Easy and flexible mixture distributions

Fosgerau, Mogens and Mabit, Stefan

Technical University of Denmark, Center for Transport Studies, Sweden

2013

Easy and flexible mixture distributions

Mogens Fosgerau* Stefan L. Mabit†

March 14, 2013

Abstract

We propose a method to generate flexible mixture distributions that are useful for estimating models such as the mixed logit model using simulation. The method is easy to implement, yet it can approximate essentially any mixture distribution. We test it with good results in a simulation study and on real data.

Keywords:
Mixture distributions; mixed logit; simulation; maximum simulated likelihood
JEL: C14, C15, C25

*Corresponding author, Technical University of Denmark and Centre for Transport Studies, Sweden, mf@transport.dtu.dk, DTU Transport, Bygningstorvet, Building 116B, 2800 Kgs. Lyngby, Denmark, +45 45256521

†Technical University of Denmark, smab@transport.dtu.dk

1 Introduction

This paper presents an easy yet powerful method for creating a mixture distribution for a random parameter in an econometric model that is estimated using simulation. The method is presented using maximum simulated likelihood estimation of the mixed logit model as an example, but can be applied in a wide range of circumstances. The advantages of the method are that essentially any distribution can be represented arbitrarily well, while implementation is very simple.

Consider a model that specifies the likelihood $P(y|x,\beta)$ of some outcome y conditional on variables x and an unobserved random parameter β having distribution F.[1] Assuming that x and β are independent, the likelihood $P(y|x)$ may be simulated given R independent draws β_r from F. This is the basis for estimation by simulation (Train, 2003; McFadden, 1989), which can be applied when the distribution F is considered as known.

Most applications of this method rely on the inversion method for generating draws from F: If u_r are draws from a standard uniform distribution, then $F^{-1}(u_r)$ are draws from F. In order to use this method, it is necessary to compute the inverse of F explicitly.[2]

There are many situations where it is not desirable to impose a specific functional form on F. Generally, this is the case whenever the choice of F has impact on the object of interest for the investigation but there is no a priori reason to choose a particular F. It is particularly undesirable to impose a specific form on F when F is the object of interest itself, e.g., when the purpose is to estimate a distribution of willingness-to-pay. Then it is preferable if the shape of F can be estimated. This can be accomplished by the method of sieves (see e.g. Chen, 2007; Gallant and Nychka, 1987), also known as series estimators. It is however necessary to guarantee that the approximation of F is actually a CDF and then it must be inverted in order to generate random draws from F using the inversion method.

Another idea is to approximate F^{-1} directly. Then inversion is unnecessary. It is however still necessary to ensure that F^{-1} is monotone, which might involve somewhat complicated restrictions on the deep parameters of F^{-1} in a series approximation.

The key insight of this paper is that approximating F or F^{-1} is actually an unnecessary complication for the present purpose. All that is required for simulating the likelihood is draws β_r from some distribution F that depends on some deep

[1] There will generally be other parameters to be estimated in the likelihood. They are suppressed in the notation here as the focus lies elsewhere.

[2] Devroye (1986) provides a comprehensive treatment of techniques for random variable generation.

parameters to be estimated. The simulated likelihood is simply

$$\frac{1}{R}\sum_{r} P\left(y|x,\beta_r\right). \tag{1}$$

It is not necessary that the draws β_r are monotone functions of standard uniform draws. It is not even necessary to know explicitly the distribution of the draws β_r in order to compute (1); the ability to generate draws from the distribution is sufficient. Being able to obtain the draws, it is always possible to estimate their distribution.

In this paper we take draws u_r from some distribution and transform them using a power series

$$f\left(u|\alpha\right) = \sum_{k=0}^{K} \alpha_k u^k \tag{2}$$

to compute random draws $\beta_r = f\left(u_r|\alpha\right)$ that depend on deep parameters $\alpha = (\alpha_0, ..., \alpha_K)$ to be estimated. The random draws are inserted into (1) and the resulting expression is very easy to implement in software. For instance, if the model contains a term βx, then that is replaced by $\sum_{k=0}^{K} \alpha_k \left(x u_r^k\right)$. This is a convenient form, since it is linear in deep parameters α that are multiplied by easily computed variables $x u_r^k$. In most cases the distribution of $f\left(u|\alpha\right)$ is not easily derived analytically. The distribution is by construction, however, very easy to simulate, which is all that is really needed.

