

# Appendix to “Mixed Logit Models for Multiparty Elections”: Estimation of Mixed Logit Models

The parameters to be estimated in a mixed logit model are  $\beta$ , the vector of fixed coefficients, and  $\Omega$ , the parameters that describe the distribution of  $\eta$ . The mixed logit log-likelihood function for given values of the parameters  $\beta$  and  $\Omega$  is:

$$\mathcal{L}(\beta, \Omega) = \sum_I \sum_J y_{ij} \log \left[ \int_{\eta_i} \left\{ \frac{e^{X_{ij}\beta_j + Z_{ij}\eta_i}}{\sum_k e^{X_{ik}\beta_k + Z_{ik}\eta_i}} \right\} g(\eta_i|\Omega) \partial\eta_i \right]$$

where  $I$  is the set of all individuals,  $J$  is the set of all alternatives, and

$$y_{ij} = \begin{cases} 1 & \text{if } i \text{ chooses } j \\ 0 & \text{otherwise} \end{cases}$$

The dimension of  $\eta$  is  $1 \times Q$ , where  $Q$  is the number of variables in  $Z$ . Thus, the log-likelihood function in Eq.(6) involves the estimation of a  $Q$ -dimensional integral.<sup>1</sup> This integral cannot be evaluated analytically since it does not have a closed-form solution. If  $Q = 1$  or  $Q = 2$  the log-likelihood can be evaluated with numerical methods such as quadrature. However, if  $Q$  is greater than two quadrature techniques cannot compute the integrals with sufficient speed and precision for maximum likelihood estimation (Hajivassiliou and Ruud 1994, Revelt and Train 1998). In this case simulation techniques must be applied to estimate the log-likelihood function.

---

<sup>1</sup>Note that this is why a “mixed probit” model (with  $g(\eta|\Omega)$  following a general distribution and  $\varepsilon$  IID standard normal) is generally regarded as impractical. Estimating this model would involve the evaluation of a  $(Q + 1)$ -dimensional integral, since the choice probabilities conditional on  $\eta$  are not closed-form as they are in MXL, but instead require the evaluation of a univariate normal density. Unless there is a strong theoretical reason to believe that the IID disturbances are normal, MXL is superior to a “mixed probit” model due to the lower dimension of integration required for estimation.

The integrals in the choice probabilities are approximated using a Monte Carlo technique, and then the resulting simulated log-likelihood function is maximized. For a given  $\Omega$  a vector of values for  $\eta$  is drawn from  $g(\eta|\Omega)$  for each individual. The values of this draw can then be used to calculate  $\hat{P}(j|\eta)$ , the conditional choice probability given in Eq. 4 of the paper. This process is repeated  $R$  times, and the integration over  $g(\eta|\Omega)$  is approximated by averaging over the  $R$  draws. Let  $\hat{P}_r(j|\eta_r)$  be the realization of the choice probability for individual  $i$  for alternative  $j$  for the  $r^{\text{th}}$  draw of  $\eta$ . The choice probabilities given the parameter vectors  $\beta$  and  $\Omega$  are approximated by averaging over the values of  $\hat{P}_r(j|\eta)$ :

$$\hat{P}(j|\beta, \Omega) = \frac{1}{R} \sum_{r=1}^R \hat{P}_r(j|\eta_r)$$

$\hat{P}(j|\beta, \Omega)$  is the simulated choice probability of individual  $i$  choosing alternative  $j$  given  $\beta$  and  $\Omega$ . This simulated choice probability is an unbiased estimator of the actual probability  $P(j)$ , with a variance that decreases as  $R$  increases. It is also twice differentiable and strictly positive for any realization of the finite  $R$  draws, which means that log-likelihood functions constructed with  $\hat{P}(j|\beta, \Omega)$  are always defined and can be maximized with conventional gradient-based optimization methods. Under weak conditions this estimator is consistent, asymptotically efficient, and asymptotically normal (Hajivassiliou and Ruud 1994; Lee 1992). When  $R$  increases faster than the square root of the number of observations, this estimator is asymptotically equivalent to the maximum likelihood estimator. However, this estimator does display some bias at low values of  $R$ , which decreases as  $R$  increases. The bias is very low when  $R = 250$  (Brownstone and Train 1999); most empirical work uses  $R$  equal to 500 or 1000.

