A guide to heterogeneity features captured by parametric and nonparametric mixing distributions for the mixed logit model

Yuan Yuan* Wen You† Kevin J Boyle‡

Selected Paper prepared for presentation at the
2015 Agricultural & Applied Economics Association and
Western Agricultural Economics Association Joint Annual Meeting,
San Francisco, CA, July 26–28

*PhD candidate, Department of Agricultural and Applied Economics, Virginia Tech
✆ 202.360.8291, ✉ y.yuan@vt.edu
†Associate Professor, Department of Agricultural and Applied Economics, Virginia Tech
‡Director, Program in Real Estate, Virginia Tech

A guide to heterogeneity features captured by parametric and nonparametric mixing distributions for the mixed logit model

Yuan Yuan, Wen You, Kevin Boyle

July 26–28

Abstract

Unobserved heterogeneity is popularly modelled using the mixed logit model, so called because it is a mixture of standard conditional logit models. Although the mixed logit model can, in theory, approximate any random utility model with an appropriate mixing distribution, there is little guidance on how to select such a distribution. This study contributes to suggestions on distribution selection by describing the heterogeneity features which can be captured by established parametric mixing distributions and more recently introduced nonparametric mixing distributions, both of a discrete and continuous nature. We provide empirical illustrations of each feature in turn using simple mixing distributions which focus on the feature at hand.

1 Introduction

Heterogeneity in behaviour is widely acknowledged to be a fundamental aspect of discrete choice modelling (Hess et al., 2005; Desarbo et al., 1997; Allenby and Rossi, 1998), arising from differences in individual tastes, attitudes and perceptions, decision strategies, and other factors. Consequently, heterogeneity can affect many different parts of the choice model specification, including, for example, the taste parameters (taste variation), values of the attributes (perceptual variation), functional form (structural variation), and the error term (scale variation). The consequences of ignoring heterogeneity are similar to the consequences of other forms of misspecification: biased estimates and misleading policy implications (Desarbo et al., 1997).

Perhaps the most popular way of taking heterogeneity into account is the mixed logit model, which is defined by a choice probability expressed as a mixture of standard logit probabilities (Train, 2009):

$$p_{ij} = \int L_{ij}(\beta) f(\beta; \theta) d\beta \tag{1}$$

where p_{ij} is the probability that individual i will choose alternative j from a choice set, $f(\beta;\theta)$ is the density of the parameters β, θ are *hyperparameters* which parameterize the density of β[1], and L_{ij} is the standard choice probability for the conditional logit model

$$L_{ij}(\beta) = \frac{\exp(\beta' x_{ij})}{\sum_k \exp(\beta' x_{ik})}. \tag{2}$$

assuming a linear-in-parameters specification for the systematic utility. Inspection of Equation 1 reveals the origin of the name 'mixed logit': the mixed logit model is a mixture model where the components are conditional logit models and the mixing distribution is given by $f(\beta)$. Hence, the mixed logit model is also known as the 'mixed multinomial logit model' and the 'logit kernel model'.

To the author's knowledge, there is no one study which comprehensively covers all the mixing distributions now available, old and new, parametric and nonparametric, continuous and discrete. Most of the relatively recent mixing distributions have seen limited use in empirical applications, possibly because many practitioners are not yet aware of them. On the other hand, those practitioners facing the full gamut of available mixing distributions may find specification an overwhelming task. In order to raise awareness of new mixing distributions and guide selection of mixing distributions, this study will describe mixing distributions in terms of the features of heterogeneity which they model. Although some studies have focused on one feature or another, this study is unique in its emphasis on features relevant to the research goal at hand and the coverage of multiple features. This study encourages practitioners to consider the *nature* of the heterogeneity which may be present, and choose the appropriate mixing distributions based on data, theory or policy relevance. For example, the data may reveal evidence of skew in the preference distribution, which can bias the mean from a policy standpoint. Another example comes from theory, which usually suggests that cost coefficients should be negatively signed, implying a distribution which is at least bounded on one side. Finally, if the policy question is about identifying different behavioural segments of customers (as in market segmentation, a fundamental concept in marketing), a preference distribution with multiple modes is indicated.

Wedel et al. (1999) points out that selecting an appropriate mixing distribution is especially difficult because specifying the form of parameter heterogeneity is largely an empirical issue. Yet his call for research into the 'theoretical underpinning of heterogeneity, with the purpose of identifying variables that need to be included in models and to assist researchers in the appropriate model specification' has not been answered in more

[1]For the rest of the text, we will use the term 'hyperparameter' to differentiate between parameters which enter the standard logit probability (parameter) and parameters which parameterize the mixing distribution (hyperparameter).

than 10 years. However, he also suggests that 'empirical generalizations could help form theoretical foundations for the description of heterogeneity'.

Thus, the main contribution of the study is to comprehensively describe common and alternative mixing distributions, including raising awareness of the more recently introduced alternative mixing distributions, and with particular emphasis on the set of heterogeneity features which each mixing distribution captures. A secondary contribution of this study is an empirical search for each feature explored in this study using a case study in health economics, thus adding to the body of work which can form the empirical generalizations Wedel et al. (1999) call for.

2 Parametric mixing distributions

The multivariate normal distribution is probably the most widely used mixing distribution for MXLs. Its appeal can probably be attributed, in part, to the same reason which makes the normal distribution so common in more general usage: the central limit theorem. When the distributional form is unknown, researchers can at least appeal to the CLT to justify their usage of the normal distribution. Besides this general rationale, the normal distribution also provides advantages specific to the mixed logit model. Identifying the normal mixing distribution only requires two hyperparameters per random coefficient, and the interpretation of the hyperparameters is intuitive since they directly correspond to mean and standard deviation. Finally, estimating a mixed logit model with a normal mixing distribution is relatively fast and stable, and estimation routines are widely available in discrete choice modelling software packages.

On the other hand, the normal mixing distribution also suffers from a number of drawbacks, including assumptions of unbounded support, unimodality, symmetry and, frequently, mutual independence among random parameters.

2.1 Unbounded support

The long tails of the normal distribution imply very positive and very negative values for taste parameters, which may be behaviourally implausible. For taste parameters which researchers have *a priori* sign expectations (*e.g.*, cost coefficient), support on both sides of zero will often imply implausibly large proportions of the distribution with the 'wrong' sign. Counterintuitive signs, particularly on cost coefficients, are a major driver of research into alternative mixing distributions (Hess et al., 2005; Train and Sonnier, 2005; Hess et al., 2006; Train, 2008; Rigby et al., 2009; Bastin et al., 2010; Campbell et al., 2010; Cirillo and Hetrakul, 2010; Hess, 2010; Chalak et al., 2012; Bastani and Weeks, 2013; Keane and Wasi, 2013). Finally, if the cost coefficient is specified as random with support which spans zero, WTP estimates will have infinite moments (Daly et al., 2012). Hensher and

Greene (2003) and Hess et al. (2005, 2006) recommend bounded mixing distributions, emphasizing that the bounds should be estimated from the data. Once the mixing distribution is bounded, behaviourally implausible extreme values for taste parameters are eliminated, and coefficients with sign expectations are forced to be consistent with theory.

Many parametric alternatives offer bounded support, including the **uniform**, **triangular**, **log-normal**, **censored or truncated normal**, and the **Johnson** S_B. The uniform and triangular have bounds which are estimated from the data, although they are frequently thought too simplistic for realism. The log-normal, censored normal and truncated normal, which all have bounds at 0, are primarily employed when sign expectations are to be met. On the other hand, their heavy tails (on the unbounded side of their support) may bias mean estimates and their bounds are theoretically dictated rather than estimated from data. The Johnson S_B is an interesting case because it is parameterized by four parameters, two of which represent bounds. The analyst may choose either to fix those parameters or to freely estimate them from the data. Due to the greater number of hyperparameters associated with the Johnson S_B compared to the other bounded distributions, empirical identification could be problematic, hence encouraging the use of fixed bounds.