A predecessor of our method is Fleishman (1978), who considers the problem of generating random variables with prespecified moments. He generates a random variable as a third-order polynomial in a standard normal random variable and provides formulae for the coefficients of the polynomial such that specific values of the first four moments are matched by such a variable. The present case is similar, except we are not concerned with matching given moments, but estimate coefficients in order to match a given dataset and may use polynomials of any degree. We present results using both uniform and normal draws.

The following section 2 presents some properties of the proposed method. It will also be argued that essentially any distribution can be approximated arbitrarily well by (2) by choosing a sufficiently large number of parameters K. This section also discusses extension to multivariate random parameter distributions. Section 3 provides simulation results that illustrate the ability of the method to recover various true distributions from binary discrete choice panel data. Section 4 presents an application to real data and section 5 concludes.

2 Some properties of the method

Let $\alpha = (\alpha_0, ..., a_K) \in \mathbb{R}^K$ be a parameter vector and let u be a random variable. Then $\beta = f(u|\alpha) = \sum_{k=0}^{K} \alpha_k u^k$ is a random variable and it is convenient for use as a random parameter. The following proposition summarises a few properties of β.

Proposition 1 *Let u follow a uniform distribution. Then the random parameter β has compact support ranging between α_0 and $\sum_{k=0}^{K} \alpha_k$, either of which may be greatest; the mean is*

$$E\beta = \sum_{k=0}^{K} \frac{\alpha_k}{1+k},$$

and the m'th raw moment $(m > 1)$ is

$$E(\beta^m) = \sum_{k_1=0,...,k_m=0}^{K,...,K} \frac{\prod_{i=1}^{m} \alpha_{k_i}}{1 + \sum_{i=1}^{m} k_i}.$$

The variance of β is

$$\begin{aligned} V(\beta) &= E\left(\beta^2\right) - (E\beta)^2 \\ &= \sum_{k=0,j=0}^{K,K} \frac{kj\alpha_k\alpha_j}{(1+k+j)(1+k)(1+j)} \end{aligned}$$

Proof. Immediate. ■

Remark 1 *It is straightforward (but quite tedious) to show that with uniform u and $K = 2$, then it is possible to attain any skewness while maintaining that $E\beta = 0$ and $E\left(\beta^2\right) = 1$.*

Remark 2 *If the first K moments are to be matched, it may be necessary to include more than K terms. The necessity of this has been shown for a third-order polynomial in a standard normal random variable (Headrick, 2002).*

Remark 3 *By the Weierstrass approximation theorem, the set of functions $\left\{f(\cdot|\alpha)\,|\alpha \in \mathbb{R}^{(\mathbb{N})}\right\}$ uniformly approximates any continuous function on the unit interval. This comprises all inverse CDF of distributions that have densities.*

Remark 4 *Consistency of series estimators has been established for a range of cases (see e.g. Geman and Hwang, 1982; Chen, 2007; Bierens, 2008; Fosgerau and Nielsen, 2010), but not formally for the present. Consistency of the proposed estimator seems highly likely, meaning that the estimated distribution of β will become arbitrarily close to the true distribution given a large enough dataset and a correspondingly large value of K. For a fixed K, the standard results regarding consistency of maximum simulated likelihood apply (Newey and McFadden, 1994; Hajivassiliou and Ruud, 1993).*

Remark 5 *Given R i.i.d. draws β_r from some distribution, its CDF F can be estimated by*

$$F(t) = E(1\{\beta \leq t\}) \simeq \frac{1}{R}\sum_r 1\{\beta_r \leq t\}.$$

As β_r are the results of simulation, we are free to choose R and hence it can be chosen to achieve any desired degree of precision of the estimate of F.

2.1 Multivariate distributions

The method can be extended to allow for a multivariate random parameter. The extension is straightforward if the random parameters are independent, so in the following we allow them to be dependent.