The choice probabilities above depend on  $\beta$  and  $\Omega$ , which need to be estimated. In order

to estimate  $\Omega$  the distributions in  $\eta$  are re-expressed in terms of standardized, independent distributions. That is,  $g(\eta|\Omega)$  is re-expressed as  $\mu + Ws$ , where  $\mu = 0$  ( the mean vector of  $\eta$ ),  $W$  is the Choleski factor of  $\Omega_\eta$ , and  $s$  consists of IID deviates drawn from standardized, independent distributions. Under this specification,  $Z\eta$  becomes  $(sZ)\Omega$ , where  $\Omega = W$  and is to be estimated. A simulated log-likelihood function can then be constructed:

$$\mathcal{S}\mathcal{L}(\beta, \Omega) = \sum_{i=1}^I \sum_{j=1}^J y_{ij} \log \left[ \hat{P}_i(j|\beta, \Omega) \right]$$

The estimated parameter vectors  $\hat{\beta}$  and  $\hat{\Omega}$  are the vectors that maximize the simulated log-likelihood function.

To summarize the estimation of MXL using this Monte Carlo technique: first select starting values for  $\beta$  and  $\Omega$ , and take  $R$  random draws for each individual from  $Ws$  (where  $W = \Omega$ ). Then calculate  $\hat{P}_r(j|\eta_r)$  for each individual for each of the  $R$  draws from  $Ws$ , and from there calculate  $P(j|\beta, \Omega)$  for each individual. Insert  $P_i(j|\beta, \Omega)$  into  $\mathcal{S}\mathcal{L}(\beta, \Omega)$ , and use a standard gradient technique to determine what changes to make in  $\beta$  and  $\Omega$ . The new values of  $\beta$  and  $\Omega$  are then used to recalculate each  $\hat{P}_r(j|\eta_r)$ , and from there  $P_i(j|\beta, \Omega)$ , which is then inserted into  $\mathcal{S}\mathcal{L}(\beta, \Omega)$ , and so on until  $\beta$  and  $\Omega$  are estimated to an adequate degree of precision.

The method described above is time intensive, although no more so that the techniques generally used to estimate multinomial probit models. The Geweke-Hajivassiliou-Keane (GHK) simulator commonly used to estimate MNP models is recursive, meaning that the range for the random draw for one alternative depends on the value of previous draws for other alternatives. The probability simulator for MXL draws simultaneously for all probabilities from unrestricted ranges, speeding the process of creating the simulated log-likelihood.

Differences in speed may also arise through differences in the dimension of integration. GHK draws from a  $(J - 1)$ -dimensional distribution of utility differences, while the MXL simulator draws from a  $Q$ -dimensional mixing distribution, where  $Q$  is the number of elements in  $Z$ . If  $Q < J - 1$  mixed logit will be at a relative disadvantage in terms of speed, while if  $Q > Z$  the opposite is true. When  $Q = J - 1$  and with normally distributed error components, Brownstone and Train (1999) found that for a fixed amount of computer time, the mixed logit simulator has lower simulation variance than the GHK simulator, leading to more accurate estimates of the choice probabilities in the same amount of time.

An alternative estimation procedure proposed by Bhat (1999) and Train (1999) dramatically reduces estimation time for mixed logit models. This alternative simulation technique uses nonrandom draws from the distributions to be integrated over, rather than random draws. By drawing from a sequence designed to give fairly even coverage over the mixing distribution, many fewer draws are needed to reduce simulation variance to an acceptable level. In both Bhat (1999) and Train (1999), Halton sequences are used to create a series of draws that are distributed evenly across the domain of the distribution to be integrated.

