with little mass in between) to a distribution looking like a Normal with a very high variance. In the final example, we again have two Normals with equal standard deviation, fixed at 0.5, along with equal probabilities \( \pi_1 = \pi_2 = 0.5 \), and a mean for the first Normal fixed at \(-2\). As the mean of the second Normal is gradually decreased from its initial value of \(2\), we move from a distribution with two separate peaks to a distribution approximating a Normal.

### 2.2.3 Other distributions

Along with the approaches from Section 2.2.1 and Section 2.2.2, we also estimated models making use of a set of standard continuous distributions, as commonly used in Mixed Logit analyses. Here, we limit the set of distributions to the Normal, the Uniform, the symmetrical Triangular and the Johnson \(S_B\). These four distributions are the most commonly used in practice.

### 3 Results

In this section, we discuss the results of the Monte Carlo analysis carried out to compare the different methods for representing random taste heterogeneity. Given the high number of models estimated, only summary results across runs can be presented here. Along with the two advanced models, identified as MMNL(2N)\(^3\) and MMNL(SNP)\(^4\), and the four more basic models, identified as MMNL(N), MMNL(U), MMNL(T) and MMNL(S\(_B\)), a standard Multinomial Logit (MNL) model was estimated on the data.

Three different criteria are used in the presentation of the results. These are the model fit obtained in estimation, the performance in recovering the shape of the true distribution, and the performance in recovering the mean of the true distribution. A combination of tables and graphs are used in the presentation of the results.

- Table 1 shows the performance of the various models in terms of final log-likelihood (LL) obtained in estimation. Here, we give the mean LL obtained across the fifty runs in each model and dataset combination, along with the 5\(^{th}\) and 95\(^{th}\) percentiles of the distribution of the LL measure across runs, giving an indication of the stability of the methods.

- The results from Table 1 are complemented by a graphical representation of the differences in model fit across the fifty runs and across the various

\(^3\)Mixture of two Normals.

\(^4\)Semi non-parametric approach from Section 2.2.1 (cf. Fosgerau and Bierlaire, 2005).
models, where, given the much lower fit obtained by the MNL models, this comparison is limited to the mixture models. This graphical analysis is shown in Figure 2 for the DM(2) data, Figure 3 for the DM(3) data, Figure 4 for the LN data, Figure 5 for the N data, Figure 6 for the NM data, Figure 7 for the 2N data, and Figure 8 for the U data. In each case, two plots are shown, the first being a time series plot showing the actual LL values obtained by each model in each of the fifty runs, while the second plot shows the CDF for the distribution of the LL measure across runs in the different models.

- Table 2 shows the performance of the various models in terms of the recovery of the mean of the true distribution. In the absence of any consistent bias, especially when the true mean is close to zero, we use the absolute difference between the estimated mean and the true mean. As in the case of the LL results, the table again shows the mean value across the fifty runs, along with 5% and 95% quantiles.

- Finally, the performance of the various methods in terms of the recovery of the shape of the true distribution is illustrated with the help of CDF plots for the true and estimated distributions, where, for the latter, the mean CDF across runs is presented alongside a 90% confidence band for the CDF. The various plots are shown in Figure 9 for the DM(2) data, Figure 10 for the DM(3) data, Figure 11 for the LN data, Figure 12 for the N data, Figure 13 for the NM data, Figure 14 for the 2N data, and Figure 15 for the U data. Vertical distances in these CDF plots correspond to the L-infinity norm; indeed, in the space of CDFs, convergence of estimates to the true distribution, as the number of terms increases, takes place in L-infinity norm. When working in the space of probability density functions (PDF), convergence takes place in L-1 norm. However, this is harder to interpret visually, such that preference is given to CDF plots and hence the L-infinity norm in the present analysis.

We will now proceed with a discussion of the results obtained in the various datasets.