One way to go is to combine the proposed method with a copula. Let c be the density of a bivariate copula function, i.e. a density on the unit cube with uniform marginal distributions. A range of such are known (Joe, 1997; Nelsen, 2006). If the conditional likelihood of an observation given (β_1, β_2) is $P(\beta_1, \beta_2)$, then we could use

$$\int_{I^2} P(\beta_1(u_1), \beta_2(u_2))\, c(u_1, u_2)\, du_1 du_2$$

to create dependence. This is however not a very attractive option, since it requires the likelihood to be extended with a new term $c(u_1, u_2)$. Note also that while c corrects the likelihood for dependence between uniform random variables u_1 and u_2, it is not the copula for the random variables $\beta_1(u_1), \beta_2(u_2)$ since the functions β_1, β_2 may not be inverse CDF.

A simpler way to go is the following. Say again for simplicity that we want a two-dimensional random parameter $\beta = (\beta_1, \beta_2)$, extension will be straightforward. Let

$$\beta_i = \sum_{j,k=0}^{K} \alpha_{i,jk} u_1^j u_2^k.$$

This is as easy to implement as what we have discussed in the univariate case. With this specification, β_1, β_2 will be dependent if $\alpha_{i,jk} \neq 0$ for some $j > 0$ or $k >$

0. It is thus possible to allow for dependence by including such cross-parameters. This is fully flexible in the limit, but in practice the curse of dimensionality will quickly prevent inclusion of all cross-terms. It is still possible to include only some cross-terms and obtain some forms of dependence.

3 Simulation exercise

This section presents the ability of the proposed method to recover various known distributions from simulated panel binary choice data. Datasets were generated using a range of distributions F, chosen to represent a challenging range of different shapes. For every distribution, 50 datasets were generated for 1000 individuals each making 8 standard binary logit choices with probability of alternative 1 given by

$$P\left(1|\beta, x\right) = \frac{1}{1 + e^{\sigma(\beta - x)}},$$

where $\sigma = 2$ is a scale parameter, β is an individual-specific parameter following a known population distribution F and x is an observed variable drawn from a standard normal distribution. Datasets were generated for the following six distributions: a) Standard normal, b) Standard uniform (support $[-1, 1]$), c) Shifted lognormal (constructed as $X/2 - 1$, where X is standard lognormal), d) Mixture of two normals (equal weight, locations -1 and 1, standard deviations 1/2), e) Beta$(2, 5)$, f) Asymmetric triangular (support $[-1, 1]$, mode 1/2).

We observe (y, x) and estimate both σ and the distribution of β, specifying the distribution of β as a third-order polynomial of a standard uniform random variable in the way described in section 2. We apply maximum simulated likelihood using 500 Halton draws.

We estimated the model on each of the 50 datasets and report plots showing the true distribution together with the pointwise mean and 90% confidence band for the estimated distribution.[3] The results of this exercise are shown in Figure 1.

Overall, the results are very satisfactory. In all cases the confidence bands are quite tight, showing that the estimated distributions do not vary much over the 50 generated datasets. The confidence bands track the true distributions quite closely which shows that a third-order polynomial is sufficient to reproduce the main features of the distributions considered. The mixture of normals does, however, stretch the ability of the third power approximation to track its shape. Repeating the simulation exercise, basing the simulation of β on standard normal draws rather than uniform, led to similar results.[4]

[3] Data generation and estimation were carried out in Ox (Doornik, 2001). The code is available from the authors on request.

[4] These results are available from the authors on request.