2.2 Unimodal

The normal distribution is limited to a single mode, but economic theory rarely provides guidance on how many modes a mixing distribution should have. On the other hand, a central concept in marketing theory is market segmentation, the view that a heterogeneous market is composed of smaller homogeneous sub-markets (Wedel and Kamakura, 2000). The policy implication of multiple modes under a marketing context can easily be seen. Consider, for example, Campbell and Doherty (2013), which studied the demand for value-added services to chicken which improve food safety and quality (*e.g.*, food testing standards, traceability standards, animal health/welfare standards). They found that the demand for value-added services came from a niche market segment; while this segment was willing to pay a price premium for the services, other consumers were not willing to pay any price premium. A preference distribution failing to account for the multimodal nature of this preference distribution would have led to a misleading marginal WTP and revenue predictions. In particular, the niche market segment would have been charged a lesser price premium than they would actually be willing to pay, and the other consumers would not be willing to pay the price premium at all. Another way to think of the segmentation problem is that if one person wants a red balloon, and the other a blue balloon, a purple balloon satisfies no one. Under a multimodal distribution, the mode estimated under a unimodal distribution is unlikely to be located on any of the true modes.

However, accommodating multiple modes is not always necessary, even if they do exist. Suppose only population-level summary statistics, such as the population mean, are called for in a given application. Then a more restrictive mixing distribution with only one mode would still yield approximately the same measures as a more flexible mixing distribution which allows multiple modes.

Among the parametric distributions, only the Johnson S_B distribution permits more than one mode. All the other parametric distributions are unimodal like the normal, except for the uniform, which technically has no mode since all values in the support range are equally likely.

2.3 Symmetrical

The normal distribution is symmetric by definition, but economic theory is usually silent on whether a mixing distribution should be symmetric or asymmetric. Balcombe et al. (2011) motivates the consideration of skew with an example attribute, 'genetic modification', which could be associated with extreme disutility by some individuals but, at the same time, unlikely to be matched by extreme positive utility by other individuals. The authors then go on to develop a new mixing distribution which is a transformation of the normal distribution, but unlike previous transformations, can accommodate positive, negative or zero skew. In their empirical study of attitudes to bovine breeding technologies, they found significant evidence of skew, even after accounting for one potential source of skew, attribute non-attendance[2]. Thus, to arrive at an accurate picture of the taste heterogeneity, a flexible mixing distribution which makes no assumption on the presence or direction of skew may be necessary.

Both the triangular and Johnson S_B distributions can accommodate positive, negative or zero skew. Often, however, the asymmetrical variant of the Johnson S_B is used while the symmetrical variant of the triangular distribution is used (Hess et al., 2006; Fosgerau and Hess, 2009; Hess, 2010). The log-, censored and truncated normal distributions are asymmetric by construction, and their direction of skew is also fixed. The log-normal is positively skewed by definition, and the censored and truncated normals are skewed depending on which part of the distribution is censored or truncated.

2.4 Mutually independent random parameters

The conventional multivariate normal mixing distribution can theoretically accommodate a full correlation matrix between random parameters, but in practice, many applications assume mutual independence for tractability (Train and Sonnier, 2005; Hynes et al., 2008;

[2] Attribute non-attendance is the behaviour of ignoring some attributes when performing a choice task. This behaviour is econometrically equivalent to a zero utility for the ignored attributes. If the true distribution of marginal utility for the attribute is symmetric, then the presence of attribute non-attendance would skew the distribution, since it confounds the true distribution with a point mass at zero.

Hess, 2014). Those applications which do permit full correlation have found the presence of significant correlation, better model fit, substantively different taste parameter estimates and important policy implications (Train and Sonnier, 2005; Hynes et al., 2008; Mabit et al., 2008; Rigby et al., 2009). Furthermore, correlation between random parameters is a consequence of scale heterogeneity (Hess and Rose, 2012). To see why, consider again the standard logit probability expression (Equation 2), except this time with the omitted scale parameter α present:

$$L_{ij}(\beta) = \frac{\exp(\alpha\beta' x_{ij})}{\sum_k \exp(\alpha\beta' x_{ik})}. \tag{3}$$

If scale heterogeneity is present, then α is individual-specific (*i.e.*, α_i). Since the scale parameter and the taste parameter are confounded and not separately identified (*i.e.*, only $\alpha\beta$ is identified), scale heterogeneity will scale all taste parameters simultaneously and equally across individuals. Thus, if taste heterogeneity is absent but scale heterogeneity is present, the random parameters will display perfect correlation. If scale heterogeneity is present but correlation is not permitted, the scale heterogeneity will manifest elsewhere, such as in the standard errors of the parameter estimates. Consequently, correlation between random parameters accommodates scale heterogeneity. As with taste heterogeneity, there is rarely an *a priori* reason not to suspect the presence of scale heterogeneity.

The parametric mixing distributions can all, in theory, be estimated with full correlation. When using a parametric mixing distribution, each random parameter is assigned its own distribution, which is tantamount to specifying the marginal distributions of the mixing distribution. Then, correlation between the marginal distributions can be induced using a decomposition of the variance-covariance matrix. However, correlation only captures linear relationships; if there are higher order relationships between the taste parameters, inducing correlation is not enough.

The alternative parametric distributions described above have been explored and compared against the multivariate normal in a number of studies (*e.g.* Hensher and Greene, 2003; Hess et al., 2005; Train and Sonnier, 2005; Hess et al., 2006; Fosgerau and Hess, 2009; Rigby et al., 2009; Cirillo and Hetrakul, 2010; Chalak et al., 2012), and in general, the bounded, more flexible distributions are preferred for their ability to avoid behaviourally implausible taste parameters and to capture complex features of the parameter distribution such as asymmetry. However, the most flexible distributions are the nonparametric distributions[3].

[3]The term 'nonparametric' is used loosely here and elsewhere in the paper to mean any distribution which is not strictly parametric, thus encompassing semiparametric and seminonparametric approaches as well as fully nonparametric techniques.

3 Nonparametric distributions

To organize this section, we categorize the nonparametric mixing distributions into discrete and continuous. All the discrete nonparametric mixing distributions are related to the latent class logit model, while the continuous nonparametric mixing distributions are all examples of sieve estimators. While the latter may be more parsimonious and less sensitive to tuning parameters (such as analyst-imposed bounds on the support of the distribution), the former can be more flexible in some cases, particularly with respect to dependence between random parameters.

3.1 Discrete nonparametric distributions

The latent class logit model is a well-established example of a nonparametric mixing distribution which has found favour in multiple disciplines because it is intuitive and easy to interpret. The latent class logit model is an example of a finite mixture model, in which the mixing distribution is discrete and there are a finite number of support points. Thus, the number of support points limits the richness of preference heterogeneity which can be described by the latent class logit model. The more support points in the distribution, the higher the resolution, so to speak, of the distribution, thus more clearly revealing potential heterogeneity features such as symmetry and modality. In practice, the number of support points which can be estimated is relatively small since the number of hyperparameters grows linearly with the number of support points specified, and estimation issues are usually encountered when there are too many hyperparameters. As a result, variants of the finite mixture theme have been introduced in order to increase the number of support points which can be estimated. We first describe the original latent class logit model and then describe the variants which have recently been introduced.

3.1.1 Latent class logit model

The **latent class logit model** can be described behaviourally as follows: suppose there are S latent segments in the population. Within the segments, preferences are homogeneous, but across segments, preferences may be heterogeneous. Econometrically, the latent class logit model can be described as a mixed logit model with a finite mixing distribution. The class membership probabilities and class-specific preference parameters are the probability masses and mass point locations, respectively, of a probability mass function (pmf) describing the distribution of a discrete random variable. This contrast is the reason the latent class logit model is seen as the discrete analog of the typical random parameters logit model, which is specified with a continuous, parametric distribution.

The flexibility of the latent class logit model grows with the number of classes. Since the number of hyperparameters grows linearly with the number of classes, classes are typically limited to less than a dozen. If the analyst specifies too many classes, then the estimation will frequently fail to converge or experience other problems. Consequently, variants on the latent class logit model which are more parsimoniously parameterized, such as the mass point MXL and the fixed point MXL, have been introduced.

3.1.2 Mass point mixed logit model

The **mass point MXL** is essentially a reparameterization of the latent class logit model. It was first introduced in Dong and Koppelman (2003) and later compared against other mixing distributions in simulation studies and empirical applications (Hess et al., 2007; Campbell et al., 2010; Dong and Koppelman, 2014). In this model, the number of support points is parameter-specific. The distinction between the mass point MXL and the latent class logit model is subtle, and lies entirely in how many support points can be specified with a given number of hyperparameters. In the latent class logit model, the analyst specifies the number of segments S, and each segment corresponds to a *joint* support point. Thus, there are as many joint support points as there are segments. In the mixed point MXL, the analyst specifies the number of *marginal* support points for each parameter, and so the number of joint support points is equal to the product of the number of support points for each parameter. Consequently, the same number of joint support points can be specified using a different number of hyperparameters.