**DM(2) data:** For the data generated by a discrete mixture with two support points, the best model fit is obtained by MMNL(2N) and MMNL(SB), ahead of MMNL(SNP), with the former two having an advantage of about 20 units in LL. The next best performance is obtained by MMNL(U), which maintains a sizeable advantage in LL over MMNL(T) and MMNL(N). The variability in performance across the fifty runs is quite moderate, with the
range being just over 3% of the mean LL, where there is little variation across the six MMNL models. This is reflected in the first plot in Figure 2, as well as in the roughly parallel lines in the second plot. The relative range for the MNL model is narrower, but the overall performance is much poorer than that for the MMNL models.

In terms of the recovery of the true mean (cf. Table 2), the best performance is obtained by MMNL(SP) and MMNL(2N), while the remaining models, including MNL, obtain roughly similar performance, apart from MMNL(U), which gives the poorest approximation.

Finally, the results in terms of model fit are very consistent with the findings from Figure 9, which shows that MMNL(2N) and MMNL(SB) outperform MMNL(SNP) in the recovery of the shape of the true distribution, while the approximation given by MMNL(U), MMNL(T) and MMNL(N) is much poorer.

**DM(3) data:** For the data generated by a discrete mixture with three support points, the best performance is obtained by MMNL(SB) and MMNL(2N), with MMNL(U) and MMNL(SNP) very close behind. Finally, MMNL(T) again outperforms MMNL(N). The variation in performance across the fifty runs is again very similar for the six models, as shown in Figure 3. Overall, the differences between the six models are much smaller than was the case with the DM(2) data, which could be an indication of the difficulty of dealing with more than two peaks. This is reflected in the plots of the CDF functions in Figure 10. Here, the use of more than two support points in the MOD model would be a possible avenue for improvements. This however leads to a very significant increase in complexity in estimation; the exploration of efficient ways of estimating DOM models with $K > 2$ is the topic of ongoing work. Finally, just as for model fit, there are also only minor differences in the performance in terms of recovering the mean of the true distribution.

**LN data:** For the data generated by a Lognormal distribution, the best performance is again obtained by MMNL(SB) and MMNL(2N), just ahead of MMNL(SNP). These three models outperform the remaining three MMNL structures by between 54.19 and 84.77 units. Again, there are virtually no differences across the six models in terms of variability across the fifty runs, as illustrated in Figure 4. The advantages of MMNL(SB), MMNL(2N) and MMNL(SNP) are also clearly visible in Figure 11. Finally, in terms of the recovery of the mean of the true distribution, notably poorer performance is obtained by MMNL(T) and especially MMNL(N).
**N data:** For the data generated with a standard Normal distribution, there are virtually no differences in performance across the various models, where, as expected, only MMNL(U) gives slightly poorer performance. This is also reflected in Figure 5, and Figure 12. All six models perform similarly well in the recovery of the true mean.

**NM data:** For the data generated with a Normal with an added mass at 0, the differences in performance are smaller than could have been expected, a fact that can partly be explained by the low probability (20%) for the mass-point at 0. Nevertheless, we can observe that MMNL(2N) and MMNL(SNP) offer the best performance, ahead of MMNL(SB), while MMNL(U) has a slight advantage over MMNL(T) and MMNL(N) (see also Figure 6). These findings are consistent with the plots in Figure 13. This time, MMNL(U) actually offers the best performance in the recovery of the true mean, ahead of MMNL(2N) and MMNL(SNP), while MMNL(N) again performs very poorly.

**2N data:** For the data generated by a mixture of two Normals, the MOD model MMNL(2N) obtains the best model fit, just ahead of MMNL(SNP) and MMNL(SB). MMNL(U) outperforms MMNL(T), while MMNL(N) again produces the poorest model fit (see also Figure 7). These findings are again consistent with the plots in Figure 14, which nicely illustrates the problems faced by the Normal and the Triangular. Finally, this time, the differences in performance in terms of the recovery of the true mean are more marginal.

**U data:** For the final dataset, generated with a Uniform distribution, the performance of the various models is very similar, where only the two symmetrical distributions MMNL(N) and MMNL(T) lead to slightly poorer model fit (see also Figure 8). All models perform equally well in the recovery of the true mean. The problems in the tails of MMNL(N) and MMNL(T) are illustrated in Figure 15.