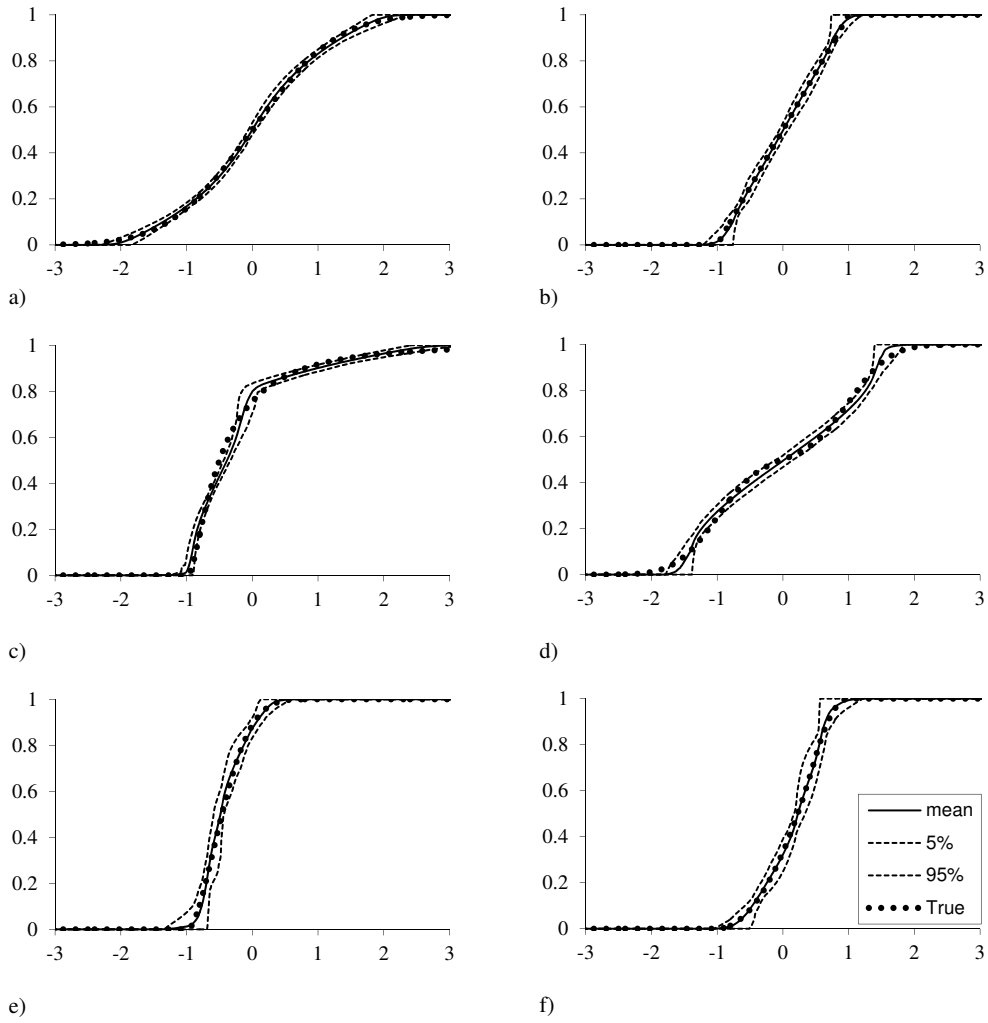

Figure 1: Simulation results using a third-order polynomial in a standard uniform random draw. a: normal, b: uniform, c: lognormal, d: mixture of two normals, e: beta, f: asymmetric triangular. For each simulation the figure shows the true distribution, the pointwise mean of the estimated distributions over 50 repetitions as well as the pointwise 5 and 95 percent quantiles of the estimates.

4 Application to real data

We use a dataset collected with the purpose of estimating the willingness-to-pay (WTP) for travel time savings (Fosgerau et al., 2007) and we adopt the data selection and model specification of Fosgerau (2006), using observations of 2,197 car drivers choosing between two car trips distinguished by cost and time only. Drivers made 8 choices each; with a few observations omitted for various reasons, 17,020 observations remained for estimation. We estimate a panel mixed binary logit model with the dependent variable defined by

$$y_{it} = 1 \Leftrightarrow \delta' x_{it} + \beta_i + \varepsilon_{it} > \lambda \ln v_{it},$$

where λ is a scale parameter to be estimated, v_{it} is the trade-off value of travel time implicitly presented for respondent i in choice occasion t, δ is a vector of parameters to be estimated, x_{it} is a vector of explanatory variables, the same as in Fosgerau (2006), β_i is an individual-specific parameter with an unknown distribution to be estimated and ε_{it} are i.i.d. standard logistic. The data are coded such that we observe $y_{it} = 1$ if the respondent prefers the faster and more expensive alternative over the slower and less expensive alternative, which we take as an indication that his individual-specific random WTP

$$\exp\left(\frac{\delta' x_{it} + \beta_i}{\lambda}\right)$$

is greater than the price of time v_{it} implicit in the offered choice.