Halton sequences are created by selecting a number  $h$  that defines the sequence (where  $h$  is a prime number) and dividing a unit interval into  $h$  equal parts.<sup>2</sup> The dividing points on this unit interval become the first  $h - 1$  elements in the Halton sequence. Each of the  $h$  subportions of the unit interval is divided as the entire unit interval was and these elements are added on to the end of the sequence. This process is continued until the desired number of elements in the sequence is reached. Halton sequences result in a far more even distribution of points across the unit interval than random draws.

---

<sup>2</sup>Prime numbers are used to define Halton sequences, since the Halton sequence for a non-prime number will divide the unit interval in the same way as the Halton sequences based on the prime numbers that constitute the non-prime number.

Halton sequences can be used in place of random draws to estimate mixed logit models. For each element of  $\eta$  a different prime number is selected, and a Halton sequence of length  $(R \times I) + 10$  is created (where  $R$  is the number of Halton draws desired for each observation and  $I$  is the number of individuals in the dataset). The first ten elements of the sequence are discarded because the first elements tend to be correlated over Halton sequences defined by different prime numbers. The first individual in the dataset is assigned the first  $R$  elements of each Halton sequence, the second individual is assigned the next  $R$  elements, and so on. For each element of each Halton sequence the inverse of the cumulative distribution for that element of  $\eta$  is calculated. The resulting values become the IID deviates  $s$  in the simulated log-likelihood function. Estimation is otherwise identical to that when using random draws to evaluate the integrals.

Both Bhat (1999) and Train (1999) found that in estimating mixed logits, the simulation error in estimated parameters is lower with 100 Halton draws than with 1000 random draws. Thus, using Halton sequences in place of random draws allows us to obtain more accurate estimates of model parameters at a fraction of the estimation cost.

For example, the model presented in Table 2 was estimated using 125 Halton draws for each of the random coefficients in order to create the simulated log-likelihood, and took 1 hour 51 minutes to converge. I estimated the same model with 1000 random draws. The results of estimating this model are presented in Table A.

**Table A here.**

A comparison of Tables 2 and A reveals that the results obtained using 125 Halton draws are substantively identical to those obtained using 1000 random draws. The time saved by using Halton draws is substantial; the model in Table A took 15 hours 52 minutes to

converge.<sup>3</sup>

Mixed logit results tend to be stable across different specifications of error components and random coefficients. A comparison of Tables 1 and 2 reveals that the fixed coefficients of the variables estimated in both models are nearly identical substantively, in sign, magnitude, and statistical significance. The same holds true when comparing the means of the random coefficients in Table 2 with the corresponding fixed coefficients in Table 1. For further comparison consider Table B, which presents an alternate mixed logit specification using the same variables and coding as were used in Tables 1 and 2. This model specifies two uniformly distributed random coefficients on the effect of home ownership on voting Conservative relative to voting Alliance, and on voting Labour relative to voting Alliance. 125 Halton draws were used in estimation.

**Table B here.**

The fixed coefficients and means of the random coefficients of this model are nearly identical to the corresponding coefficients estimated in Tables 1 and 2, despite specifying random coefficients on different variables with different distributions. Mixed logit is stable across different specification of error components and random coefficients because these random terms are estimated as additional parameters in the likelihood function which are independent of the fixed coefficients and the means of the random coefficients. In fact, the fixed parameters and mean effects in MXL should be nearly identical substantively to the coefficients estimated by the MNL derived by setting  $\eta$  to zero (all random coefficients are instead specified as fixed). This MNL is presented in Table C.

---

<sup>3</sup>Both models were estimated using a vector of zeros for the starting values, except the starting values for the distances between the means and the endpoints of the random coefficients, which were set to one. All estimation was done using Gauss 3.2.52 and Maxlik 4.0.24 on an IMB PC with a 550MHz processor and 128MB of RAM.