### 4 Conclusions

This paper has reported the findings of a systematic study using Monte Carlo experiments aimed at comparing the performance of various methods in retrieving random taste heterogeneity in a discrete choice context. Specifically, the analysis has compared the performance of the four most commonly used continuous distribution functions, the Normal, symmetrical Triangular, Uniform and Johnson SB, to that of two more advanced approaches discussed in this paper. The first of
these two approaches improves on the flexibility of a base distribution by adding in Legendre polynomials, where, in the present study, the Normal distribution was chosen as the base. The second approach uses a discrete mixture of multiple continuous distributions, where again, in the present study, the base distributions are all Normal.

The framework for the analysis compared the performance of the six resulting models across seven separate case studies, making use of different assumptions for the true distribution of the single random parameter in the model. In each case study, fifty random versions of the data were generated to allow us to gauge the stability of the various approaches.

When looking solely at the performance in terms of model fit, some interesting observations can be made:

- The model making use of the MOD approach, MMNL(2N), obtains the best performance in 4 out of the 7 case studies, while it comes second in the remaining 3 case studies.

- The model making use of the Normal distribution, MMNL(N), obtains the poorest performance in 5 out of the 7 case studies, ranking 4th and 5th (out of six models) in the remaining two case studies.

- With two exceptions, MMNL(N) is outperformed by MMNL(U), the model making use of the Uniform distribution, while MMNL(T), the model making use of the symmetrical Triangular distribution, offers better performance than MMNL(N) in all but one of the seven case studies.

- Although not the most scientific indicator of model performance, when looking at the average LL across datasets and runs (i.e. the 350 estimated models), MMNL(2N) obtains the best performance, just ahead of MMNL(SB) and MNL(SNP). MMNL(U), which follows at a distance of some 26 units, has an advantage of around 8 units over MMNL(T), which in turn obtains marginally better average LL than MMNL(N).

The findings in terms of model fit are broadly consistent with the findings in terms of the recovery of the shape of the true distribution, while there is no overly consistent pattern in terms of the recovery of the true mean.

Perhaps the most striking observation from this analysis is the poor performance of the Normal distribution. This is very worrying indeed, given the overwhelming reliance on this distribution in current Mixed Logit practice. On the other hand, the two advanced approaches discussed in this paper seem to perform very well across all case studies, suggesting that they can deal with a variety
of true distribution, ranging from the most trivial (Uniform) to more complex multi-modal distributions.

Here, it should be said that very good performance was also obtained by the models based on the Johnson SB distribution. However, as was the case in previous studies (e.g. Hess et al., 2006a; Hess and Rose, 2006), the use of the SB led to slow convergence, and occasional problems with parameter significance.

In a direct comparison between the two advanced approaches discussed in this paper, we can conclude that they are very similar in their ability to approximate smooth distributions. While the MOD approach has a slight advantage over the proposed SNP approach in terms of model fit, it seems that, on the basis of the various CDF plots, it has a slightly higher variance. For non-smooth distributions, the MOD approach has the ability to become degenerate and have a point mass. This may be viewed as an advantage if one believes in mass-points, a concept that, in a discrete choice context, only really makes sense for a mass-point at zero. However, this degeneracy is also a problem for the ability of the estimator to approximate smooth distributions and the estimator must be constrained in some way (cf. Coppejans, 2001).

The flexibility of either of the two approaches can be increased by estimating additional parameters, in terms of additional Legendre polynomials, or additional support distributions in the MOD approach. This does however invariably lead to increased estimation cost, and, with the MOD approach especially, issues arise with convergence to local minima. Here, an advantage of the SNP approach is that we only need to add one parameter at a time, while, with the MOD approach, we need to add three parameters at the same time (two distributional parameters and one probability). Finally, both approaches are not restricted to being based on the Normal distribution, but can use any conditional distribution as the base.

