The Journal of Choice Modelling 19 (2016) 40–53

Contents lists available at ScienceDirect

The Journal of Choice Modelling

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

Mixed logit with a flexible mixing distribution ☆

Kenneth Train

Department of Economics, University of California, Berkeley, United States

ARTICLE INFO

Article history:
Received 28 February 2016
Received in revised form
19 July 2016
Accepted 21 July 2016
Available online 3 August 2016

Keywords:
Mixed logit
Mixing distribution
Nonparametric

ABSTRACT

This paper presents a flexible procedure for representing the distribution of random parameters in mixed logit models. A logit formula is specified for the mixing distribution, in addition to its use for the choice probabilities. The properties of logit assure positivity and provide the normalizing constant for the mixing distribution. Any mixing distribution can be approximated to any degree of accuracy by this specification. The researcher defines variables to describe the shape of the mixing distribution, using flexible forms such as polynomials, splines, and step functions. The gradient of the log-likelihood is easy to calculate, which facilitates estimation. The procedure is illustrated with data on consumers' choice among video streaming services.

1. Introduction

A mixed logit model (e.g., Revelt and Train, 1998) contains two parts: a logit specification of a person's probability of choosing a given alternative, which depends on parameters that enter the person's utility function; and a specification of the distribution – often called the mixing distribution – of these utility-parameters over people. McFadden and Train (2000) have shown that a mixed logit model can, under benign conditions, approximate any choice model to any degree of accuracy. However, this theoretical generality is constrained in practice by the difficulty of specifying and estimating parameter distributions that are sufficiently flexible and yet feasible from a computational perspective. The vast majority of studies have used normal and lognormal distributions, with a few using Johnson's S_b, gamma, and triangular distributions. However, these distributions are limiting, and most researchers will probably agree that: whatever parametric distribution the researcher specifies, he/she quickly becomes dissatisfied with its properties.

The current paper introduces a new way of specifying the distribution of random parameters that is relatively simple numerically and yet allows a high degree of flexibility. For a finite parameter space, the probability of each parameter value is given by a logit function with terms that are defined by the researcher to describe the shape of the distribution. The researcher can specify polynomials, splines, steps, and other functions that have been developed for general approximation. The exponential in the logit numerator assures that the probability is positive, and the summation in the denominator assures that the probabilities sum to one. Sampling from the parameter space is facilitated by the fact that the logit probability takes the same form on subsets as the full set. When used in a mixed logit, the model contains two logits: one for the decision-maker's choice among alternatives, and another for the "selection" of parameters for the decision-maker. The procedure is easy to program and fast computationally.

Procedures for flexible mixing distributions have been previously proposed by Bajari et al. (2007), Fosgerau and Bierlaire

☆ Matlab codes to implement the procedures in this paper are available at http://eml.berkeley.edu/~train/software.html.
E-mail addresses: kenneth.train@nera.com, train@econ.berkeley.edu

http://dx.doi.org/10.1016/j.jocm.2016.07.004

(2007), Train (2008), Fox et al. (2011), Burda et al. (2008) and Fosgerau and Mabit (2013). The current method is an approximate generalization of the first four of these papers. Burda et al. (2008) obtain flexibility through a convolution of a normal kernel with a skewing function. Fosgerau and Mabit (2013) suggest an approach that approximates a different function than the density of the utility parameters. In particular, random terms from a standardized distribution (such as uniform or standard normal) are transformed as they enter utility, and this transformation is approximated by, e.g., a polynomial.

This movement toward greater flexibility, which the current paper augments, shifts the emphasis for future research from overcoming distributional constraints to developing richer datasets that can more clearly distinguish among the variety of shapes that the parameter distribution might take.

Sections 2 and 3 describe the model form and its estimation. Section 4 gives examples of how to specify variables to describe the mixing distribution. Section 5 extends the procedure in two fairly obvious ways. Section 6 provides an application, and Section 7 concludes.

2. A Logit-mixed logit (LML) model

Consider a situation in which each decision-maker makes only one choice, since the double-use of logits is most apparent in this situation; generalization to multiple choices by each decision-maker is described in the next section. Let the utility that person n obtains from alternative j in choice set J be denoted in the usual way as $U_{nj} = \beta_n' x_{nj} + \varepsilon_{nj}$ where x_{nj} is a vector of observed attributes, β_n is a corresponding vector of utility coefficients that vary randomly over people, and ε_{nj} is a random term that represents the unobserved component of utility. The unobserved term ε_{nj} is assumed to be distributed iid extreme value. Under this assumption, the probability that person n chooses alternative i, conditional on β_n, is the logit formula:

$$Q_{ni}(\beta_n) = \frac{e^{\beta_n' x_{ni}}}{\sum_{j \in J} e^{\beta_n' x_{nj}}} \tag{1}$$

The researcher does not observe the utility coefficients of each person and knows that the coefficients vary over people. The cumulative distribution function of β_n in the population is $F(\beta)$ which is called the mixing distribution. Let F be discrete with finite support set S. This specification is not restrictive since a continuous distribution can be approximated to any degree of accuracy by a discrete distribution with a sufficiently large and dense S. The probability mass at any $\beta_r \in S$ is expressed as a logit formula:

$$\text{Prob}(\beta_n = \beta_r) \equiv W(\beta_r|\alpha) = \frac{e^{\alpha' z(\beta_r)}}{\sum_{s \in S} e^{\alpha' z(\beta_s)}} \tag{2}$$

where $z(\beta_r)$ is a vector-valued function of β_r and α is a corresponding vector of coefficients. The z variables are chosen to capture the shape of the probability mass function; their specification is discussed in Section 4 below.

The unconditional choice probability is then:

$$\text{Prob}(n \text{ chooses } i) = \sum_{r \in S} W(\beta_r|\alpha) \cdot Q_{ni}(\beta_r) = \sum_{r \in S} \left(\frac{e^{\alpha' z(\beta_r)}}{\sum_{s \in S} e^{\alpha' z(\beta_s)}} \right) \cdot \left(\frac{e^{\beta_r' x_{ni}}}{\sum_{j \in J} e^{\beta_r' x_{nj}}} \right) \tag{3}$$

The model contains a logit formula for the probability that the decision-maker chooses alternative i *and* a logit formula for the probability that the decision-makers has utility coefficients β_r. The researcher's task is to specify the x variables that describe the probability of each alternative and the z variables that describe the probability of each β_r.

The advantage of using a logit model for the mixing distribution is that it allows for easy and flexible specification of relative probabilities. The researcher specifies z variables that describe the shape of the distribution, without needing to be concerned about assuring positivity or summation to one: the exponential in the logit numerator guarantees that the probability at each point is positive, and the sum in the denominator guarantees that probabilities sum to one over points.

The specification is entirely general in the sense that any choice model with any mixing distribution can be approximated to any degree of accuracy by a model of the form of Eq. (3). McFadden's (1975) "mother logit" theorem shows that any model that describes the choice among alternatives can be represented by a logit formula of the form in Eq. (1). An analogous derivation applies for representing the mixing distribution as a logit formula.