We estimate nine different models using a standard off-the-shelf software for estimation of discrete choice models.[5] The first uses just a normal distribution for β, such that the WTP becomes lognormal, which provided the best fit to the data in Fosgerau (2006). The next four models use a polynomial in a standard uniform random variable with powers up to 1, 2, 3 and 4, respectively.

The parameter estimates are shown in Table 1. The parameters in the linear index $\delta' x_{it}$ are relatively insensitive to the distribution of β that is imposed. Compared on the loglikelihood to the model based on the normal distribution, the fit of the models with a polynomial in a uniform is worse for first- and second-order polynomials but better with third- and fourth-order polynomials. The Aikaike information criterion (AIC) prefers the model with a third-order polynomial.

Table 2 shows the estimation results for the base model and four additional models using a polynomial in the normal distribution. Again we find that parameters in the linear index are relatively insensitive to the distribution of β that

[5]The models were estimated in Biogeme 2.0 (Bierlaire, 2005) with 100 Halton draws. A test with 500 Halton draws led to no significant change in the results. The code is available from the authors on request.

Table 1: ESTIMATION RESULTS FOR THE BASE AND FOUR MODELS BASED ON THE UNIFORM DISTRIBUTION

	ML normal		Uniform 1		Uniform 2		Uniform 3		Uniform 4	
Variables	estimate	z test	estimate	z test	estimate	z test	estimate	z test	estimate	z test
Age/10	0.02	0.12	-0.004	-0.02	-0.05	-0.25	-0.01	-0.07	-0.02	-0.10
Age squared/100	-0.03	-1.68	-0.03	-1.38	-0.02	-1.19	-0.03	-1.42	-0.03	-1.41
Commute dummy	0.34	3.49	0.34	3.22	0.42	4.00	0.35	3.66	0.36	3.71
ConShare	0.52	1.66	0.43	1.39	0.33	1.06	0.56	1.68	0.55	1.60
Education dummy	0.27	1.62	0.18	1.16	0.19	1.22	0.26	1.32	0.24	1.24
Female dummy	-0.29	-3.41	-0.27	-3.05	-0.27	-3.18	-0.29	-3.39	-0.29	-3.43
Ln(income)	0.71	7.73	0.66	6.68	0.66	6.90	0.68	7.18	0.67	7.22
Income NA dummy	0.86	4.93	0.87	4.75	0.87	4.99	0.78	4.55	0.78	4.57
Time difference	0.37	8.81	0.34	8.04	0.33	7.89	0.37	9.10	0.38	9.12
Trip duration	0.44	8.17	0.40	7.39	0.43	8.07	0.42	7.54	0.42	7.52
Constant	1.18	2.43	-1.07	-2.15	-0.55	-1.10	-2.31	-4.05	-2.57	-3.98
α_1	1.56	33.35	5.00	34.31	2.10	2.94	18.4	8.63	22.8	4.05
α_2	-	-	-	-	2.76	4.12	-34.2	-7.38	-52.4	-2.44
α_3	-	-	-	-	-	-	23.4	8.03	50.0	1.63
α_4	-	-	-	-	-	-	-	-	-12.8	-0.87
λ	1.15	35.17	1.13	35.09	1.13	35.15	1.15	35.37	1.15	35.36
DoF	13		13		14		15		16	
No. observations	17020		17020		17020		17020		17020	
No. individuals	2197		2197		2197		2197		2197	
Final LL	-9051.9		-9102.9		-9088.6		-9045.5		-9045.1	
Adjusted ρ^2	0.232		0.227		0.228		0.232		0.232	
AIC	18129.8		18231.8		18205.2		18121		18122.2	