Both approaches are relatively easy to implement, where the SNP approach has already been implemented in BIOGEME (Bierlaire, 2003), and where estimation code for the MOD approach is available from the first author on request. Here, it should also be noted that the potential of these approaches is not limited solely to the estimation of models. Indeed, they can also be seen as a diagnostic tool that can be used to get an idea of the shape of the true distribution; this knowledge can then be used in the choice of an appropriate continuous distributions, which may lead to reductions in the cost of model application. This concept has already been exploited by Fosgerau and Bierlaire (2005) with the SNP approach discussed here.

Several avenues for further research can be identified. These relate to the testing of the two approaches in more complex scenarios, such as in the presence of multiple random coefficients with potential correlation between them. Additionally, it is of interest to test the methods in the face of more complex
true distributions than those used here. Finally, the estimation of more complex versions of the two methods, making use of more Legendre polynomials or more support points for the mixture, remains an important topic of research. Here, one aim would be to analyse the relative benefit of the more complex formulations when taking into account the rise in estimation cost.

Acknowledgements

Part of the work described in this paper was carried out during a guest stay by the first author in the Institute of Transport and Logistics Studies at the University of Sydney.

References


Figure 1: CDF plots for various mixtures of two Normal distributions
<table>
<thead>
<tr>
<th>Data</th>
<th>$5^{th}$ perc.</th>
<th>mean</th>
<th>$95^{th}$ perc.</th>
<th>$5^{th}$ perc.</th>
<th>mean</th>
<th>$95^{th}$ perc.</th>
<th>$5^{th}$ perc.</th>
<th>mean</th>
<th>$95^{th}$ perc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMNL(N)</td>
<td>-4610.75</td>
<td>-3597.03</td>
<td>-3539.93</td>
<td>-3588.84</td>
<td>-3459.84</td>
<td>-3458.78</td>
<td>-4416.58</td>
<td>-3845.56</td>
<td>-3828.10</td>
</tr>
<tr>
<td>MMNL(U)</td>
<td>-4359.76</td>
<td>-3779.46</td>
<td>-3765.29</td>
<td>-3774.11</td>
<td>-3763.92</td>
<td>-3764.05</td>
<td>-4319.65</td>
<td>-3718.80</td>
<td>-3701.53</td>
</tr>
<tr>
<td>MMNL(T)</td>
<td>-4359.76</td>
<td>-3779.46</td>
<td>-3765.29</td>
<td>-3774.11</td>
<td>-3763.92</td>
<td>-3764.05</td>
<td>-4319.65</td>
<td>-3718.80</td>
<td>-3701.53</td>
</tr>
<tr>
<td>MMNL(SB)</td>
<td>-4359.76</td>
<td>-3779.46</td>
<td>-3765.29</td>
<td>-3774.11</td>
<td>-3763.92</td>
<td>-3764.05</td>
<td>-4319.65</td>
<td>-3718.80</td>
<td>-3701.53</td>
</tr>
<tr>
<td>MMNL(2N)</td>
<td>-4359.76</td>
<td>-3779.46</td>
<td>-3765.29</td>
<td>-3774.11</td>
<td>-3763.92</td>
<td>-3764.05</td>
<td>-4319.65</td>
<td>-3718.80</td>
<td>-3701.53</td>
</tr>
<tr>
<td>MMNL(SNP)</td>
<td>-4359.76</td>
<td>-3779.46</td>
<td>-3765.29</td>
<td>-3774.11</td>
<td>-3763.92</td>
<td>-3764.05</td>
<td>-4319.65</td>
<td>-3718.80</td>
<td>-3701.53</td>
</tr>
<tr>
<td>LN</td>
<td>-4180.21</td>
<td>-3830.51</td>
<td>-3858.48</td>
<td>-3840.27</td>
<td>-3781.38</td>
<td>-3780.31</td>
<td>-4121.12</td>
<td>-3774.34</td>
<td>-3802.94</td>
</tr>
<tr>
<td>N</td>
<td>-4068.89</td>
<td>-3721.06</td>
<td>-3750.64</td>
<td>-3731.79</td>
<td>-3671.59</td>
<td>-3669.97</td>
<td>-4170.18</td>
<td>-3654.83</td>
<td>-3665.71</td>
</tr>
<tr>
<td>NM</td>
<td>-4045.70</td>
<td>-3542.04</td>
<td>-3539.53</td>
<td>-3542.45</td>
<td>-3525.01</td>
<td>-3519.86</td>
<td>-4004.33</td>
<td>-3501.18</td>
<td>-3494.42</td>
</tr>
<tr>
<td>2N</td>
<td>-3961.25</td>
<td>-3447.67</td>
<td>-3437.88</td>
<td>-3445.89</td>
<td>-3428.16</td>
<td>-3424.99</td>
<td>-4750.37</td>
<td>-3677.86</td>
<td>-3654.00</td>
</tr>
<tr>
<td>U</td>
<td>-4030.36</td>
<td>-3781.03</td>
<td>-3779.44</td>
<td>-3779.13</td>
<td>-3776.79</td>
<td>-3776.65</td>
<td>-4148.12</td>
<td>-3896.62</td>
<td>-3893.18</td>
</tr>
</tbody>
</table>