Result: For any mixing distribution, there exists a sequence of probability distributions in the form of Eq. (2) that converges weakly (i.e., in distribution) to that mixing distribution. Stated more intuitively, any mixing distribution can be approximated to any degree of accuracy by a logit model of the form given in Eq. (2). Proof: For any distribution F*, there exists a sequence of discrete distributions, labeled F_m, $m = 1, 2, \ldots$, that converges weakly to F^* (e.g. Chamberlain, 1987).[1] Let the probability mass function associated with distribution F_m be denoted $k_m f_m(\beta)$ at each $\beta \in S_m$, where k_m is the normalizing constant, $f_m(\beta)$ is the kernel, and S_m is the support set. Define $g_m(\beta) = \ln(f_m(\beta))$, which exists for each $\beta \in S_n$; that is, function

[1] Dan McFadden has suggested (personal communication) a simple demonstration: Take m draws from F^* and define $F_m(\beta) = \sum_{i=1}^{m} I(\beta_i \le \beta)/m$. By the strong law of large numbers, this sequence converges weakly to F^*.

g_m is the log of the kernel of the probability mass function. Eq. (2) with $\alpha = 1$ and $z(\beta_r) = g_m(\beta_r)$ returns $k_m f_m(\beta_r)$:

$$\frac{e^{\alpha z(\beta_r)}}{\sum_{s\in S_m} e^{\alpha z(\beta_s)}} = \frac{f_m(\beta_r)}{\sum_{s\in S_m} f_m(\beta_r)} = \frac{k_m f_m(\beta_r)}{\sum_{s\in S} k_m f_m(\beta_r)} = k_m f_m(\beta_r). \tag{4}$$

If $\ln f_m(\beta)$ cannot be calculated exactly, then it can be approximated by a linear-in-parameters form $\alpha' z(\beta)$ with appropriate selection of z variables, such as step functions, splines or polynomials; see, e.g., Theorems 4.7.1–4.7.3 respectively in Sohrab (2003). This derivation points out that the goal of the researcher is to specify z variables that capture the shape of the log of the density of the random coefficients.[2]

3. Estimation

We now allow for multiple choices by each decision-maker. Index choice situations by t and denote the attributes of alternative j for person n in choice situation t as x_{njt}. Suppose the person chose alternative i_t in choice situation t, and consider the person's sequence of chosen alternatives $i_1 \ldots i_T$ in T choice situations. The probability that person n made this sequence of choices, conditional on β_n, is:

$$L_n(\beta_n) = \prod_{t=1,\ldots,T} Q_{ni_tt}(\beta_n) \tag{5}$$

where $Q_{ni_tt}(\beta_n)$ is given by Eq. (1) appropriately modified to represent choice situation t. The (unconditional) probability of the sequence of choices is then:

$$P_n = \sum_{r\in S} L_n(\beta_r) W(\beta_r|\alpha) \tag{6}$$

The parameters to be estimated are the vector α that describe the mixing distribution. The log-likelihood function for α given a sample indexed by $n = 1, \ldots, N$ is

$$\mathrm{LL} = \sum_{n=1,\ldots,N} \ln\left(\sum_{r\in S} L_n(\beta_r) W(\beta_r|\alpha) \right) \tag{7}$$

S might be so large that calculating LL is infeasible. If so, then the log-likelihood function can be simulated in the usual way by using random draws of β_r for each person (McFadden, 1989; Hajivassiliou and Ruud, 1994; Train, 2009). Let $S_n \subset S$ be a subset of R randomly selected values of β, with all elements of S having the same probability of being selected.[3] The simulated log-likelihood function is:

$$\mathrm{SLL} = \sum_n \ln\left(\sum_{r\in S_n} L_n(\beta_r) w_n(\beta_r|\alpha) \right) \tag{8}$$

where w_n is the logit formula based on subset S_n:

$$w_n(\beta_r|\alpha) = \frac{e^{\alpha' z(\beta_r)}}{\sum_{s\in S_n} e^{\alpha' z(\beta_s)}} \tag{9}$$

The estimator is the value of α that maximizes SLL. This simulation-based estimator is consistent, asymptotically normal, and asymptotically equivalent to the non-simulated maximum likelihood estimator if (i) R rises faster than $\sqrt{N}$ and (ii) the non-simulated maximum likelihood estimator is itself consistent and asymptotically normal (Gourieroux and Monfort, 1993; Lee, 1995; Train, 2009).[4,5]

[2] As mentioned above, Fosgerau and Mabit (2013) proposed an approach for flexible mixing distributions that is based on an alternative representation of the mixed logit formula. The difference can now be explained. For random u with a standardized density $g(u)$, the utility coefficients are represented by transformation $\beta = T(u|\theta)$ which depends on parameters θ. The mixed logit formula becomes $\int L_{ni}(T(u|\theta))g(u)du$. The researcher selects g, such as uniform, and a flexible form for T, such as a polynomial, and estimates the parameters θ. The procedure is easy to code and fast computationally. The density of the utility parameters, $f(\beta)$, is implied by the estimated transformation and the specified g. This density is not easy to derive, but it can be readily simulated for any estimated transformation and specific g. However, the relation of T and g to f is complex, and it is often not clear whether, or how, particular forms of T and g restrict the possible shapes of f. Clarifying this relation is a worthwhile goal for future research.

[3] The generalization to unequal selection probabilities is discussed in Section 5.2 below.

[4] The use of a subset of β's in estimation is similar to McFadden's (1978) suggestion to estimate standard (i.e., fixed-coefficient) logit models on a subset of alternatives. However, there is a difference. The chosen alternative is necessarily included in the subset of alternatives, which necessitates a correction term in estimation or, as usually applied, a sampling procedure that satisfies the "uniform conditioning property." In contrast, the decision-maker's actual β (which might be called the "chosen" β) is not observed and need not be included in the subset S_n. The conditions for consistency are therefore the same as those for any maximum simulated likelihood estimator.

[5] Note that $w_n(\beta_r)$ is an overestimate of $W(\beta_r)$ by the ratio of denominators. However, for any $R > 0$, the simulated mean of a statistic $t(\beta)$ using w_n is unbiased for its mean based on W: $E\left[\sum_{r\in S_n} t(\beta_r) w_n(\beta_r)\right] = \sum_{r\in S} t(\beta_r) W(\beta_n)$. Other statistics, such as the mode, are more difficult to simulate.

An important advantage of using a logit formula for the mixing distribution is that it gives easy-to-calculate gradients for maximization of the log-likelihood:

$$\frac{\partial SLL}{\partial \alpha} = \sum_n \left[\sum_{r \in S_n} (h_n(\beta_r|\alpha) - w_n(\beta_r|\alpha)) z(\beta_r) \right] \tag{10}$$

where

$$h_n(\beta_r|\alpha) = \frac{L_n(\beta_r) w_n(\beta_r|\alpha)}{\sum_{s \in S_n} L_n(\beta_s) w_n(\beta_s|\alpha)} \tag{11}$$

is the probability mass at β_r conditional on person n's sequence of choices. The first term in the gradient is the difference between the conditional and unconditional probability of β_r. The estimated parameters are those at which this difference becomes uncorrelated with the z variables; stated alternatively, if the z variables are correlated with this difference at any given value of the parameters, then the model can be improved by changing the parameters.[6]

For computation, note the following. First, and very importantly, $L_n(\beta_r)$, the probability of the person's choices conditional on β_r, does not depend on the parameters α that are being estimated. This term therefore does not change during the optimization process for estimation. Given a set S_n of points for a person, this probability is calculated once at each point in the set, and does not need to be recalculated at each iteration of the optimization procedure. The optimization procedure changes the weights $w_n(\beta_r|\alpha)$ associated with each value of β_r, but not the probability $L_n(\beta_r)$ of the person's choices at that β_r. This feature reduces computation time considerably. Second, the parameters α are estimated by a standard logit model with an "alternative" for each β_r and with the "dependent variable" for each alternative being the conditional weight associated with that value of β_r. The dependent variable (i.e., conditional weights) changes with each iteration, unlike a standard logit, but otherwise, the speed of standard logit estimation applies.

The covariance matrix of the estimator can be calculated as the inverse of the Hessian or the BHHH matrix, or by using both in the sandwich estimator, with the BHHH matrix easily calculated as the sample covariance of the bracketed term in Eq. (10).[7] Usually, however, the researcher will be interested in statistics that depend on α, rather than being concerned about α itself. The delta method can be used for these statistics, or the sampling distribution can be simulated by the bootstrap. The latter is conceptually straightforward and, importantly, avoids the need to derive and code-up derivative formulas.