3.1.3 Fixed point mixed logit model

In the latent class logit model and mass point MXL, both the locations of and the probability masses at each support point are freely estimated. In this next model, the locations of the support points are fixed, and only the probability mass at each point is estimated. Accordingly, we term this model the **fixed point MXL**. The fixed point MXL was first introduced by Bajari et al. (2007), and compared against other mixing distributions in simulation studies and empirical applications by Train (2008) and Bastani and Weeks (2013). The chief benefit of the fixed point MXL compared to the other discrete mixing distributions is the ability to estimate many more joint mass points. Whereas latent class logit models and mass point MXL often fail to converge or yield degenerate solutions[4] when the number of support points is only a handful, the fixed point MXL is fast and stable even when estimating probability masses for hundreds of thousands support points (Train, 2008). However, the results are sensitive to the analyst's specification of the support point locations, and in particular the range of the parameter space. In response to this weakness, Bastani and Weeks (2013) developed some heuristics to address range selection.

[4]Solutions in which probability masses are close to zero for some segments or some segment locations are very close together.

All these variants on the latent class logit model have similar properties. They all have bounded support, since the number of support points is specified by the analyst and therefore finite. They can accommodate multiple modes, asymmetry and arbitrary dependence between random parameters because the joint distribution is directly estimated.

3.2 Continuous nonparametric distributions

The mass point MXL, fixed point MXL and fixed mass MXL arose in response to the limited heterogeneity which traditional latent class logit models could capture. At the same time, a separate branch of the literature reacted to the same problem in a different way, by using *continuous* nonparametric mixing distributions rather than discrete ones.

3.2.1 Mixture of distributions mixed logit model

The most obvious approach for increasing the heterogeneity captured by the latent class logit model is to extend it so that each segment has random rather than fixed taste parameters. This model has been called by different names, including random parameters latent class logit model and latent class mixed multinomial logit model, which recognize its origins with the latent class logit model (Bujosa et al., 2010; Campbell and Doherty, 2013; Greene and Hensher, 2013). However, the formulation is equivalent to specifying the mixing distribution as a **mixture of distributions** itself (Dong and Koppelman, 2003; Train, 2008; Fosgerau and Hess, 2009; Campbell et al., 2010, 2014; Fosgerau, 2014). Setting the base distribution to be normal, $G(\cdot) = \Phi(\cdot)$, is particularly appealing because any continuous distribution can be approximated arbitrarily well by a finite mixture of normals. A further advantage of using a mixture of distributions is the ability to accommodate point masses, since the distribution within any given segment may become degenerate. Thus, the mixture of distributions can represent both discrete and continuous types of heterogeneity. The complexity of this model increases with the number of mixtures included, and most applications use only two or three components.

In fact, the mixture of distributions model is an example of a sieve estimator, defined as an estimator which approximates unknown functions with a series of basis functions. In the case of the mixture of distributions MXL, the basis functions are the base distribution of the mixture, such as the normal in a mixture of normals. The quality of the approximation depends on the basis functions and the number of terms in the series. If the basis function is a good approximation to the unknown function, then only a small number of terms should be necessary (Fosgerau, 2014). The method of sieves has been a popular approach for developing nonparametric MXL mixing distributions, and below we describe other sieve estimators which have also been proposed.

3.2.2 Other sieve estimators

Fosgerau and Bierlaire (2007) introduced a **sieve estimator with Legendre polynomials** as the basis function. The series of Legendre polynomials seminonparametrically approximate the derivative of a transformation function, rather than the mixing distribution directly.

Bastin et al. (2010) introduced a different **sieve estimator with B-splines** as the basis functions. In this approach, the spline function approximates the inverse cdf of a random coefficient, hence the domain is $[0,1]$. The knot locations are fixed on $[0,1]$ by the analyst and the coefficients on the basis functions are estimated. The spline function is guaranteed to be nondecreasing (and hence a proper inverse cdf) if the coefficients on the basis functions are also nondecreasing, a property which is achieved during estimation through constrained optimization.

3.3 Discrete or continuous?

Historically, the debate between continuous and discrete representations of heterogeneity has centred around the MXL with a parametric mixing distribution and the latent class logit model. The argument against the continuous representation is that it is parametric and therefore subject to misspecification, whereas the argument against the discrete approach is that it is too restrictive because it assumes homogeneity within each segment. Given the advanced mixing distributions which have been discussed in the previous two sections, these arguments are clearly outdated. The sieve estimators with distributions, Legendre polynomials and B-splines represent heterogeneity in a continuous manner while avoiding strict parametric assumptions. The fixed point MXL estimated on a fine grid of support points is a discrete representation which can capture far more detailed heterogeneity than any analytical distribution can display. What remains at the heart is the question: what is the true nature of heterogeneity, continuous or discrete?

This question is, practically speaking, unanswerable. At the same time, there are other, more relevant questions we could ask to determine whether to choose a continuous or discrete representation of heterogeneity. In reflecting on this debate, Wedel and Kamakura (2000, p 329) have the following to say:

> In applying models to segmentation, one should recognize that every model is at best a workable approximation of reality. One cannot claim that segments really exist or that the distributional form of unobserved heterogeneity is known.

The more relevant question we should ask ourselves is: which is the more *useful* representation, continuous or discrete? Wedel and Kamakura (2000) suggest that continuous representations are suitable for individual level forecasting, while discrete representations are particularly useful for understanding the structure of heterogeneity in the population. Segmentation is a core concept in marketing, not because it is necessarily more true than a continuous representation of heterogeneity, but because it has proven to be useful over and over again: it is accessible, compelling and actionable to managers and other end users of information resulting from marketing studies. On the other hand, when applied economists wish to generate population-level welfare estimates to be used in policymaking techniques such as cost-benefit analysis, segments are unnecessary, confusing and can lead to bias due to oversimplification from reducing a continuous distribution to a discrete distribution with a finite number of support points. If, however, policymakers are interested in the composition of 'winners' and 'losers' of a policy change, then segments once again become useful. To choose a continuous or a discrete representation, analysts should identify which levels of aggregation are appropriate for their context, research question and audience.

Continuous and discrete representations of heterogeneity can be complementary, each enriching the insight which can be gained from the other. In one example, Hynes et al. (2008) investigated preference heterogeneity among kayakers for whitewater sites in Ireland by estimating both a conventional MXL and a latent class logit model. With the conventional MXL, they established the presence of heterogeneity in the sample. With the latent class logit model, they were able to match the latent classes revealed by the model to specializations within the kayaking sport, with intuitive taste parameter estimates. For managers of the whitewater sites, this type of information is useful because it allows them to identify the mix of kayakers patronizing different sites and tailor responses to their preferences. In another example, Arunotayanun and Polak (2011) investigated mode choice heterogeneity among freight shippers. They used conventional MXL to establish the presence of heterogeneity even after the sample was split by commodity, the standard practice in this context, revealing the inadequacy of this segmentation scheme. They used the latent class logit model to identify alternative segments which were behaviourally driven instead, and systematically related them to shipper and shipment characteristics, leading to a new segmentation scheme. These examples illustrate how continuous and discrete mixing distributions can both be used to identify behaviourally and policy relevant heterogeneity.

For completeness, we bring to the reader's attention a nonparametric approach developed by Rouwendal et al. (2010) which does not rely on a statistical model and is not fully identified, but rather seeks to identify individual-specific valuations for attribute levels within a given dataset[5].

[5] Rouwendal et al. (2010) introduces a method of locating individual-specific valuations of attributes by viewing each choice as revealing an inequality in valuations between attribute levels. For concreteness, we illustrate the core concept using a very simple example. Suppose that two bundles differ only in one attribute. If a respondent chooses bundle A over B, then he must have a higher valuation for the attribute level in

4 Parametric or nonparametric?

In Table 1, we summarise the key properties of each mixing distribution which have been discussed above. If flexibility and capturing as many features of heterogeneity as possible were the main goals of the analyst, then the nonparametric mixing distributions such as the fixed point and mixture of distributions MXL would be appropriate. However, more flexibility is not always better. Flexible mixing distributions may be overly complex, suffering from weak identification, overfitting and difficulty of interpretation. If the structure of the heterogeneity is simple, a simple mixing distribution would be more parsimonious, efficient, robust and interpretable (Keane and Wasi, 2013).