**Table C here.**

The coefficients of this MNL agree with the fixed coefficients and the means of the random coefficients of all three MXLs considered here in sign, magnitude, and statistical significance. Notice that the fixed coefficients and means of the random coefficients are generally slightly larger in magnitude than the corresponding coefficients in MNL. This is due to differences in scaling between MNL and MXL. In MNL, all unobserved portions of utility are absorbed into  $\varepsilon$ , which sets the scale of utility. However, in MXL some portion of the unobserved utility is estimated in  $\eta$  rather than  $\varepsilon$ . This reduces the variance of  $\varepsilon$  in MXL relative to MNL, and since utility is scaled so that  $\varepsilon$  has the variance of an extreme value distribution, this scales down the coefficients in MXL relative to MNL. This result was also noticed in Brownstone and Train (1999) and Revelt and Train (1998).

Incidentally, the mixed logit models presented in Tables 2 and A are a significantly better fit for this data than an equivalently specified MNL. A  $\chi^2$  test yielded a value of 19.89 for Table 2 and 18.76 for Table A (both statistically significant at the 99% level with 4 d.f.). The  $\chi^2$  value for Table 1 was 5.72 (statistically significant at the 90% level with 3 d.f.), and for Table B it was 0.04 (not statistically significant).

I am aware of two software packages currently available that allow for estimation of mixed logit models. All mixed logit models in this paper were estimated using GAUSS. The GAUSS code used to estimate the models in this paper is available on the Political Analysis website or at <http://www.polsci.ucsb.edu/faculty/glasgow>. This code is a modified version of the GAUSS code made available by Kenneth Train on his website at <http://elsa.berkeley.edu/users/train>. Mixed logits can also be estimated in Limdep, although the options for estimation are more limited than those in the GAUSS code — only normally or lognormally distributed random coefficients are permitted. In comparing both software

packages I found that in many applications Limdep produced results that agreed with the equivalent specification in the GAUSS code. However, in some applications Limdep failed to converge while the GAUSS code did. Overall, the GAUSS code seems more reliable than Limdep, and offers more options for estimation (more distributions are available for the random components, error-components specifications are possible, and Halton draws are available as an estimation option).

**Table A: Mixed Logit Estimates,  
1987 British General Election  
(Alliance Coefficients Normalized to Zero)  
(1000 Random Draws)**

Independent Variables	Conservatives/Alliance		Labour/Alliance	
Defense			-0.21**	(0.02)
Phillips Curve			-0.13**	(0.03)
Taxation			-0.19**	(0.03)
Nationalization			-0.21**	(0.02)
Redistribution			-0.09**	(0.02)
Crime			-0.11*	(0.05)
Welfare			-0.16**	(0.02)
Constant	0.53	(0.76)	2.83**	(0.80)
South	-0.13	(0.19)	-0.49*	(0.23)
Midlands	-0.30	(0.19)	-0.23	(0.23)
North	-0.07	(0.20)	0.71**	(0.21)
Wales	-0.59	(0.40)	1.46**	(0.34)
Scotland	-0.48*	(0.28)	0.78**	(0.28)
Public Sector Employee	0.10	(0.17)	-0.03	(0.17)
Female	0.38**	(0.16)	-0.03	(0.16)
Age	0.05	(0.05)	-0.24**	(0.06)
Home Ownership	0.54**	(0.21)	-0.59**	(0.18)
Family Income	0.08**	(0.03)	-0.07*	(0.03)
Education	-0.82**	(0.34)	-0.69*	(0.37)
Inflation	0.28**	(0.11)	-0.05	(0.13)
Taxes	0.02	(0.07)	-0.10	(0.07)
Unemployment	0.31**	(0.07)	0.01	(0.08)
Working Class (Mean)	0.09	(0.19)	0.81**	(0.18)
Working Class (Mean - Endpoint)	4.10**	(0.73)	2.45**	(0.82)
Union Member (Mean)	-0.57**	(0.19)	0.42**	(0.18)
Union Member (Mean - Endpoint)	0.31	(1.77)	0.07	(1.20)
Number of Observations	2131			
Log-Likelihood	-1468.18			

Standard errors in parentheses. \*\* indicates statistical significance at the 99% level; \* indicates statistical significance at the 95% level. Random coefficients have triangular distributions.