4. Variables for the mixing distribution

The critical issue is: how to specify the z variables that describe the mixing distribution? By representing the mixing distribution as a logit function, which assures positivity and summation to one, the researcher can utilize the numerous procedures that have been developed for approximation of standard functions. The possibilities are most readily described through examples.

4.1. Approximate normal and lognormal

It is doubtful that a researcher would want to use this type of mixed logit model for only normals and lognormals, since these distributions can be readily handled with the usual mixed logit software. However, it is instructive to see how normals and lognormals can be accommodated in the current setup. Also, there can be an advantage to using the logit representation of a normal or lognormal, since it eliminates (though specification of S) the long tails of the normal and lognormal, which can be unrealistic in real-world choice situations.

The normal density for β with mean b and variance V contains a linear combination of variables in its exponential, with the coefficients depending on b and V and the variables depending on β:

$$f(\beta) = m(V)\exp(-(\beta - b)'V^{-1}(\beta - b)/2) = m(V)\exp(-(\beta' V^{-1}\beta)/2 + b'V^{-1}\beta - (b'V^{-1}b)/2) \tag{12}$$

where $m(V)$ is the normalizing constant. The exponentiated term is linear in the elements of β and the unique elements of $\beta\beta'$, which means that this density can be represented exactly by the logit of Eq. (3). If the elements of β are independent (i.e., V diagonal), then the z variables are each element of β and each element squared. For correlation among elements of β, cross-products of the elements are included as additional z variables.[8] The logit denominator provides the normalizing

[6] Note that the average difference $\sum_{r \in S_n}(h_n(\beta_r|\alpha) - w_n(\beta_r|\alpha))$ is necessarily zero for each person, since each probability, $h_n(\beta_r|\alpha)$ and $w_n(\beta_r|\alpha)$, sums to one over $\beta_r \in S_n$ such that their difference sums to zero.

[7] Most optimization codes calculate a numerical Hessian at convergence, which avoids the need to derive and code-up the analytic Hessian.

[8] The number of z variables equals the number of unique parameters in b and V, and the coefficients α are functions of b and V. Estimates of b and V can be calculated from the estimate of α. However, as stated above, it is doubtful that the method would be used to represent just a normal distribution; with higher-order polynomials or in combination with other functional forms, translation of the estimated coefficients of the second-order terms to b and V (i.e., to the normal component of the more general density) is not particularly useful or meaningful.

constant. And since the term $(b'V^{-1}b)/2$ in the normal density does not vary over β's, it drops out of the logit formula. Log-normal distributions are represented analogously with $\ln(\beta)$ replacing β.

4.2. Higher-order polynomials

The example above consists of specifying z to be a second-order polynomial in β. The polynomial can be extended to higher order to gain greater flexibility. Orthogonal polynomials provide the added advantage of reducing collinearity among the terms. Legendre polynomials are commonly used for numerical approximation of functions in general and are particularly useful in the current setup. Consider one-dimensional β. The support of Legendre polynomials is $[-1, 1]$, and so the transformation $\tilde{\beta} = -1 + 2(\beta - a)/(b - a)$ is used, where a and b are the lower and upper bounds of β in S. Let $\text{LGP}(k, q)$ be the k-th order Legendre polynomial of degree k for q. The z variables are $z_k(\beta) = \text{LGP}(k, \tilde{\beta})$ for $k = 1, \ldots, K$ where K is the highest order specified by the researcher. The logit numerator, $\exp(\alpha' z(\beta))$, based on these variables can represent a very wide variety of shapes depending on the value of α, and the flexibility rises with more terms. Dependence among the elements of multi-dimensional β is captured though cross-products of the terms of each element's polynomial. All cross-products need not be included; e.g., the cross-product of only the first-order terms provides correlation among the elements of β without greatly increasing the number of distributional parameters. Chebyshev, Bernstein, or other polynomials can be used instead. Most coding packages, including Matlab, contain commands to calculate the various polynomials.

This approach is very similar to that described by Fosgerau and Bierlaire (2007). The difference is that the current approach inserts the polynomials into a logit function, while Fosgerau and Bierlaire define the probability at a point to be the squared sum of weighted polynomials divided by a normalizing constant that depends on the parameters. The logit function has a more convenient gradient and allows the polynomials to be combined with other z variables.

4.3. Step functions

Let the set S be partitioned into (possibly overlapping) subsets labeled H_g for $g = 1, \ldots G$. Let $W(\beta)$ be the same for all points within each subset, but different over subsets. The logit formula for the probability masses is:

$$W(\beta_r) = \frac{e^{\sum_g \alpha_g I(\beta_r \in H_g)}}{\sum_{s \in S} e^{\sum_g \alpha_g I(\beta_s \in H_g)}} \tag{13}$$

The z variables are the G indicators of which subsets contain β_r. If the subsets do not overlap, then one of the coefficients is normalized to zero. With overlapping subsets, one coefficient is normalized to zero for each possible way of covering the set S.

The use of overlapping sets H_g can reduce the number of parameters with little, if any, loss in interpretation. An example is to use subsets for each single coefficient (to provide a step function for each marginal) and then use more coarsely defined subsets for each pair of coefficients to provide a coarser multi-dimensional step function that captures correlation. Fig. 1 depicts an example in two dimensions. There are six intervals in each dimension, giving 36 partitioned squares. Estimating a parameter for each square would result in 35 parameters (with the summation to 1 providing the 36th share), which can be considered the saturated specification. Consider instead overlapping partitions, with six rectangles differentiating one of the dimensions, as in Fig. 1a; six rectangles differentiating the other dimension, as in Fig. 1b; and four large squares differentiating over both dimensionals, as in Fig. 1c. This specification contains 5+5+3=13 parameters and yet provides as flexible a distribution in each dimension as the saturated specification as well as correlation over the dimensions (through the parameters for the four large squares). The extra 22 parameters in the saturated specification provide extra multi-dimensional flexibility, but might not warrant the large number of extra parameters. The reduction in number of parameters from using overlapping subsets in this way is greater with more dimensions and more steps in each dimension. Essentially. the researcher's selective use of subsets can allow for finely defined grids without a proliferation of parameters and with meaningful, interpretable relations among the probabilities at all points in S.

Bajari et al. (2007), Train (2008) and Fox et al. (2011) proposed estimation of the probability mass at each point in S, with S specified as a grid. Their model is a type of latent class model, with each point representing a class, but with the location of the points being specified rather estimated as in a standard latent class model. The LML model generalizes their approach by allowing each H_g to contain more than one point and to overlap. A numerical limitation of their approach is that the number of parameters is the number of grid points, which becomes unwieldy or infeasible, even for moderately coarse grids. For example, a grid with only six points in each of seven dimensions contains 279,936 points, each of which, in their approach, represents a parameter. As well as presenting numerical difficulties for optimization and calculation of standard errors, it is doubtful that a model with so little structure is really needed. In general, the LML model can be seen as a latent class model where: (1) the points are specified instead of estimated and (2) the share at each point need not be treated as an individual parameter but rather can be related to the shares at other points through the z variables. Both of these features allow far more points to be included in LML models than traditional latent class models.