Table 1: Model fit statistics across datasets and models
<table>
<thead>
<tr>
<th>Data</th>
<th>5&lt;sup&gt;th&lt;/sup&gt; perc.</th>
<th>Mean</th>
<th>95&lt;sup&gt;th&lt;/sup&gt; perc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNL</td>
<td>0.0009</td>
<td>0.0173</td>
<td>0.0385</td>
</tr>
<tr>
<td>MMNL(N)</td>
<td>0.0008</td>
<td>0.0186</td>
<td>0.0415</td>
</tr>
<tr>
<td>MMNL(U)</td>
<td>0.0024</td>
<td>0.0228</td>
<td>0.0537</td>
</tr>
<tr>
<td>MMNL(T)</td>
<td>0.0016</td>
<td>0.0188</td>
<td>0.0448</td>
</tr>
<tr>
<td>MMNL(SB)</td>
<td>0.0014</td>
<td>0.0191</td>
<td>0.0395</td>
</tr>
<tr>
<td>MMNL(2N)</td>
<td>0.0005</td>
<td>0.0111</td>
<td>0.0274</td>
</tr>
<tr>
<td>MMNL(SNP)</td>
<td>0.0010</td>
<td>0.0107</td>
<td>0.0275</td>
</tr>
<tr>
<td>DM(2)</td>
<td>0.0018</td>
<td>0.0124</td>
<td>0.0269</td>
</tr>
<tr>
<td>5&lt;sup&gt;th&lt;/sup&gt; perc.</td>
<td>0.0002</td>
<td>0.0113</td>
<td>0.0331</td>
</tr>
<tr>
<td>Mean</td>
<td>0.0018</td>
<td>0.0157</td>
<td>0.0348</td>
</tr>
<tr>
<td>95&lt;sup&gt;th&lt;/sup&gt; perc.</td>
<td>0.0009</td>
<td>0.0142</td>
<td>0.0294</td>
</tr>
<tr>
<td>DM(3)</td>
<td>0.0013</td>
<td>0.0130</td>
<td>0.0283</td>
</tr>
<tr>
<td>5&lt;sup&gt;th&lt;/sup&gt; perc.</td>
<td>0.0009</td>
<td>0.0120</td>
<td>0.0282</td>
</tr>
<tr>
<td>Mean</td>
<td>0.0013</td>
<td>0.0120</td>
<td>0.0264</td>
</tr>
<tr>
<td>95&lt;sup&gt;th&lt;/sup&gt; perc.</td>
<td>0.0008</td>
<td>0.0120</td>
<td>0.0264</td>
</tr>
<tr>
<td>LN</td>
<td>0.0020</td>
<td>0.0150</td>
<td>0.0323</td>
</tr>
<tr>
<td>5&lt;sup&gt;th&lt;/sup&gt; perc.</td>
<td>0.0025</td>
<td>0.0194</td>
<td>0.0386</td>
</tr>
<tr>
<td>Mean</td>
<td>0.0032</td>
<td>0.0189</td>
<td>0.0448</td>
</tr>
<tr>
<td>95&lt;sup&gt;th&lt;/sup&gt; perc.</td>
<td>0.0013</td>
<td>0.0198</td>
<td>0.0407</td>
</tr>
<tr>
<td>N</td>
<td>0.0020</td>
<td>0.0150</td>
<td>0.0323</td>
</tr>
<tr>
<td>5&lt;sup&gt;th&lt;/sup&gt; perc.</td>
<td>0.0025</td>
<td>0.0194</td>
<td>0.0386</td>
</tr>
<tr>
<td>Mean</td>
<td>0.0032</td>
<td>0.0189</td>
<td>0.0448</td>