However, knowing whether heterogeneity is complex or simple prior to choosing a mixing distribution is difficult. Flexible mixing distributions can play a role here, by acting as a preliminary tool diagnosing which features of heterogeneity are present (Fosgerau and Hess, 2009)[6]. For example, if multiple modes are identified, then the Johnson S_B or nonparametric mixing distributions may be indicated. If a point mass at zero is identified, but the rest of the distribution appears smooth, then a censored normal or a latent class with fixed and random segments may be indicated. Campbell and Doherty (2013) implemented the latter approach to investigate WTP for value-added services to chicken. They hypothesized that some consumers would be indifferent to the value-added services, while others would have non-zero preferences. They accommodated these types by assigning each to a latent class: in one class, WTP was fixed to 0 for the former type of consumer, and in the other, WTP was allowed to be random. This approach allowed them to identify market niches, estimate demand responses to price premiums in those niches,

bundle A than bundle B, thus revealing an inequality. Over the sequence of multiple choices, the space spanned by the inequalities shrinks, and ideally, is exactly identified. In an empirical application of the method, valuations were identified exactly in a few cases, and in most cases only bounds (which may be very wide) were identified. Furthermore, many respondents chose inconsistently, so that the space satisfying all inequalities was empty. Unlike the previously described approaches, this method avoids approximation and attempts to discover the specific valuations of each respondent in the dataset. It is not a statistical model which extends the valuations beyond the dataset at hand, and for most empirical cases the valuations will not be fully identified.

[6]Flexible mixing distributions are not the only such preliminary tool. Hensher and Greene (2003) propose an entirely different approach to 'reveal' the empirical distribution of heterogeneity. In this approach, a series of 'leave-one-out' MNL models are estimated, with each individual in the dataset 'left out' a different MNL model. The difference between the parameter estimate under a leave-one-out MNL model and under the pooled MNL model represents the individual's contribution to the mean parameter estimate. Considering all the differences from all of the leave-one-out MNL models presents a picture of individual heterogeneity. As a different approach, Hess (2010) suggests using conditional distributions (*i.e.*, respondent-specific coefficient distributions, conditional on the respondents' responses) to form a picture of the empirical distribution. However, he cautions against simply using the means of the conditional distributions, since doing so fails to take into account the heterogeneity around those means, within each conditional distribution. Conceptually, this approach is similar to checking one's prior with the estimated posterior in a Bayesian analysis.

Mixing distribution	Support	Number of modes	Symmetry	Dependence
Parametric				
Normal	unbounded	1	symmetrical	correlation
Uniform	bounded	–	symmetrical	correlation
Triangular	bounded	1	a/symmetrical	correlation
Log-normal	bounded on one side	1	asymmetrical	correlation
Censored or truncated normal	bounded on one side	1	asymmetrical	correlation
Johnson S_B	bounded	2	a/symmetrical	correlation
Nonparametric discrete				
Latent class logit	bounded	n	a/symmetrical	arbitrary
Mass point	bounded	n	a/symmetrical	arbitrary
Fixed point	bounded	n	a/symmetrical	arbitrary
Nonparametric continuous				
Mixture of distributions	depends on component distributions	n	a/symmetrical	arbitrary
Sieve estimator with Legendre polynomials	depends on base distribution	n	a/symmetrical	independent
Sieve estimator with B-splines	bounded	n	a/symmetrical	independent

Table 1: Summary of properties by mixing distribution.

and describe demographics of those niches which could aid in marketing efforts.

Using ten datasets, Keane and Wasi (2013) compared some of the mixing distributions described in this study, including the latent class logit model and the mixture of distributions approach. Although the emphasis of their study was on finding models of best fit, they found that the mixture of distributions model was superior to simpler mixing distributions in its ability to accommodate complex heterogeneity, in which small subpopulations of individuals have strong preferences for attributes to which most other individuals are indifferent. However, the mixture of distributions model accommodated these individuals by placing them into the tails of the component distributions. The latent class logit model, on the other hand, distinguished these individuals and placed them into their own segments. Even though the latent class logit model consistently scored among the lowest in terms of log-likelihood, it provided insight into the structure of heterogeneity hidden by other mixing distributions. Thus, more flexible distributions do not always provide more insight.

Practical considerations may also affect the choice of mixing distribution. In particular, mixing distributions vary in how easy they are to estimate, due to differences in empirical identifiability and available estimation strategies.

4.1 Estimation issues

Generally speaking, the more parameters a mixing distribution has, the more difficult it may be to empirically identify. For example, the Johnson S_B is a transformed normal with four parameters: two represent the mean and variance of the underlying normal and the other two represent the lower and upper bound of the distribution. Train and Sonnier (2005) suggests that the bounds are difficult to identify because the difference between them is closely related to the variance of the underlying normal. Consequently, Rigby et al. (2009) imposed bounds instead of allowing them to be freely estimated, Chalak et al. (2012) conducted a repetitive search for the bounds, and Cirillo and Hetrakul (2010) could not always achieve convergence during model estimation. The Johnson S_B is not the only parametric mixing distribution to suffer from estimation issues: the log-normal distribution occasionally does so as well (Hensher and Greene, 2003; Hess et al., 2006).

The nonparametric mixing distributions also vary in terms of empirical identification, but the degree to which they suffer does not seem to be closely related to the number of parameters involved. The latent class logit model, which is relatively parsimonious compared to, say, the fixed point MXL, often cannot be estimated beyond a handful of latent classes, as mentioned earlier. The mixture of distributions approach suffers from the same problem (*e.g.* Fosgerau and Hess, 2009), and few applications have attempted estimating more than two components in this model. The sieve estimators with Legendre polynomials and B-splines, on the other hand, have not presented analysts with any convergence problems. Similarly, the virtue and motivating reason to consider the fixed point MXL is its ability to remain fast and stable even under a huge number of points to estimate (Train, 2008).

Another factor affecting ease of estimation is what strategies are available for each mixing distribution. All of the continuous mixing distributions, parametric and nonparametric, are estimated by maximum simulated likelihood estimation (MSLE). The discrete mixing distributions, on the other hand, do not require simulation and can be estimated directly via maximum likelihood. Avoiding simulation is desirable because simulation increases computation time, introduces simulation bias, and requires a 'sufficient' number of draws in order to produce stable estimates. Unfortunately, determining this number is primarily an empirical matter (Hensher and Greene, 2003).

Train (2008) suggested using the expectation-maximization (EM) algorithm to avoid numerical problems which may be encountered during maximum likelihood estimation of models with a large number of parameters (*i.e.*, as typically found in nonparametric mixing distributions). He developed and illustrated the EM algorithm for several nonparametric mixing distributions, including the latent class logit model, the mixture of distributions (normals) MXL and the fixed point MXL. However, in a comparison of MSLE, the EM algorithm, and one other estimation strategy, Cherchi and Guevara (2012) found that as long as sample sizes were sufficient, MSLE outperformed the EM algorithm with respect to estimation time and estimation error. At the same time, as expected, MSLE suffered from weak empirical identification with insufficient sample sizes, while the EM algorithm was more robust to this issue. Other applications of the EM algorithm in the literature are limited, although Bastani and Weeks (2013) used it to estimate a fixed point MXL and latent class logit model, and Pacifico (2013) used it to estimate a latent class logit model. Neither study reported problems with using the EM algorithm.

5 Empirical illustration

In this section we seek empirical evidence for the properties discussed above using a selection of simple and advanced mixing distributions. We use as our baseline the multivariate normal distribution with mutual independence, which could be considered the most frequently used, 'default' MXL mixing distribution. We then compare this baseline model against a set of MXL with alternative mixing distributions. Each alternative distribution is chosen with an eye towards a particular property identified as a potentially relevant feature of heterogeneity while still maintaining as much simplicity as possible so as to focus the comparison on the target property.

Previous empirical studies of mixing distributions have been dominated by the transportation, environmental and marketing domains, but this empirical illustration uses a choice experiment in health economics. The health domain is particularly interesting in this context because many of the heterogeneity features described previously are differently applicable to health preferences. For example, in transportation economics, signed cost coefficients have been a prime driver in considering bounded supports, but in health economics, cost/reward attributes may have counterintuitive signs because consumers may load them with other meanings beyond the purely financial. Thus bounded supports, or at least supports which have bounds not estimated from data, may be less desirable in a health context. Another example is multimodality: in the environmental context, benefit measures are important tools for policymaking, and in those cases, it is the benefit across large populations which are relevant. However, policymaking can often be more customized in the health context, due to the ability of individual programs, hospitals, and care providers to tailor their care to the particular subpopulation (which may not be representative of the general population) which utilizes their services, or which they hope to target with their services. Thus multimodality can be an actionable policy goal in the health context, allowing customization and targeting of desired subpopulations.