4.4. Splines

Linear splines can be written in the form $\alpha' z(\beta)$ as needed for our specification. Consider, as an example, the spline $f(\beta)$ for one-dimensional β with starting point $\bar{\beta}_1$, knots at $\bar{\beta}_2$ and $\bar{\beta}_3$, and endpoint $\bar{\beta}_4$. The distributional parameters are the height of

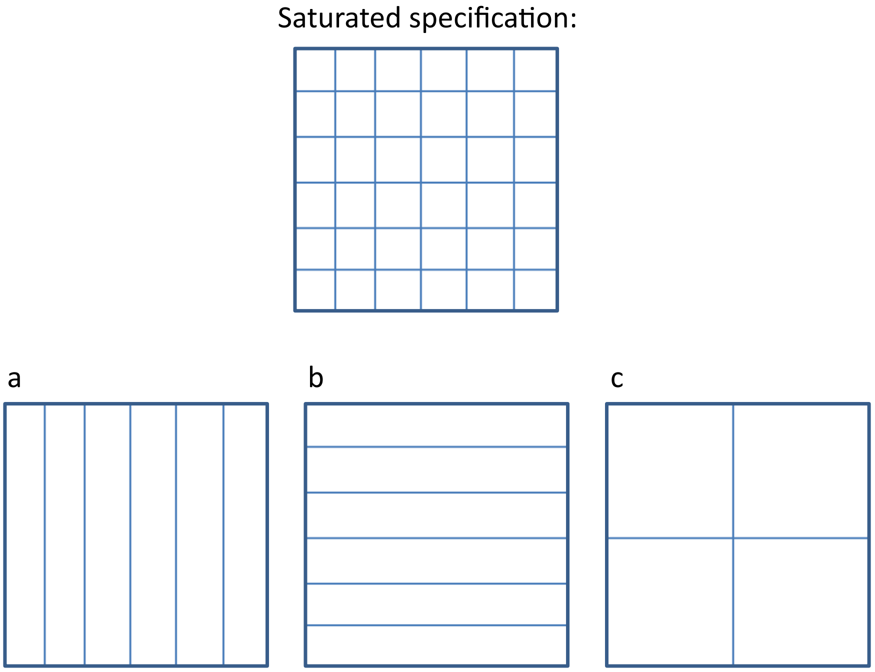

Fig. 1. Overlapping step functions.

the spline at these points, which are labeled $\alpha_1, \dots, \alpha_4$. The formula for the spline is

$$f(\beta) = \left\{ \begin{array}{ll} \alpha_1 + \dfrac{\alpha_2 - \alpha_1}{\bar{\beta}_2 - \bar{\beta}_1}(\beta - \bar{\beta}_1) & \text{if } \beta \le \bar{\beta}_2 \\ \alpha_2 + \dfrac{\alpha_3 - \alpha_2}{\bar{\beta}_3 - \bar{\beta}_2}(\beta - \bar{\beta}_2) & \text{if } \bar{\beta}_2 < \beta \le \bar{\beta}_3 \\ \alpha_3 + \dfrac{\alpha_4 - \alpha_3}{\bar{\beta}_4 - \bar{\beta}_3}(\beta - \bar{\beta}_3) & \text{if } \bar{\beta}_3 < \beta \end{array} \right\} = \alpha' z(\beta) \tag{14}$$

where z contains four elements:

$$z_1(\beta) = \left(1 - \frac{\beta - \bar{\beta}_1}{\bar{\beta}_2 - \bar{\beta}_1}\right) I(\beta \le \bar{\beta}_2) z_2(\beta) = \frac{\beta - \bar{\beta}_1}{\bar{\beta}_2 - \bar{\beta}_1} I(\beta \le \bar{\beta}_2) + \left(1 - \frac{\beta - \bar{\beta}_2}{\bar{\beta}_3 - \bar{\beta}_2}\right) I(\bar{\beta}_2 < \beta \le \bar{\beta}_3) z_3(\beta) = \frac{\beta - \bar{\beta}_2}{\bar{\beta}_3 - \bar{\beta}_2} I(\bar{\beta}_2 < \beta \le \bar{\beta}_3)$$
$$+ \left(1 - \frac{\beta - \bar{\beta}_3}{\bar{\beta}_4 - \bar{\beta}_3}\right) I(\bar{\beta}_3 < \beta) z_4(\beta) = \frac{\beta - \bar{\beta}_3}{\bar{\beta}_4 - \bar{\beta}_3} I(\bar{\beta}_3 < \beta)$$

and $I(\cdot) = 1$ if the statement in parentheses is true and $=0$ otherwise.

The discrete probability mass at any point in S is defined by the logit Eq. (3) evaluated at the spline function. The exponentiation changes the shape of the spline, but its flexibility remains. Since the overall height is irrelevant, one of the parameters (heights) is normalized, such that, in the above example, there are three free parameters to be estimated. Multidimensional splines are defined analogously.

4.5. Combinations

The z variables described above can be combined to take advantage of the properties of each. One particularly useful combination is to specify a step-function or spline for each single coefficient, and then capture correlation over coefficients with a second-order polynomial. This procedure utilizes fewer parameters for correlations than creating multi-dimensional step-functions or splines. The specification can be viewed as consisting of a joint normal distribution (created by the second-order polynomial) with the marginals re-shaped based on a step function or spline.

4.6. Method of sieves

The use of polynomials, step functions and linear splines within the LML model can be viewed as a type of nonparametric (or semi-nonparametric) estimation of the mixing distribution, by the method of sieves (Chen, 2007). The number of α

parameters (i.e., the order of the polynomial and/or the number of steps/nodes) rises with sample size, providing ever-more flexibility in fitting the true distribution. The requirements for consistency and asymptotic normality, and the rates of convergence, are important topics that are beyond the scope of this paper. These issues depend greatly on the statistics that are being evaluated. For example, requirements for consistency are less stringent and the rates of convergence faster when evaluating summary statistics of the mixing distribution (where the rising number of α parameters are treated as nuisance parameters) than when assessing the maximum error over the density's support.

5. Extensions

5.1. WTP space

Train and Weeks (2005) describe how the utility parameters in a mixed logit can be represented as the person's willingness-to-pay rather than the coefficients entering the person's utility. They call the former representation “models in WTP space” and the latter “models in preference space.” The procedure in the current paper for specifying the distribution of utility parameters is described above for models in preference space. However, the procedure is equally applicable to models in WTP space. The only change is that utility entering the logit probability of choice is changed from $\beta_n' x_{nit}$ to $-\sigma_n(r_{nit} + wtp_n' x_{nit})$ where r_{nit} is price, wtp_n is a vector of willingness to pay for each non-price attribute, and σ_n is a random scalar. The vector β is re-defined as $<\sigma, wtp>$, after which β enters the same in all subsequent equations.

5.2. Unequal probability sampling

The simulated log-likelihood function is defined above for a sample of β's selected with equal probability from S. Non-equal probabilities can be used in the random selection of β's for each person's simulated probability, using importance sampling. In particular, let $q(\beta)$ be a probability mass function over $\beta_r \in S$ that is determined by the researcher. The LL function can be re-written as

$$\text{LL} = \sum_{n=1,\ldots N} \ln\left(\sum_{r \in S} (L_n(\beta_r)/q(\beta_r))W(\beta_r|\alpha)q(\beta_r) \right) \tag{15}$$

which is the same as (7) but with the sampling probability made explicit. Simulation points in S_n are selected randomly from S with probability $q(\beta_r)$. The simulated log-likelihood then becomes:

$$\text{SLL} = \sum_n \ln\left(\sum_{r \in S_n} (L_n(\beta_r)/q(\beta_r))w_n(\beta_r|\alpha) \right). \tag{16}$$

The term $L_n(\beta_r)/q(\beta_r)$ does not depend on α and is calculated only once, rather than in each iteration of maximization. The use of q for sampling of the β_r's allows the researcher to expand the size of S (i.e., use wider ranges in each dimension) while, for example, giving greater importance to those points near the edge of the ranges where polynomials can take extreme and inaccurate shapes.

6. Application

To illustrate the procedure, I utilize data from a conjoint experiment designed and described by Glasgow and Butler (forthcoming) for consumers' choice among video streaming services. Their experiments included the monthly price of the service and the non-price attributes given in Table 1.