</tr>
<tr>
<td>95&lt;sup&gt;th&lt;/sup&gt; perc.</td>
<td>0.0013</td>
<td>0.0198</td>
<td>0.0407</td>
</tr>
<tr>
<td>NM</td>
<td>0.0272</td>
<td>0.0578</td>
<td>0.0839</td>
</tr>
<tr>
<td>5&lt;sup&gt;th&lt;/sup&gt; perc.</td>
<td>0.0130</td>
<td>0.0479</td>
<td>0.0786</td>
</tr>
<tr>
<td>Mean</td>
<td>0.0011</td>
<td>0.0173</td>
<td>0.0406</td>
</tr>
<tr>
<td>95&lt;sup&gt;th&lt;/sup&gt; perc.</td>
<td>0.0076</td>
<td>0.0360</td>
<td>0.0581</td>
</tr>
<tr>
<td>2N</td>
<td>0.0009</td>
<td>0.0182</td>
<td>0.0377</td>
</tr>
<tr>
<td>5&lt;sup&gt;th&lt;/sup&gt; perc.</td>
<td>0.0011</td>
<td>0.0207</td>
<td>0.0477</td>
</tr>
<tr>
<td>Mean</td>
<td>0.0018</td>
<td>0.0248</td>
<td>0.0602</td>
</tr>
<tr>
<td>95&lt;sup&gt;th&lt;/sup&gt; perc.</td>
<td>0.0007</td>
<td>0.0211</td>
<td>0.0507</td>
</tr>
<tr>
<td>U</td>
<td>0.0021</td>
<td>0.0126</td>
<td>0.0250</td>
</tr>
<tr>
<td>5&lt;sup&gt;th&lt;/sup&gt; perc.</td>
<td>0.0003</td>
<td>0.0138</td>
<td>0.0293</td>
</tr>
<tr>
<td>Mean</td>
<td>0.0009</td>
<td>0.0130</td>
<td>0.0288</td>
</tr>
<tr>
<td>95&lt;sup&gt;th&lt;/sup&gt; perc.</td>
<td>0.0022</td>
<td>0.0136</td>
<td>0.0327</td>
</tr>
</tbody>
</table>

Table 2: Recovery of true mean by various models (absolute difference between actual and estimated mean)
Figure 2: Model fit statistics on DM(2) data

Figure 3: Model fit statistics on DM(3) data
Figure 4: Model fit statistics on LN data

Figure 5: Model fit statistics on N data
Figure 6: Model fit statistics on NM data

Figure 7: Model fit statistics on 2N data
Figure 8: Model fit statistics on U data
Figure 9: CDF plots for $\alpha$ in models estimated on DM(2) data
Figure 10: CDF plots for $\alpha$ in models estimated on DM(3) data.
Figure 11: CDF plots for $\alpha$ in models estimated on LN data.
Figure 12: CDF plots for $\alpha$ in models estimated on N data.
Figure 13: CDF plots for $\alpha$ in models estimated on NM data.
Figure 14: CDF plots for $\alpha$ in models estimated on 2N data
Figure 15: CDF plots for $\alpha$ in models estimated on U data.