This study uses a choice experiment on the design of financial incentives for a weight loss program to illustrate common and alternative mixing distributions. The financial incentives were described using five attributes: reward amount, program location, payment form, reward condition and payment frequency. There were 1,296 respondents, each of which were presented with four choice tasks. A total of 96 choice sets were constructed according to the D-efficiency criterion.

5.1 Baseline distribution

The baseline model uses a normal mixing distribution with mutually independent random parameters. We specify the systematic utility to be a simple linear form, with attributes as the only terms. We specify the coefficients on reward amount and program locations to be random while the coefficients on the other attributes are left fixed. We limit the number of random parameters in order to facilitate clarity in comparison across the alternative distributions. We chose reward amount and program location to have random coefficients because previous work has indicated that are the two most important attributes, and focus group discussions have suggested heterogeneous preferences for their levels.

Table 2 displays the coefficient estimates for the baseline distribution. Note that all variables are effect-coded, except for reward amount, which is continuous. We can see that there is significant preference heterogeneity for reward amount and all program locations except for the community center.

	Baseline model
Fixed parameters	
ASC	-0.16 (0.10)
Mag.0	0.09 (0.11)
Form: gym / cash	-0.14 (0.05)**
Form: medical / cash	-0.19 (0.05)***
Form: debit / cash	0.14 (0.05)**
Condition: weight / attendance	0.08 (0.05)
Condition: compliance / attendance	-0.05 (0.05)
Condition: attendance and compliance / attendance	-0.12 (0.05)*
Frequency: weekly / once	0.15 (0.05)**
Frequency: monthly / once	0.00 (0.05)
Frequency: quarterly / once	0.05 (0.05)
Random parameter means	
log(amount + 1)	0.35 (0.04)***
Location: workplace / clinic	-0.22 (0.05)***
Location: community center / clinic	0.18 (0.04)***
Location: church / clinic	-0.21 (0.05)***
Random parameter standard deviations	
s.log(amount + 1)	0.57 (0.03)***
s.Location: workplace / clinic	0.60 (0.10)***
s.Location: community center / clinic	0.24 (0.18)
s.Location: church / clinic	-0.65 (0.10)***
Log-likelihood	-4454.83
N	4994.00

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2: Coefficient estimates for baseline model, normal mixing distribution with mutually independent random parameters.

5.2 Bounded alternative

Since we have a positive sign expectation on reward amount and no such expectations for the location levels, we expect boundedness to be a more acute need for the reward amount coefficient than the location coefficients. The bounded alternative distribution we use is the uniform mixing distribution. Like the normal mixing distribution, it is also described by two parameters, the minimum a and maximum b of the support. Alternatively, the uniform mixing distribution can also be parameterized by the center c and spread s of the support. This alternative distribution is widely available in discrete choice modelling software.

Table 3 displays the coefficient estimates for both the baseline model and the uniform mixing distribution. We can see that across all random coefficients, the normal and uniform mixing distributions are quite similar to each other in terms of location and spread. Additionally, they all display both negative and positive preferences. Some readers may find a noticeable proportion of reward amount preferences in the negative domain unusual. However, focus groups held during the development of the survey instrument indicated that some respondents felt offended at the presence of a financial incentive in a weight loss program. Moreover, financial incentives in weight loss programs have raised some degree of ethical controversy in both the popular press and academic literature, and there are several other examples of transactions for which monetization is culturally distasteful (Roth, 2007; Halpern et al., 2009; Schmidt, 2011; Lunze and Paasche-Orlow, 2013).

For this particular application, the estimated uniform and normal mixing distributions were very similar across most aspects we considered, although the standard deviation of the estimated uniform mixing distribution was smaller, a natural result of the uniform's bounded nature. These results suggest that unless policy questions are concerned with the extreme values of the distributions, policy interpretations in this case are likely to be quite similar whichever distribution is used.

5.3 Correlated alternative

Correlation between random parameters can provide important policy insights on which attributes and attribute levels work in synergy and which work in opposition to each other. Unanticipated correlation may produce unintended policy consequences, and so identifying the sign and significance of correlations between random parameters has policy relevance. The correlated alternative distribution we use is the correlated normal mixing distribution. In the baseline model, the random parameters are assumed to be mutually independent, but in the correlated normal mixing distribution, the random parameters are permitted to have an arbitrary variance-covariance matrix. In practice, the variance-covariance matrix is estimated by estimating the terms of the Cholesky factor in

	Baseline	Uniform
Fixed parameters		
ASC	-0.16 (0.10)	-0.19 (0.10)
Mag.0	0.09 (0.11)	0.06 (0.10)
Form: gym / cash	-0.14 (0.05)**	-0.14 (0.05)**
Form: medical / cash	-0.19 (0.05)***	-0.19 (0.05)***
Form: debit / cash	0.14 (0.05)**	0.14 (0.05)**
Condition: weight / attendance	0.08 (0.05)	0.08 (0.05)
Condition: compliance / attendance	-0.05 (0.05)	-0.05 (0.05)
Condition: attendance & compliance / attendance	-0.12 (0.05)*	-0.12 (0.05)*
Frequency: weekly / once	0.15 (0.05)**	0.15 (0.05)**
Frequency: monthly / once	0.00 (0.05)	-0.00 (0.05)
Frequency: quarterly / once	0.05 (0.05)	0.05 (0.05)
Random parameter means/centers		
log(amount + 1)	0.35 (0.04)***	0.34 (0.04)***
Location: workplace / clinic	-0.22 (0.05)***	-0.21 (0.05)***
Location: community center / clinic	0.18 (0.04)***	0.18 (0.04)***
Location: church / clinic	-0.21 (0.05)***	-0.22 (0.05)***
Random parameter standard deviations/spreads		
s.log(amount + 1)	0.57 (0.03)***	0.88 (0.04)***
s.Location: workplace / clinic	0.60 (0.10)***	-0.97 (0.17)***
s.Location: community center / clinic	0.24 (0.18)	0.48 (0.28)
s.Location: church / clinic	-0.65 (0.10)***	1.14 (0.15)***
Log-likelihood	-4454.83	-4469.75
N	4994.00	4994.00

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3: Coefficient estimates for baseline model and uniform mixing distribution.

a Cholesky decomposition of the variance-covariance matrix (Croissant, 2013). Like the bounded alternative distribution, the correlated alternative distribution was also readily available in software packages.

Of most interest are the estimated covariance hyperparameters[7](Table 4). These indicate that there is no significant correlation in preferences between reward amount and location. However, there are significant negative correlations between location levels. In this case, preferences for different attributes are uncorrelated, but preferences for attributes within the same level are negatively correlated. In other words, an individual who prefers one location tends not to prefer the other locations. This result is intuitive, although we might expect some complementary location preferences if more location levels were available in the experiment. However, the location levels presented in the current application appear to be different enough to elicit preferences which view location levels more as substitutes.

Using a mixing distribution which permits correlation, we uncovered significant correlation between random coefficients for levels with the same attribute, as well as evidence of bias in the estimated standard deviation for those random coefficients affected by correlation. However, correlation only captures linear dependence. Nonlinear dependence may be present, but in practice, dependent random variables are generated using the Cholesky decomposition of a covariance matrix. Possible approaches to assessing the presence of higher order relations between random coefficients is using copulas to describe the dependence relation between parametric random coefficients, using a mixture of distributions, or using a nonparametric discrete mixture distribution which directly estimates the joint distribution and hence the dependence structure.

5.4 Discrete, nonparametric alternative

A discrete, nonparametric mixing distribution can help in identifying minority preferences which may otherwise become 'lost' in the tails of continuous mixing distributions. To ensure a broad reach of the incentivized weight loss program, it may be necessary to consider satisfying the wants of small minorities, in addition to designing incentives which attract the majority of the target population. The discrete, nonparametric alternative distribution we use is the latent class logit model (LCL). We have already discussed how the LCL can be seen as the discrete analog to the commonly encountered mixed logit, with a continuous, parametric mixing distribution. Estimating the class-specific coefficients and the class membership probabilities is essentially estimating the location and mass of mass points in a pmf.

[7]The covariance matrix hyperparameters were computed from the estimated Cholesky factor (as $\Sigma = C'C$), and the standard errors and Z-tests for significance of the hyperparameters were approximated using the delta method.