Each choice experiment included four alternative video streaming services with specified price and attributes plus a fifth alternative of not subscribing to any video streaming service. Each respondent was presented with 11 choice situations. Butler and Glasgow obtained choices from 300 respondents and implemented an estimation procedure that accounts for protestors (mainly consumers who never chose a service that shared data). For the present use of their data, I do not include the 40 respondents whom they identified as protestors, so that the sample consists of 260 respondents.

6.1. Normal distribution

The first column of Table 2 gives a model in WTP-space where the WTP's are specified to be jointly normal and the price/scale coefficient σ is lognormal.[9] I first estimated the model by the hierarchical Bayes (HB) procedure in Train (2009) modified for WTP space as in Train and Weeks (2005, Ch. 12) and Scarpa et al. (2008) using my own Matlab codes. I then

[9] These estimates are the same as those in Table 10 of Ben-Akiva et al. (2015). The estimated correlation coefficients, not shown here, are reported by Ben-Akiva et al.

Table 1
Non-price attributes.

Attribute	Levels
Commercials between content	Yes (“commercials”) No (baseline category)
Speed of content availability	TV episodes next day, movies in 3 months (“fast content”) TV episodes in 3 months, movies in 6 months (baseline)
Catalog	5000 movies and 2500 TV episodes (baseline) 10,000 movies and 5000 TV episodes (“more content”) 2000 movies and 13,000 TV episodes (“more TV, fewer movies”)
Data-sharing policies	Information is collected but not shared (baseline) Usage info is shared with third parties (“share usage”) Usage and personal info shared (“share usage and personal”)

used the HB estimates as the starting values for maximum simulated likelihood in Stata using the command “mixlogitwtp”. Estimation in Stata took a little over 4 h on my PC. The simulated log-likelihood rose from −4017.10 with the HB estimates to −3903.47 with the maximum likelihood estimates.

The estimates indicate that people are willing to pay $1.56 per month on average to avoid commercials. Fast availability is valued highly, with an average WTP of $3.94 per month in order to see TV shows and movies soon after their original showing. On average, people do not want to have more TV shows with fewer movies. But consumers are WTP $2.96 on average to obtain twice as much content of both kinds. Consumers are estimated to have a WTP of 62 cents per month to avoid having their usage data shared in aggregate form; however, the hypothesis of zero average WTP cannot be rejected. Consumers are much more concerned about their personal information being shared along with their usage information: the average WTP to avoid such sharing is estimated to be $2.70 per month.

6.2. Polynomials

Consider now the procedure described above for representing the distribution with polynomials. Using the estimated means and standard deviations in the first column of Table 2, the parameter space S was specified to extend two standard deviations above and below the mean of each WTP and from 0 to 2 for the price/scale coefficient.[10] Each dimension was divided into 1000 evenly-spaced points, such that S contained a total of 10^{24} multi-dimensional grid points (which is a trillion trillions.) A random sample of 2000 points was drawn independently for each person in the sample. The z variables were specified as a sixth-order polynomial on each utility parameter and, to capture correlations, a second-order polynomial on each pair of WTPs. Notice that even though the parameter space contains many points (approximating a continuous space), only 69 parameters are being estimated.[11]

The parameters were estimated by maximum simulated likelihood, with standard errors calculated by bootstrapping. Estimation was very fast: I wrote the code in Matlab using its parallel processing toolbox and ran on a computer with a Tesla K20 gpu, which operates well with Matlab's gpu processing. Estimation on the original sample (i.e. not bootstrapped) took 18 s, converging in 615 iterations, with starting values of zero. Bootstrapping (i.e., resampling the respondents, taking new draws, recreating the data based on the new draws, and estimating) took 16 min for 20 resamples. Without the parallel processing toolbox, which allows for gpu processing, run times were about ten times longer, which is still sufficiently fast to allow extensive testing and exploration of alternative specifications.[12]

[10] I always estimate a standard logit first and then a mixed logit with all normal coefficients, in order to check for any problems or anomalies in the estimation process and results. The estimates from the model with all normals can be used to assist in specifying the parameter space for the more flexible procedures.

[11] The model with a continuous normal distribution (column 1) was estimated with 100 Halton draws per person; 2000 draws for each person is probably more than needed when the continuous distribution is replaced with a discrete grid and the number of parameters raised from 37 to 69.

[12] These times cannot be directly compared to the four hour run-time for the model with all normally distributed WTPs because different codes (Stata and Matlab) were used for the two models.

Table 2
Model in WTP space.

	Normal	Polynomial	Spline
Means			
Commercials	−1.562 (0.4214)	−2.808 (0.5469)	−2.736 (0.4913)
Fast availability	3.945 (0.4767)	3.155 (0.4275)	2.979 (0.4264)
More TV, fewer movies	−0.6988 (0.4783)	−0.2231 (0.4464)	−0.1274 (0.4323)
More content	2.963 (0.4708)	2.514 (0.3716)	2.820 (0.2688)
Share usage only	−0.6224 (0.4040)	0.4827 (0.4097)	0.3265 (0.4481)
Share personal and usage	−2.705 (0.5844)	−2.813 (0.6153)	−3.263 (0.5742)
No service	−27.26 (2.662)	−34.09 (2.065)	−34.10 (2.361)
Price/scale	0.2378 (0.0236)	0.3006 (0.0356)	0.2943 (0.0269)
Standard deviations			
Commercials	3.940 (0.5307)	3.537 (0.2771)	3.305 (0.2836)
Fast availability	3.631 (0.4138)	4.304 (0.3643)	4.348 (0.4117)
More TV, fewer movies	4.857 (0.5541)	4.728 (0.4719)	4.727 (0.4943)
More content	2.524 (0.4434)	2.684 (0.1397)	2.816 (0.2172)
Share usage only	2.494 (0.4164)	2.025 (0.2404)	1.905 (0.2508)
Share personal and usage	6.751 (0.7166)	6.452 (0.5171)	6.764 (0.4950)
No service	19.42 (2.333)	23.08 (1.518)	24.06 (1.392)
Price/scale	0.2539 (0.0391)	0.4276 (0.0593)	0.4166 (0.0433)

The second column of Table 2 gives the mean and standard deviation of each utility parameter under this specification. The left side of Fig. 2 graphs the estimated marginal distribution for each utility parameter. Note that the means and standard deviations are fairly similar to those obtained with a normal distribution, but the shape of the distribution for each utility coefficient is quite different from a normal. Several coefficients are bimodal. For example, most people do not care much about whether their personal and usage information is shared, but a group of respondents are willing to pay a considerable amount to avoid this form of sharing. Interestingly, some people are estimated to like having their personal and usage data shared, presumably because they value the promotional information that is targeted to them based on their