Table 2: ESTIMATION RESULTS FOR THE MODELS BASED ON THE NORMAL DISTRIBUTION

	ML normal		Normal 2		Normal 3		Normal 4		Normal 5	
Variables	estimate	z test	estimate	z test	estimate	z test	estimate	z test	estimate	z test
Age/10	0.02	0.12	0.001	0.00	0.001	0.01	-0.005	-0.02	0.003	0.01
Age squared/100	-0.03	-1.68	-0.03	-1.55	-0.03	-1.53	-0.03	-1.52	-0.03	-1.59
Commute dummy	0.34	3.49	0.38	3.83	0.35	3.63	0.36	3.73	0.36	3.81
ConShare	0.52	1.66	0.47	1.49	0.52	1.61	0.51	1.56	0.46	1.41
Education dummy	0.27	1.62	0.28	1.67	0.32	1.80	0.31	1.72	0.30	1.73
Female dummy	-0.29	-3.41	-0.29	-3.46	-0.29	-3.43	-0.29	-3.51	-0.29	-3.58
Ln(income)	0.71	7.73	0.71	7.79	0.72	7.89	0.72	7.91	0.72	7.94
Income NA dummy	0.86	4.93	0.85	4.95	0.79	4.63	0.80	4.70	0.80	4.73
Time difference	0.37	8.81	0.36	8.74	0.38	9.09	0.38	9.15	0.38	9.14
Trip duration	0.44	8.17	0.45	8.43	0.44	8.07	0.44	8.08	0.44	8.10
Constant	1.18	2.43	1.11	2.31	1.13	2.33	1.10	2.27	1.04	2.19
α_1	1.56	33.35	1.54	32.95	1.29	16.67	1.25	14.26	1.35	9.68
α_2	-	-	0.10	2.72	0.08	1.70	0.24	4.12	0.43	2.24
α_3	-	-	-	-	0.13	3.59	0.16	3.28	-0.01	-0.08
α_4	-	-	-	-	-	-	-0.05	-4.67	-0.13	-1.59
α_5	-	-	-	-	-	-	-	-	0.05	1.06
λ	1.15	35.17	1.15	35.18	1.15	35.29	1.15	35.31	1.15	35.31
DoF	13		14		15		16		17	
No. observations	17020		17020		17020		17020		17020	
No. individuals	2197		2197		2197		2197		2197	
Final LL	-9051.9		-9047.6		-9038.9		-9036.7		-9035.9	
Adjusted ρ^2	0.232		0.232		0.233		0.233		0.233	
AIC	18129.8		18123.2		18107.8		18105.4		18105.8	

is imposed. Of course, since the base model uses the normal distribution, now the loglikelihood improves with each additional power of the normal distribution. The AIC prefers the model with a fourth-order polynomial among all models estimated. The models with four and five powers of the normal distribution required many iterations to converge; this seems to be related to collinearity of the second and the fourth powers of the normal. Our reason for using the powers of the normal random variable is that we want to show that the simple implementation works. It is, however, possible to replace the powers of the normal random variable by orthogonal polynomials in the normal random variable and this could plausibly resolve the collinearity issue.

Figure 2 plots the estimated cumulative distributions for three models, namely the base model, the third-order polynomial model which is the best model based on the uniform distribution and the fourth-order polynomial model which is the best model based on the normal distribution. The scale parameter changes only little so the cumulative distributions are comparable. In this case, the deviations from the normal distribution do not appear large, although the polynomial terms clearly improve the model fit.