	Baseline	Correlated
Fixed parameters		
ASC	-0.16 (0.10)	-0.22 (0.11)
Mag.0	0.09 (0.11)	0.09 (0.11)
Form: gym / cash	-0.14 (0.05)**	-0.15 (0.05)**
Form: medical / cash	-0.19 (0.05)***	-0.20 (0.05)***
Form: debit / cash	0.14 (0.05)**	0.16 (0.05)**
Condition: weight / attendance	0.08 (0.05)	0.10 (0.05)
Condition: compliance / attendance	-0.05 (0.05)	-0.06 (0.05)
Condition: attendance & compliance / attendance	-0.12 (0.05)*	-0.12 (0.05)*
Frequency: weekly / once	0.15 (0.05)**	0.13 (0.06)*
Frequency: monthly / once	0.00 (0.05)	0.01 (0.05)
Frequency: quarterly / once	0.05 (0.05)	0.04 (0.05)
Random parameter means		
log(amount + 1)	0.35 (0.04)***	0.38 (0.04)***
Location: workplace / clinic	-0.22 (0.05)***	-0.25 (0.06)***
Location: community center / clinic	0.18 (0.04)***	0.18 (0.05)***
Location: church / clinic	-0.21 (0.05)***	-0.24 (0.05)***
Random parameter standard deviations		
s.log(amount + 1)	0.57 (0.03)***	
s.Location: workplace / clinic	0.60 (0.10)***	
s.Location: community center / clinic	0.24 (0.18)	
s.Location: church / clinic	-0.65 (0.10)***	
Random parameter covariance matrix terms		
amount.amount		0.36 (0.04)***
amount.workplace		0.01 (0.05)
amount.community		-0.02 (0.05)
amount.church		0.05 (0.05)
workplace.workplace		0.89 (0.25)***
workplace.community		-0.35 (0.12)**
workplace.church		-0.23 (0.12)*
community.community		0.49 (0.16)**
community.church		-0.25 (0.13)
church.church		0.88 (0.21)***
Log-likelihood	-4454.83	-4435.63
N	4994.00	4994.00

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4: Coefficient estimates for baseline model and correlated normal mixing distribution.

Rather than estimating the membership probabilities directly, we estimate the parameters θ_c, which are used in a multinomial membership model as follows

$$\pi_c = \frac{\exp(\theta_c)}{\sum_k \exp(\theta_k)}$$

where π_c are the class membership probabilities. Estimating θ_c rather than the membership probabilities directly affords a computational advantage, because θ_c are unconstrained whereas the class membership probabilities must be proper probabilities.

One of the most important aspects of the LCL specification is the number of latent classes. This parameter is not estimated but rather user-specified. For many applications of the LCL, there is no theoretical guidance on the number of latent classes. Since there are also no statistical tests available to select the number of classes, the practitioner instead uses a combination of goodness-of-fit measures and individual discretion. Table 5 presents goodness-of-fit measures for the LCL estimated under one to six classes. The measures include the well-known information criteria AIC and BIC, as well as McFadden's pseudo R-squared, adjusted for degrees of freedom, and entropy (Morey et al., 2006). The entropy statistic is bounded between 0 and 1, and a value closer to 1 indicates that the model successfully differentiates individuals into latent classes. Note that a one-class LCL is equivalent to a standard conditional logit model, and that the 'infinite'-class LCL refers to the baseline model.

Number of classes	AIC	BIC	$\bar{\rho}^2$	Entropy
1	10260.52	10337.77	0.05	
2	9052.70	9155.70	0.16	0.16
3	8947.54	9076.29	0.17	0.22
4	8887.46	9041.96	0.18	0.13
5	8845.68	9025.93	0.18	0.14
6	8873.50	9079.50	0.18	0.14
∞	8947.67	9045.51	0.17	

Table 5: Goodness-of-fit measures for the LCL under different numbers of classes.

According to three of the four goodness-of-fit measures, the LCL with five latent classes has the best fit, even compared to the baseline model. Usually LCL models do not fit as well as a continuous mixture, so this result can be seen as evidence for the appropriateness of a discrete mixing distribution. On the basis of these goodness-of-fit measures and the plausibility of five classes, we select the five-class LCL for further comparison to the baseline model.

Reassuringly, the modes under the estimated LCL pmf and the estimated baseline pdf

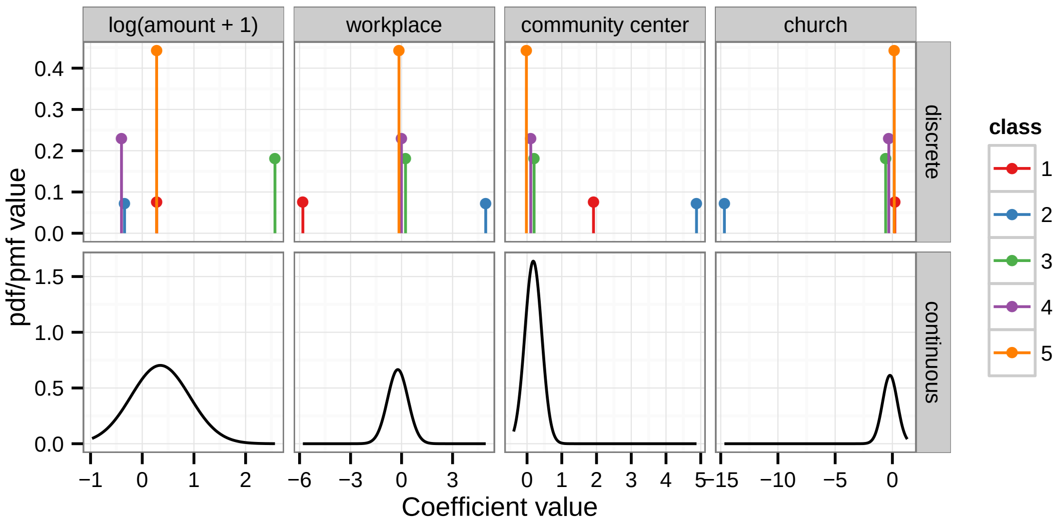

Figure 1: Discrete mixing distribution pmf and continuous mixing distribution pdf for each random parameter. The mass points are grouped by colours, representing the classes to which they belong. The colours and numbers associated with the classes are arbitrary.

are very similar for each random parameter (Figure 1). However, the LCL pmf often reveals preference masses in the tails of the normal mixing distribution, thus biasing the mean and standard deviations of the LCL pmf with respect to the normal mixing distribution. Revealing these masses in the tails has both advantages and disadvantages. On the one hand, preferences not readily observable in the normal mixing distribution were revealed. On the other hand, these masses bias summary statistics like means and standard deviations. Thus, which mixing distribution is more appropriate depends on the policy motivation; if 'minority' preferences are policy relevant, then the LCL may be more suitable. However, if the policy questions are primarily concerned with 'average' preferences across the whole population, then the normal mixing distribution may be a better, less biased choice.

6 Conclusion

In this study, we have described an extensive list of options for the choice of mixing distribution in a mixed logit model. Our descriptions focused on the theoretical ability of the distributions to describe heterogeneity features such as boundedness, multimodality, asymmetry and dependence between random parameters, as well as discussing practical considerations such as estimation issues and the importance of the policy context in determining relevant heterogeneity features.

Each mixing distribution differs in its theoretical properties, practical considerations and policy context. The commonly proposed parametric mixing distributions typically have relatively inflexible theoretical properties, but can be easier to estimate and interpret. In many policy contexts, the additional flexibility in theoretical properties is unnecessary, because the policy relevant questions are at a high level, and the detailed features of the preference heterogeneity present do not affect the policy implications and may only complicate matters.

However, when the policy context calls for detailed consideration of the preference heterogeneity features, nonparametric mixing distributions which are more flexible than the parametric mixing distributions may be appropriate. When market segments are policy relevant, the latent class logit model can yield rich insights into minority preferences which would be glossed over in less flexible distributions, especially with respect to arbitrarily complex dependence structures. At the same time, population-wide conclusions based on the latent class logit model may give too much weight to the minority preferences. Moreover, in practice, the complexity of the latent class logit model is strongly limited by estimation issues which arise when too many latent classes are specified. This limitation has given rise to a number of variations on the latent class logit model, which also estimate a pmf for the mixing distribution. However, the way in which they parameterize the support points is different, and so fewer hyperparameters are necessary to estimate the same number of support points. One weakness common to all the discrete nonparametric mixing distributions is the large extent to which results depend on analyst-specified tuning parameters, such as number and range of support points.