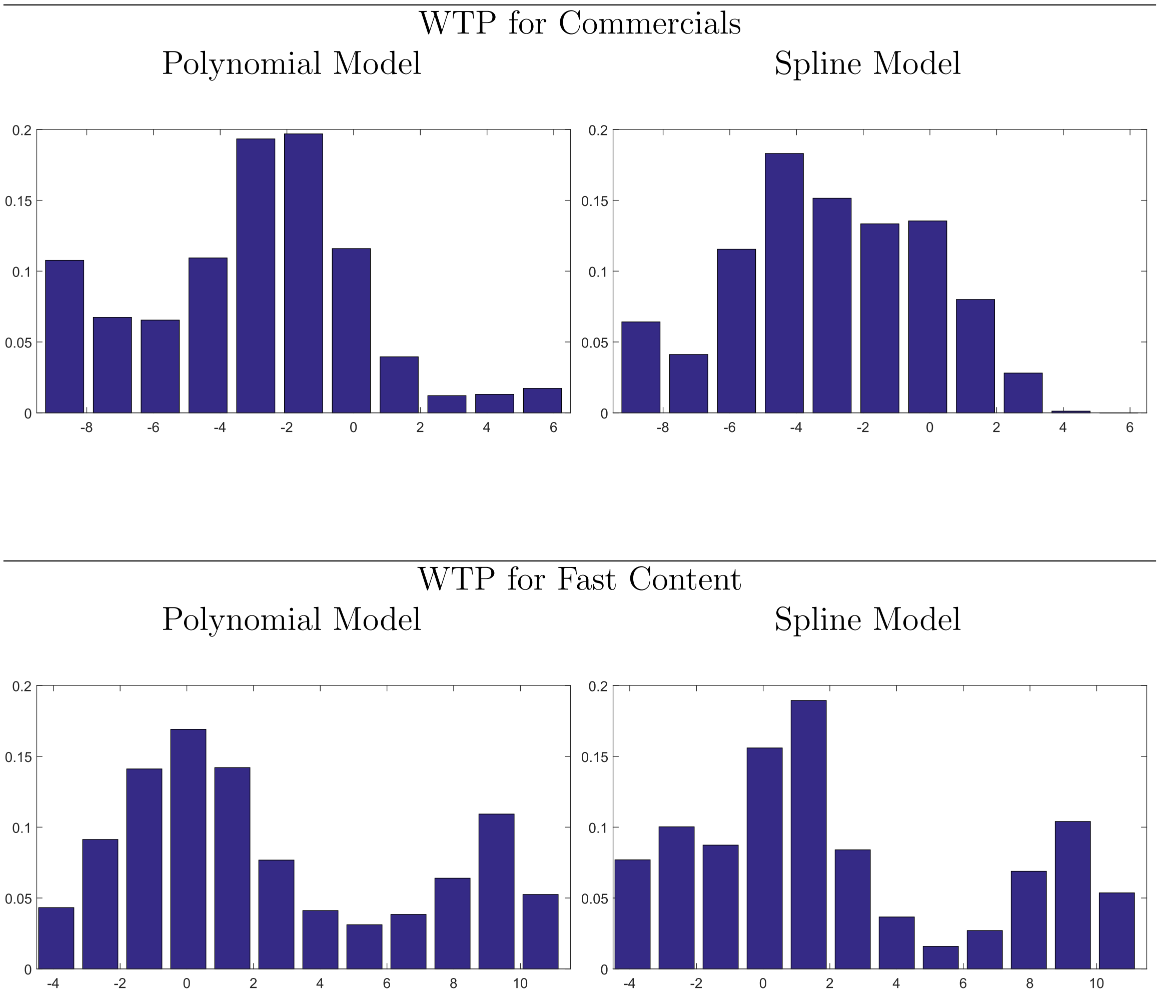

Fig. 2. Distribution of utility parameters.

information. Sharing usage information alone (without personal information) is valued positively by an even greater share of the population, since this form of sharing provides benefits in the form of recommended content, with little loss of confidentiality.

The following pairs of attributes have estimated correlations (not shown) that are statistically significant:

- WTP for sharing usage information and for sharing personal and usage information (0.443): people who like having their usage information shared also tend to like having their personal and usage information shared, and similarly for people who dislike sharing.
- WTP for fast content and for more TV with fewer movies (0.410), indicating that people who want content to be available quickly also want to have more TV shows. Stated more intuitively: people who like TV shows more than movies also want the TV shows to be available soon after they are originally shown.
- WTP for commercials and for more content (0.372): people who want more content also don't mind seeing commercials as much as other people.
- WTP for sharing personal and usage information and no service (-0.285): people who dislike having their personal and usage information shared are also more willing to do without the video service altogether.

In the model with normals, the first and second of these correlations are significant, and the other two are not.

The model attained a SLL at convergence of -3864.85, compared to -3903.47 for the model with a normal distribution. In the latter, the normal distribution was estimated with unbounded support, which means that the two models are not precisely nested. For direct comparison, an LML model was estimated with a second-order polynomial instead of six-order. The SLL was -3912.86, giving a likelihood ratio test statistic of 96.02. The critical value of chi-squared with 32 degrees of freedom (the number of additional parameters in the more general model) is 47.40 at the 95% confidence level. The hypothesis that the extra parameters are zero can be rejected.

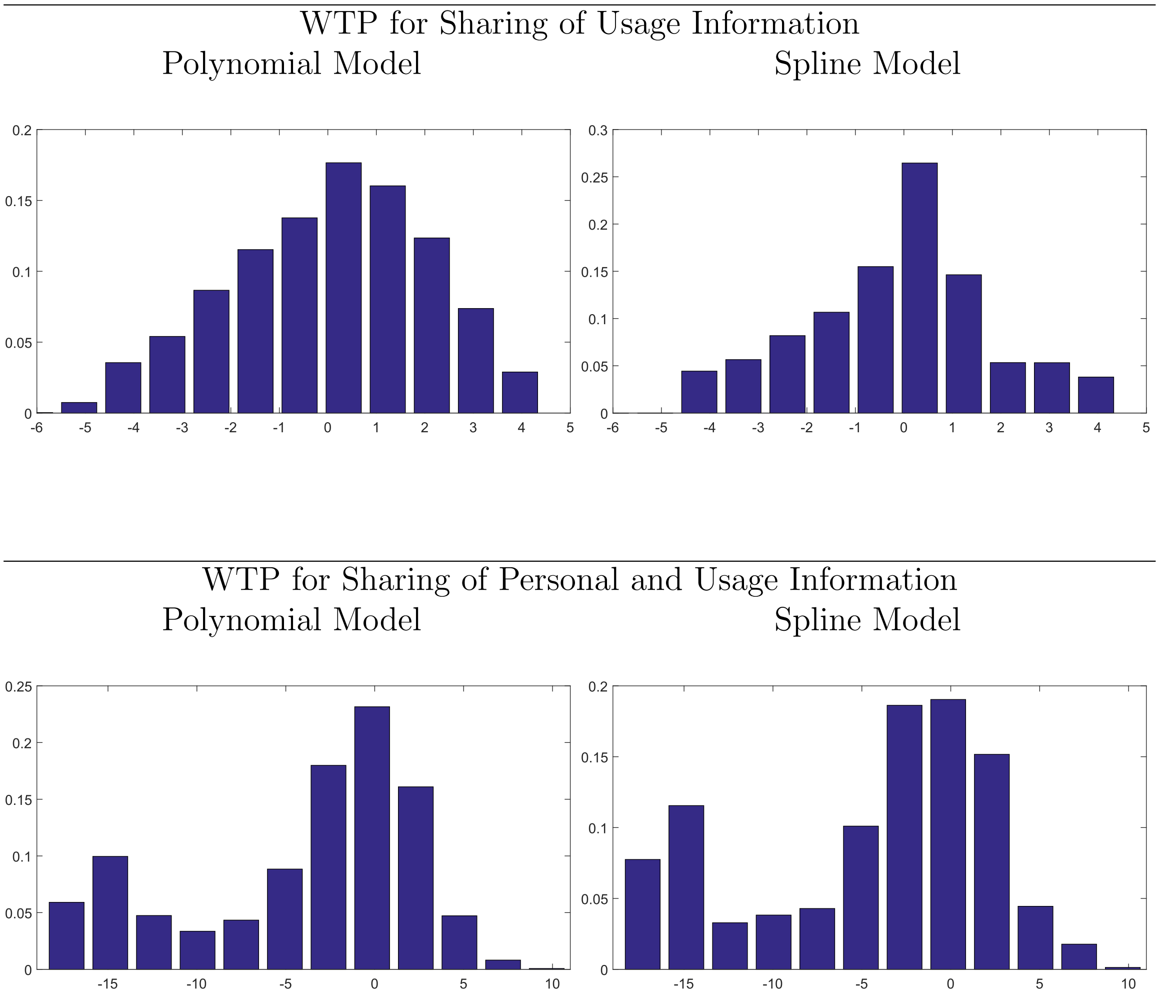

Fig. 2. (*continued*)

6.3. Splines

Using the same specification of *S*, a model was estimated using the combination described above, namely, a six-segment spline for each utility parameter and a second-order polynomial in the WTPs. The model contains 83 parameters: 6 for each dimension's spline, giving 48 in total, plus 35 for the polynomial in the 7 WTPs. Using the gpu, estimation took about 50 s on the original sample, with convergence in 1669 iterations, and the bootstrapping took 37 min.