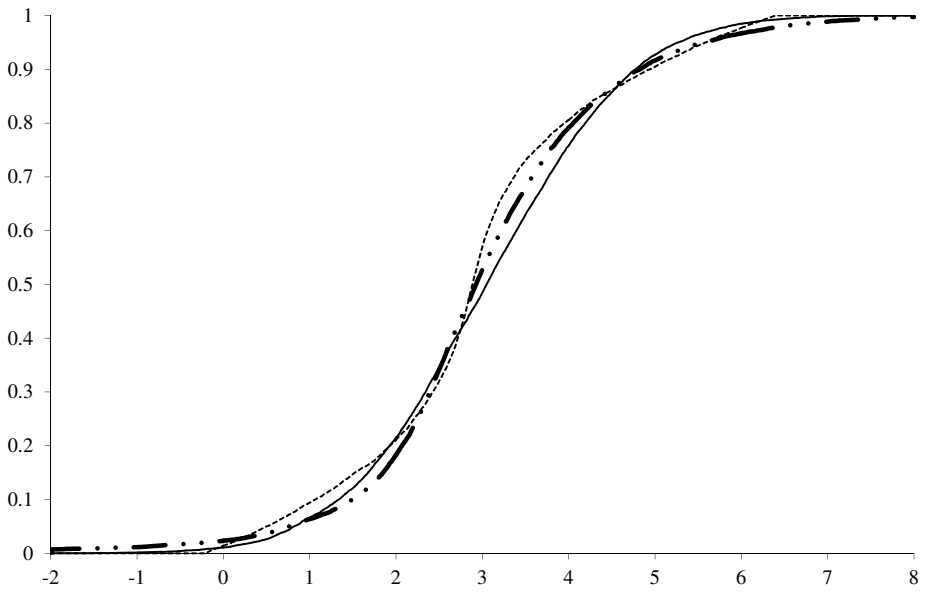

Figure 2: Estimation results using a normal distribution (solid line), a third-order polynomial in a standard uniform random draw (dots) and a fourth-order polynomial in a standard normal random draw (dots and dashes).

5 Conclusion

This paper has developed and applied a simple method for creating flexible mixing distributions. It is easy to implement and the mixing distributions can be arbitrarily flexible. The method has been applied successfully in a simulation study as well as to real data, both using the mixed logit model estimated with maximum simulated likelihood. The application to real data was carried out in a freely available and much used package for estimation of discrete choice models, demonstrating that the method is readily applicable and does not require specialised programming.

Acknowledgement

Support from the Danish Strategic Research Council is gratefully acknowledged.

References

Bierens, H. J., 2008. Semi-nonparametric interval-censored mixed proportional hazard models: Identification and consistency results. Econometric Theory 24 (3), 749–794.

Bierlaire, M., 2005. An introduction to biogeme. biogeme.epfl.ch.

Chen, X., 2007. Large sample sieve estimation of semi-nonparametric models. In: Heckman, J. J., Leamer, E. E. (Eds.), Handbook of Econometrics. Vol. 6, Part B. pp. 5549–5632.

Devroye, L., 1986. Non-Uniform Random Variate Generation. Springer-Verlag, New York.

Doornik, J., 2001. Ox: An Object-Oriented Matrix Language. Timberlake Consultants Press, London.

Fleishman, A. I., 1978. A method for simulating non-normal distributions. Psychometrika 43 (4), 521–532.

Fosgerau, M., 2006. Investigating the distribution of the value of travel time savings. Transportation Research Part B: Methodological 40 (8), 688–707.

Fosgerau, M., Hjorth, K., Vincent Lyk-Jensen, S., 2007. The danish value of time study - final report. Tech. rep., www.dtf.dk.
URL http://www.transport.dtu.dk/upload/institutter/

dtu%20transport/pdf_dtf/rapporter/the%20danish%20value%20of%20time%20study_250208.pdf

Fosgerau, M., Nielsen, S. F., 2010. Deconvoluting preferences and errors: a model for binomial panel data. Econometric Theory 26 (6), 1846–1854.

Gallant, A., Nychka, D., 1987. Semi-nonparametric maximum likelihood estimation. Econometrica 55 (2), 363–390.

Geman, S., Hwang, C.-R., 1982. Nonparametric maximum likelihood estimation by the method of sieves. The Annals of Statistics 10 (2), 401–414.

Hajivassiliou, V. A., Ruud, P., 1993. Classical estimation methods for LDV models using simulation. In: Engle, R., McFadden, D. L. (Eds.), Handbook of Econometrics. Vol. IV. pp. 2383–2440.

Headrick, T. C., 2002. Fast fifth-order polynomial transforms for generating univariate and multivariate nonnormal distributions. Computational Statistics & Data Analysis 40 (4), 685–711.

Joe, H., 1997. Multivariate Models and Dependence Concepts. Chapman & Hall, London.

McFadden, D., 1989. A method of simulated moments for estimation of the multinomial probit without numerical integration. Econometrica 57 (5), 995–1026.

Nelsen, R. B., 2006. An introduction to copulas. Vol. 2 of Springer Series in Statistics.

Newey, W. K., McFadden, D., 1994. Large sample estimation and hypothesis testing. In: Engle, R., McFadden, D. L. (Eds.), Handbook of Econometrics. Vol. IV. pp. 2111–2245.

Train, K., 2003. Discrete Choice Methods with Simulation. Cambridge University Press, Cambridge, MA.