The continuous nonparametric mixing distributions are more forgiving, with results fairly robust to tuning parameters such as number and choice of basis function. These mixing distributions are all sieve estimators, which approximate unknown functions with a series of basis functions. They can also be more parsimonious than the discrete nonparametric mixing distributions in terms of numbers of estimated hyperparameters, while still maintaining a high degree of flexibility. Their chief weakness is their inability, at present, to accommodate dependence between random coefficients. The discrete nonparametric mixing distributions, in contrast, directly estimate the joint mixing distribution, and dependence between random coefficients is revealed as a by-product.

In our empirical comparison of the baseline mixing distribution, a multivariate normal with mutually independent random parameters, and alternative distributions which relax the theoretical properties of the baseline distribution, we found the biggest differences were in dependence structure. When we compared the baseline distribution with a correlated normal mixing distribution, we found significant correlation between different levels within the same attribute. Furthermore, when we compared the baseline distribution with the latent class logit model, we found evidence of dependence relationships which were more complex than the pairwise, linear relationships modelled by the correlated normal. The market segments identified by the latent class logit model suggested different designs to attract different segments.

In contrast, our empirical comparison of the baseline distribution with a bounded alternative, the uniform distribution, did not reveal substantial differences in policy implication. Although technically there were significant differences in standard deviation between the baseline and uniform distributions, these differences did not have much effect on the policy implications. The amount of the normal density found outside of the uniform density was modest, and evenly distributed between the upper and lower tails. This even distribution can be attributed to the symmetrical nature of both distributions and the similarity in their locations. Consequently, considering only a specific portion of the distribution, such as the negative preferences, again showed very small differences between the baseline and alternative distributions. Thus, unless the extreme values of the distribution (*e.g.*, top or bottom 1%) are policy relevant, there is little difference in policy implications between unbounded and bounded distributions.

When analysts have a choice of which heterogeneity features to model, it makes sense to start with a more flexible mixing distribution and then simplify as features are seen to be absent. Although we discussed this role flexible mixing distributions could play in section 4, doing so is not yet common practice. Future work could focus on this role, establishing which of the flexible distributions are most appropriate for aiding the specification search.

The importance of dependence structure, particularly moving beyond correlation into higher-order relationships, suggests that approaches for capturing these relationships is an important future direction for theoretical study. Given the large amount of data needed to estimate the joint pmf directly in nonparametric discrete mixing distributions, approaches which are elegant and parsimonious could be complementary to the existing techniques for detecting complex dependence structure. However, given the limitations of the copula approach and the difficulty with estimating highly multidimensional nonparametric surfaces, this direction appears to be quite challenging.

References

Greg M. Allenby and Peter E. Rossi. Marketing models of consumer heterogeneity. *Journal of Econometrics*, 89(1–2):57–78, 1998. ISSN 0304-4076. doi: 10.1016/S0304-4076(98)00055-4. URL http://www.sciencedirect.com/science/article/pii/S0304407698000554.

Kriangkrai Arunotayanun and John W. Polak. Taste heterogeneity and market segmentation in freight shippers' mode choice behaviour. *Transportation Research Part E: Logistics and Transportation Review*, 47(2):138 – 148, 2011. ISSN 1366-5545. doi: http://dx.doi.org/10.1016/j.tre.2010.09.003. URL http://www.sciencedirect.com/science/article/pii/S1366554510000827.

Patrick Bajari, Jeremy T. Fox, and Stephen P. Ryan. Linear regression estimation of discrete choice models with nonparametric distributions of random coefficients. *The American Economic Review*, 97(2):pp. 459–463, 2007. ISSN 00028282. URL http://www.jstor.org/stable/30034495.

Kelvin Balcombe, Michael Burton, and Dan Rigby. Skew and attribute non-attendance within the bayesian mixed logit model. *Journal of Environmental Economics and Management*, 62(3):446 – 461, 2011. ISSN 0095-0696. doi: http://dx.doi.org/10.1016/j.jeem.2011.04.004. URL http://www.sciencedirect.com/science/article/pii/S0095069611000635.

Parisa Bastani and Melvyn Weeks. Nonparametric estimation of mixing distributions in discrete choice models: Heterogeneity and WTP. In *European Economic Association & Econometric Society Parallel Meetings*, Gothenburg, August 2013. URL http://www.eea-esem.com/eea-esem/2013/prog/viewpaper.asp?pid=2232.

Fabian Bastin, Cinzia Cirillo, and Philippe L. Toint. Estimating nonparametric random utility models with an application to the value of time in heterogeneous populations. *Transportation Science*, 44(4):537–549, 2010. doi: 10.1287/trsc.1100.0321. URL http://dx.doi.org/10.1287/trsc.1100.0321.

Angel Bujosa, Antoni Riera, and Robert L. Hicks. Combining discrete and continuous representations of preference heterogeneity: A latent class approach. *Environmental and Resource Economics*, 47(4):477–493, 2010. ISSN 0924-6460. doi: 10.1007/s10640-010-9389-y. URL http://dx.doi.org/10.1007/s10640-010-9389-y.

Danny Campbell and Edel Doherty. Combining discrete and continuous mixing distributions to identify niche markets for food. *European Review of Agricultural Economics*, 40 (2):287–312, 2013. doi: 10.1093/erae/jbs018. URL http://erae.oxfordjournals.org/content/40/2/287.abstract.

Danny Campbell, Edel Doherty, Stephen Hynes, and Tom M. van Rensburg. Combining discrete and continuous mixing approaches to accommodate heterogeneity in price sensitivities in environmental choice analysis. In *84th Annual Conference*, Edinburgh, Scotland, March 29-31 2010. Agricultural Economics Society. URL http://purl.umn.edu/91723.

Danny Campbell, Suzanne Elizabeth Vedel, Bo Jellesmark Thorsen, and Jette Bredahl Jacobsen. Heterogeneity in the WTP for recreational access: distributional aspects. *Journal of Environmental Planning and Management*, 57(8):1200–1219, 2014. doi: 10.1080/09640568.2013.793173. URL http://dx.doi.org/10.1080/09640568.2013.793173.

Ali Chalak, Jonathan Hecht, Scott Reid, and Mohamad G. Abiad. Willingness-to-pay for greenhouse gas reductions: A Bayesian investigation of distributional patterns. *Environmental Science & Policy*, 19–20(0):147–157, 2012. ISSN 1462-9011. doi: http://dx.doi.org/10.1016/j.envsci.2012.02.003. URL http://www.sciencedirect.com/science/article/pii/S1462901112000330.

Elisabetta Cherchi and Cristian Angelo Guevara. A Monte Carlo experiment to analyze the curse of dimensionality in estimating random coefficients models with a full variance-covariance matrix. *Transportation Research Part B: Methodological*, 46(2):321–332, 2012. ISSN 0191-2615. doi: http://dx.doi.org/10.1016/j.trb.2011.10.006. URL http://www.sciencedirect.com/science/article/pii/S0191261511001469. Emerging and Innovative Directions in Choice Modeling.

Cinzia Cirillo and Pratt Hetrakul. Continuous random coefficient logit models: A comparison of parametric and non-parametric methods to estimate individual preferences over cybernetic transportation systems. In *Transportation Research Board 89th Annual Meeting*, 2010. URL http://trid.trb.org/view/2010/C/909334.

Yves Croissant. *mlogit: multinomial logit model*, 2013. URL http://CRAN.R-project.org/package=mlogit. R package version 0.2-4.

Andrew Daly, Stephane Hess, and Kenneth Train. Assuring finite moments for willingness to pay in random coefficient models. *Transportation*, 39(1):19–31, 2012. ISSN 0049-4488. doi: 10.1007/s11116-011-9331-3. URL http://dx.doi.org/10.1007/s11116-011-9331-3.

Wayne Desarbo, Asim Ansari, Pradeep Chintagunta, Charles Himmelberg, Kamel Jedidi, Richard Johnson, Wagner Kamakura, Peter Lenk, Kannan Srinivasan, and Michel Wedel. Representing heterogeneity in consumer response models 1996 choice conference participants. *Marketing Letters*, 8(3):335–348, 1997. ISSN 0923-0645. doi: 10.1023/A:1007916714911. URL http://dx.doi.org/10.1023/A%3A1007916714911.

X. Dong and F. S. Koppelman. Comparison of continuous and discrete representations of unobserved heterogeneity. *J Market Anal*, 2(1):43–58, Mar 2014. ISSN 2050-3318. URL http://dx.doi.org/10.1057/jma.2014.5.

Xiaojing Dong and Frank S. Koppelman. Comparison of methods representing heterogeneity in logit models. In *10th International Conference on Travel Behaviour Research*, Lucerne, 2003. URL https://edit.ethz.ch/ivt/news/archive/20030810_IATBR/dong.pdf.