The estimated means and standard deviations are given in the third column of Table 2, and the marginal distributions are shown on the right side of Fig. 2. The means and standard deviations are fairly similar to those for the other two models. The shapes of the distributions are similar to those obtained with polynomials, and very different from normal.

The correlations that were significant in the polynomial model are also significant in the spline model and have similar magnitudes (0.476 v. 0.443; 0.448 v. 0.410; 0.486 v. 0.372; −0.228 v. −0.285). One additional correlation is significant in the spline model, namely, WTP for fast content and for sharing personal and usage information (−0.356): people who want fast content also dislike having their personal and usage data shared.

7. Discussion

The procedure allows for fast and easy specification of flexible mixing distributions. However, the flexibility places additional burden on the researcher to specify the range of each utility parameter and the *z* variables that describe the shape of the distribution. The researcher's data might not be sufficiently informative to provide authoritative guidance in making these specification decisions. Different distributions can provide very similar log-likelihood values, which leaves the researcher with the unsatisfying option, given current data, of choosing one distribution over another based on tiny differences in log-likelihood.

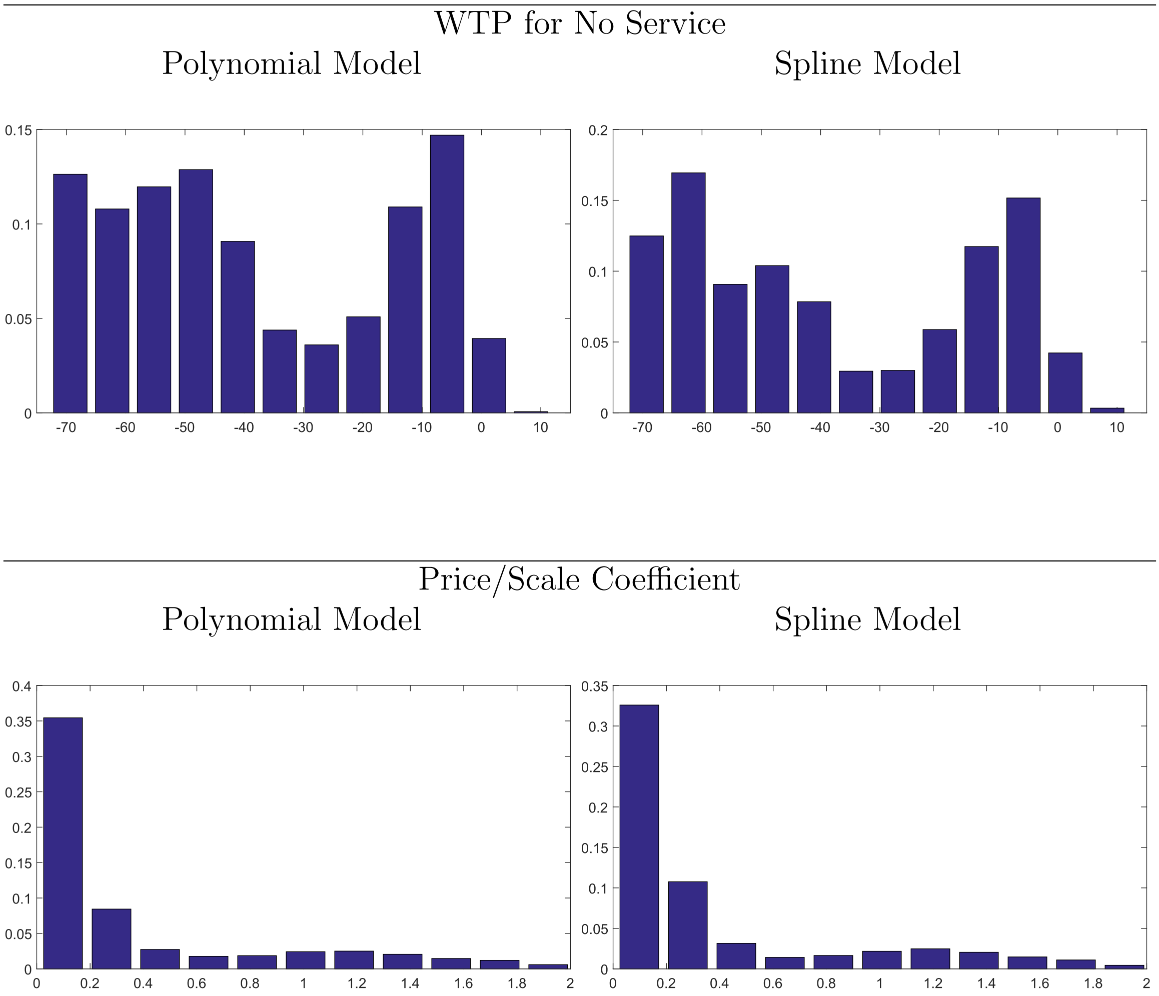

Fig. 2. (*continued*)

The range of values that is allowed for each coefficient or WTP is also an issue. Parameters might be expected, by rational choice, to be non-negative, or non-positive, for all people, and yet a model that specifies a distribution with the theoretically expected support for these parameters might (and in my experience, often does) fit worse than a model that allows a share of values with the "wrong" sign. In the current application, for example, WTP for commercials, fast content, and more content can be theoretically signed. The spline model was re-estimated with the specified end-point for the range of WTPs being zero (at the high end for commercials and the low end for fast content and more content), and the log-likelihood dropped from -3858.74 to -3886.70. This reduction is not evidence that the procedure for estimating the mixing distribution is problematic. Rather, the problem arises because the data do not contain enough information to rule out theoretically implausible behavior.

These issues point to the need for richer data that can more meaningfully distinguish among distributions and better align empirical results with theoretical expectations or requirements. The availability of procedures that accommodate flexible distributions shifts the focus for future research away from concern about overcoming specification constraints to developing forms of data that more clearly reveal the distribution of consumers' preferences.