**Table B: Mixed Logit Estimates,  
1987 British General Election  
(Alliance Coefficients Normalized to Zero)  
(Alternate Specification, 125 Halton Draws)**

Independent Variables	Conservatives/Alliance		Labour/Alliance	
Defense			-0.18**	(0.02)
Phillips Curve			-0.11**	(0.02)
Taxation			-0.16**	(0.03)
Nationalization			-0.18**	(0.02)
Redistribution			-0.08**	(0.02)
Crime			-0.10*	(0.05)
Welfare			-0.14**	(0.02)
Constant	0.79	(0.70)		
South	-0.15	(0.17)		
Midlands	-0.30	(0.17)		
North	-0.07	(0.18)		
Wales	-0.51	(0.36)		
Scotland	-0.41	(0.26)		
Public Sector Employee	0.09	(0.15)		
Female	0.29*	(0.14)		
Age	0.02	(0.05)		
Family Income	0.07*	(0.03)		
Education	-0.81*	(0.32)		
Inflation	0.28**	(0.10)		
Taxes	0.02	(0.07)		
Unemployment	0.28**	(0.06)		
Working Class	0.11	(0.15)		
Union Member	-0.51**	(0.17)		
Home Ownership (Mean)	0.38*	(0.18)		
Home Ownership (Mean - Endpoint)	0.47	(1.28)		
Number of Observations			2131	
Log-Likelihood			-1477.54	

Standard errors in parentheses. \*\* indicates statistical significance at the 99% level; \* indicates statistical significance at the 95% level. Random coefficients have uniform distributions.

**Table C: Multinomial Logit Estimates,  
1987 British General Election  
(Alliance Coefficients Normalized to Zero)**

Independent Variables	Conservatives/Alliance		Labour/Alliance	
Defense			-0.18**	(0.02)
Phillips Curve			-0.11**	(0.02)
Taxation			-0.16**	(0.02)
Nationalization			-0.18**	(0.02)
Redistribution			-0.08**	(0.02)
Crime			-0.10*	(0.05)
Welfare			-0.14**	(0.02)
Constant	0.81	(0.69)	2.53**	(0.75)
South	-0.15	(0.17)	-0.43*	(0.21)
Midlands	-0.29*	(0.17)	-0.19	(0.20)
North	-0.06	(0.18)	0.64**	(0.19)
Wales	-0.48	(0.36)	1.35**	(0.31)
Scotland	-0.41	(0.25)	0.69**	(0.25)
Union Member	-0.50**	(0.16)	0.37*	(0.16)
Public Sector Employee	0.09	(0.15)	-0.02	(0.16)
Working Class	0.11	(0.15)	0.70**	(0.16)
Female	0.28*	(0.14)	0.00	(0.15)
Age	0.02	(0.05)	-0.22**	(0.05)
Home Ownership	0.37*	(0.18)	-0.54**	(0.16)
Family Income	0.07*	(0.03)	-0.06*	(0.03)
Education	-0.82**	(0.32)	-0.60*	(0.35)
Inflation	0.28**	(0.10)	-0.03	(0.11)
Taxes	0.01	(0.07)	-0.09	(0.07)
Unemployment	0.28**	(0.06)	0.01	(0.07)
Number of Observations	2131			
Log-Likelihood	-1477.56			

Standard errors in parentheses. \*\* indicates statistical significance at the 99% level;  
\* indicates statistical significance at the 95% level.