Mogens Fosgerau. *Handbook of choice modelling*, chapter Nonparametric approaches to describing heterogeneity, pages 257–267. Edward Elgar Publishing, 2014. doi: 10.4337/9781781003152.00018. URL http://www.elgaronline.com/view/9781781003145.00018.xml.

Mogens Fosgerau and Michel Bierlaire. A practical test for the choice of mixing distribution in discrete choice models. *Transportation Research Part B: Methodological*, 41(7):784 – 794, 2007. ISSN 0191-2615. doi: http://dx.doi.org/10.1016/j.trb.2007.01.002. URL http://www.sciencedirect.com/science/article/pii/S0191261507000136.

Mogens Fosgerau and Stephane Hess. A comparison of methods for representing random taste heterogeneity in discrete choice models. *European Transport*, 42:1–25, 2009. URL http://www.openstarts.units.it/dspace/bitstream/10077/6127/1/42_1_FosgerauHess.pdf.

William H. Greene and David A. Hensher. Revealing additional dimensions of preference heterogeneity in a latent class mixed multinomial logit model. *Applied Economics*, 45(14):1897–1902, 2013. doi: 10.1080/00036846.2011.650325. URL http://www.tandfonline.com/doi/abs/10.1080/00036846.2011.650325.

Scott D. Halpern, Kristin M. Madison, and Kevin G. Volpp. Patients as mercenaries?: The ethics of using financial incentives in the war on unhealthy behaviors. *Circulation: Cardiovascular Quality and Outcomes*, 2(5):514–516, 2009. doi: 10.1161/CIRCOUTCOMES.109.871855. URL http://circoutcomes.ahajournals.org/content/2/5/514.short.

David A. Hensher and William H. Greene. The mixed logit model: The state of practice. *Transportation*, 30(2):133–176, 2003. ISSN 0049-4488. doi: 10.1023/A:1022558715350. URL http://dx.doi.org/10.1023/A%3A1022558715350.

S. Hess, M. Bierlaire, and J. W. Polak. A systematic comparison of continuous and discrete mixture models. *European Trans*, 8(37):35–51, 2007. URL http://hdl.handle.net/10077/5957.

Stephane Hess. Conditional parameter estimates from mixed logit models: distributional assumptions and a free software tool. *Journal of Choice Modelling*, 3(2):134–152, 2010. URL http://www.sciencedirect.com/science/article/pii/S1755534513700393.

Stephane Hess. *Handbook of Choice Modelling*, chapter Latent class structures: taste heterogeneity and beyond, pages 311–330. Edward Elgar Publishing, 2014. doi: 10.4337/9781781003152.00021. URL http://www.elgaronline.com/view/9781781003145.00021.xml.

Stephane Hess and JohnM. Rose. Can scale and coefficient heterogeneity be separated in random coefficients models? *Transportation*, 39(6):1225–1239, 2012. ISSN 0049-4488. doi: 10.1007/s11116-012-9394-9. URL http://dx.doi.org/10.1007/s11116-012-9394-9.

Stephane Hess, Michel Bierlaire, and John W. Polak. Estimation of value of travel-time savings using mixed logit models. *Transportation Research Part A: Policy and Practice*, 39 (2Ű3):221 – 236, 2005. ISSN 0965-8564. doi: http://dx.doi.org/10.1016/j.tra.2004.09.007. URL http://www.sciencedirect.com/science/article/pii/S0965856404001028. Positive Utility of Travel Positive Utility of Travel.

Stephane Hess, Kay W. Axhausen, and John W. Polak. Distributional assumptions in mixed logit. In *Transportation Research Board 85th Annual Meeting*, 2006. URL http://trid.trb.org/view.aspx?id=777238.

Stephen Hynes, Nick Hanley, and Riccardo Scarpa. Effects on welfare measures of alternative means of accounting for preference heterogeneity in recreational demand models. *American Journal of Agricultural Economics*, 90(4):1011–1027, 2008. doi: 10.1111/j.1467-8276.2008.01148.x. URL http://ajae.oxfordjournals.org/content/90/4/1011.abstract.

Michael Keane and Nada Wasi. Comparing alternative models of heterogeneity in consumer choice behavior. *Journal of Applied Econometrics*, 28(6):1018–1045, 2013. ISSN 1099-1255. doi: 10.1002/jae.2304. URL http://dx.doi.org/10.1002/jae.2304.

Karsten Lunze and Michael K. Paasche-Orlow. Financial incentives for healthy behavior: Ethical safeguards for behavioral economics. *American Journal of Preventive Medicine*, 44 (6):659 – 665, 2013. ISSN 0749-3797. doi: http://dx.doi.org/10.1016/j.amepre.2013.01.035. URL http://www.sciencedirect.com/science/article/pii/S0749379713001864.

Stefan Mabit, Stephane Hess, and Sebastiàn Cassaude. Correlation and willingness-to-pay indicators in transport demand modelling. In *Transportation Research Board Annual Meeting*, 2008. URL http://trid.trb.org/view.aspx?id=847647.

Edward Morey, Jennifer Thacher, and William Breffle. Using angler characteristics and attitudinal data to identify environmental preference classes: A latent-class model. *Environmental and Resource Economics*, 34(1):91–115, 2006. ISSN 0924-6460. doi: 10.1007/s10640-005-3794-7. URL http://dx.doi.org/10.1007/s10640-005-3794-7.

Daniele Pacifico. On the role of unobserved preference heterogeneity in discrete choice models of labour supply. *Empirical Economics*, 45(2):929–963, 2013. ISSN 0377-7332. doi: 10.1007/s00181-012-0637-6. URL http://dx.doi.org/10.1007/s00181-012-0637-6.

Dan Rigby, Kelvin Balcombe, and Michael Burton. Mixed logit model performance and distributional assumptions: Preferences and gm foods. *Environmental and Resource Economics*, 42(3):279–295, 2009. ISSN 0924-6460. doi: 10.1007/s10640-008-9227-7. URL http://dx.doi.org/10.1007/s10640-008-9227-7.

Alvin E. Roth. Repugnance as a constraint on markets. *Journal of Economic Perspectives*, 21 (3):37–58, 2007. doi: 10.1257/jep.21.3.37. URL http://www.aeaweb.org/articles.php?doi=10.1257/jep.21.3.37.

Jan Rouwendal, Arianne de Blaeij, Piet Rietveld, and Erik Verhoef. The information content of a stated choice experiment: A new method and its application to the value of a statistical life. *Transportation Research Part B: Methodological*, 44(1):136 – 151, 2010. ISSN 0191-2615. doi: http://dx.doi.org/10.1016/j.trb.2009.04.006. URL http://www.sciencedirect.com/science/article/pii/S0191261509000599.

Harald Schmidt. Wellness incentives, equity, and the 5 groups problem. *Am J Public Health*, 102(1):49–54, November 2011. ISSN 0090-0036. doi: 10.2105/AJPH.2011.300348. URL http://dx.doi.org/10.2105/AJPH.2011.300348.

Kenneth Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2009.

Kenneth Train and Garrett Sonnier. *Mixed Logit with Bounded Distributions of Correlated Partworths*, volume 6 of *The Economics of Non-Market Goods and Resources*, pages 117–134. Springer Netherlands, 2005. ISBN 978-1-4020-3683-5. doi: 10.1007/1-4020-3684-1_7. URL http://dx.doi.org/10.1007/1-4020-3684-1_7.

Kenneth E. Train. EM algorithms for nonparametric estimation of mixing distributions. *Journal of Choice Modelling*, 1(1):40–69, 2008. ISSN 1755-5345. doi: http://dx.doi.org/10.1016/S1755-5345(13)70022-8. URL http://www.sciencedirect.com/science/article/pii/S1755534513700228.

Michel Wedel and Wagner. Kamakura. *Market segmentation: conceptual and methodological foundations*. Kluwer Academic Publishers, second edition edition, 2000.

Michel Wedel, Wagner Kamakura, Neeraj Arora, Albert Bemmaor, Jeongwen Chiang, Terry Elrod, Rich Johnson, Peter Lenk, Scott Neslin, and CarstenStig Poulsen. Discrete and continuous representations of unobserved heterogeneity in choice modeling. *Marketing Letters*, 10(3):219–232, 1999. ISSN 0923-0645. doi: 10.1023/A:1008054316179. URL http://dx.doi.org/10.1023/A%3A1008054316179.