In regard to the LML model itself, four interrelated issues seem to me to be the most relevant topics for further investigation and improvement. (1) As stated above, delineating the relation of LML models to nonparametric estimation would greatly enhance our understanding of how best to specify and interpret the models. (2) Non-equal probability sampling of points from S has the potential to improve the fit of the mixing distribution and increase the range of the coefficients in S. This potential can be explored through applications and Monte Carlo exercises. (3) It would be useful to determine the extent to which summary statistics for the estimated mixing distribution, such as the mean and standard deviation, depend on the range of coefficients that is used to define S, and how the range can be specified to obtain more accurate and robust estimates of the relevant summary statistics. (4) The tails of distributions are often vitally important for policy and marketing purposes, e.g., what share of households are willing to pay more than some specified amount for an improvement in an attribute. Flexible methods, such as LML, allow the tails to be determined by the data rather than by

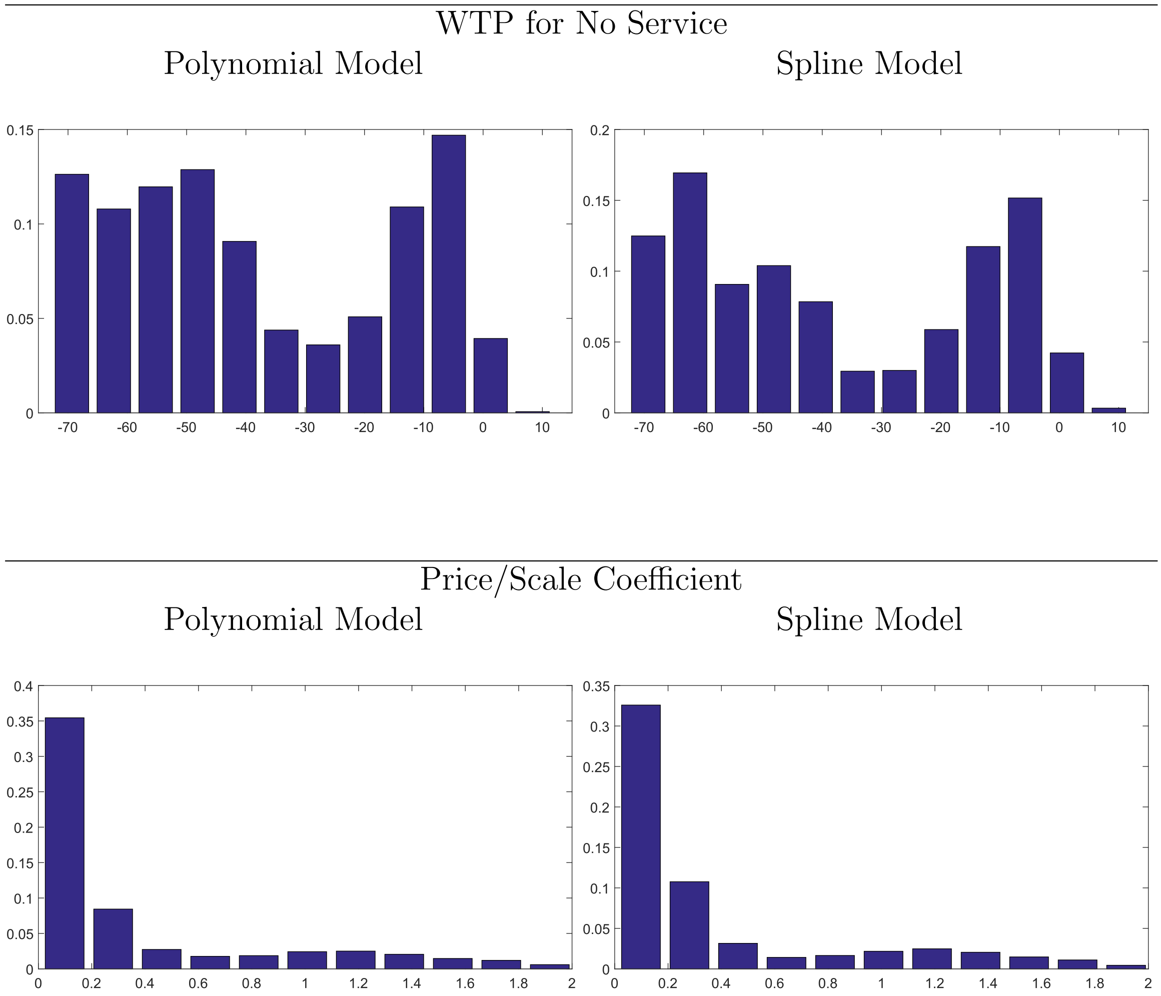

Fig. 2. (*continued*)

assumption, but increase the requirement for data that are in the tails (see, for example, Borjesson et al., 2012). Topic (2) is related to (3) and (4), since unequal probability sampling from S expands the possibilities for feasible specification of S and identification of the tails.

References

Bajari, P., Fox, J., Ryan, S., 2007. Linear regression estimation of discrete choice models with nonparametric distributions of random coefficients. Am. Econ. Rev. 97 (2), 459–463.

Ben-Akiva, M., McFadden, D., Train, K., 2015. Foundations of stated preference elicitation, working paper, Department of Economics, University of California, Berkeley. http://eml.berkeley.edu/train/foundations.pdf.

Borjesson, M., Fosgerau, M., Algers, S., 2012. Catching the tail: empirical indentification of the distribution of the value of travel time. Transp. Res. Part A: Policy Pract. 46 (2), 378–391.

Burda, M., Harding, M., Hausman, J., 2008. A Bayesian mixed logit-probit model for multinomial choice. J. Econ. 147 (2), 232–246.

Chamberlain, G., 1987. Asymptotic efficiency in estimation with conditional moment restrictions. J. Econ. 34, 305–334.

Chen, X., 2007. Large sample sieve estimation of semi-nonparametric models. In: Heckman, J., Leamer, E. (Eds.), Handbook of Econometrics, vol. 6A. , North-Holland, New York. (Chapter 76).

Fosgerau, M., Bierlaire, M., 2007. A practical test for the choice of mixing distribution in discrete choice models. Transp. Res. Part B: Methodol. 41 (7), 784–794.

Fosgerau, M., Mabit, S., 2013. Easy and flexible mixing distributions. Econ. Lett. 120, 206–210.

Fox, J., Kim, K., Ryan, S., Bajari, P., 2011. A simple estimator for the distribution of random coefficients. Quant. Econ. 2, 381–418.

Glasgow, G., Butler, S., 2016. The value of non-personally identifiable information to consumers of online services: evidence from a discrete choice experiment, Applied Economics Letters, (forthcoming).

Gourieroux, C., Monfort, A., 1993. Simulation-based inference: a survey with special reference to panel data models. J. Econ. 59, 5–33.

Hajivassiliou, V., Ruud, P., 1994. Classical estimation methods for LDV models using simulation. In: Engle, R., McFadden, D. (Eds.), Handbook of Econometrics, North-Holland, New York, pp. 2383–2441.

Lee, L., 1995. Asymptotic bias in simulated maximum likelihood estimation of discrete choice models. Econ. Theory 11, 437–483.

McFadden, D., 1978. Modelling the choice or residential location. In: Karlqvist, A., Lundqvist, L., Snickars, F., Weibull, J. (Eds.), Spatial Interaction Theory and Planning Models, North-Holland, New York, pp. 105–142.

McFadden, D., 1989. A method of simulated moments for estimation of discrete response models without numerical integration. Econometrica 57, 995–1026.

McFadden, D., Train, K., 2000. Mixed MNL models of discrete response. J. Appl. Econ. 15, 447–470.

McFadden, D., 1975. On independence, structure and simultaneity in transportation demand analysis, Working Paper No. 7511, Institute of Transportation and Traffic Engineering, University of California, Berkeley.

Revelt, D., Train, K., 1998. Mixed logit with repeated choices. Rev. Econ. Stat. 80, 647–657.

Scarpa, R., Theine, M., Train, K., 2008. Utility in willingness to pay space: a tool to address confounding random scale effects in destination choice in the Alps. Am. J. Agric. Econ. 90 (4), 994–1010.

Sohrab, H., 2003. Basic Real Analysis, Birkhuser, Boston. second ed., 2014, Springer, Dordrecht.

Train, K., 2008. EM algorithms for nonparametric estimation of mixing distributions. J. Choice Model. 1, 40–69.

Train, K., 2009. In: Discrete Choice Methods with Simulation, 2nd ed. Cambridge University Press, New York.

Train, K., Weeks, M., 2005. Discrete choice models in preference space and willingness-to-pay space. In: Scarpa, R., Alberini, A. (Eds.), Applications of Simulation Methods in Environmental and Resource Economics, Springer, Dordrecht, pp. 1–17.