An Empirical Bayes Procedure for Improving Individual-Level Estimates and Predictions From Finite Mixtures of Multinomial Logit Models

Wagner A. Kamakura; Michel Wedel

The Fuqua School of Business, Duke University, Durham, NC 27708 (kamakura@duke.edu)

Michel Wedel
University of Michigan Business School, Ann Arbor, MI 48109 (Wedel@umich.edu)

Unobserved heterogeneity in random utility choice models can be dealt with by specifying either a multinomial or a normal distribution of the coefficients, leading to finite mixture logit and mixed logit models. Focusing on the former, we show that individual-level estimates and predictions of finite mixtures estimated by maximizing the likelihood function can be improved through integration over the estimation error of the hyperparameters, using an empirical Bayes approach. We investigate the conjecture that this approach is more robust against departures of the underlying assumptions of the finite mixture model in two Monte Carlo studies. We show that our approach improves the performance of the finite mixture model in representing individual-level parameters and producing hold-out forecasts. We illustrate with two examples that our approach may offer advantages in empirical applications involving the analysis of heterogeneous choice data.

KEY WORDS: Finite mixtures; Multinomial logit; Random utility models.

1. INTRODUCTION: HETEROGENEITY IN RANDOM UTILITY MODELS

One of the most salient features of consumer demand at the micro level is the large variation in tastes among consumers. Because of this heterogeneity, consumers with relatively similar observable characteristics often show very different choice behavior. Neglecting this heterogeneity leads to seemingly weak relations between explanatory variables and observed consumption, as well as to a biased assessment of consumers' response to marketing factors, such as price and sales promotions. Therefore, consumer heterogeneity in random utility models (RUMs) of consumer choice has been recognized as a topic of considerable interest in the economics literature (Revelt and Train 1998; McFadden 1989).

The multinomial logit (MNL) model (McFadden 1974) is frequently used to describe stated or revealed preference data, despite its drawbacks related to the independence of irrelevant alternatives property. Heterogeneous MNL models can be formulated as follows. Assume that a consumer \( i \) makes \( T_i \) repeated choices among \( J_i \) alternatives, with \( y_{i,j,t} = 1 \) if alternative \( j \) is chosen on occasion \( t \) and 0 otherwise, in response to a \( K \)-dimensional row vector \( X_{i,j,t} \) of predictors (including constants). The conditional distribution of the choices of subject \( i \), collected in \( y_i \), given the subject-specific preference parameter vector \( \theta_i \), is provided by

\[
\pi(y_i|X_i, \theta_i) = \frac{\exp(\sum_{j=1}^{J_i} y_{i,j,t}(X_{i,j,t}, \theta_i))}{\prod_{j} \sum_{t} \exp(X_{i,j,t}, \theta_i)}.
\]

(1)

If \( \Theta \) is a set of (hyper) parameters indexing the distribution of individual-level parameters, \( \pi(\theta_i|\Theta) \), then we have

\[
\pi(y_i|\Theta) = \int \pi(y_i|X_i, \theta_i)\pi(\theta_i|\Theta)d\theta_i.
\]

(2)

Many simple approaches to heterogeneity in economics treat heterogeneity as identifiable from the effects of exogenous variables. That approach is overly restrictive. Other early approaches treated heterogeneity as a nuisance; fixed-effects models were applied, in which individual-level constants—the parameter vector, partitioned as \( \theta_i = (\theta_{0i}, \theta_{1i}, \ldots, \theta_{Ki-1})' \)—were included in the model, which could be estimated directly. Later, conditional likelihood approaches were used in which the choice model was formulated conditional on sufficient statistics for the individual-level parameters, eliminating the individual-level constants from the models and simplifying the estimation task. Heckman and Singer (1984) first approximated the distribution of those individual-level constants by \( S \) support points and probability masses. Their specification leads to a multinomial distribution of the mixing distribution of the intercept, \( \pi(\theta_{0i} = \theta_{0s}|\Theta) = \omega_s \) for \( s = 1, \ldots, S \). This formulation was later extended to the finite mixture MNL models by Kamakura and Russell (1989).

Then economists recognized that heterogeneity is of fundamental interest by itself (for recent discussion, see, e.g., Allenby and Rossi 1999; Wansbeek, Meijer, and Wedel 2001). First, the support point approach was extended to all parameters in the choice model, which reduces (2) to a finite mixture (FM) model,

\[
\pi(y_i|X_i, \Theta) = \sum_s \omega_s \pi(y_i|X_i, \theta_{0s}).
\]

(3)

These models have received considerable attention in marketing and transportation research, because they have an elegant interpretation in terms of market segments underlying consumer choice. A technical advantage is that the integral is replaced by a sum, which makes maximum likelihood estimation straightforward. Developments on FM models have been reviewed recently by, for example, McLachlan and Peel (2000).
The assumption of a (multivariate) normal mixing distribution, \( \pi(\theta|\Theta) = \phi(\mu, \Sigma) \) leads to random coefficient logit models, or mixed logit models,

\[
\pi(y_i|X_i, \Theta) = \int \pi(y_i|X_i, \theta_i) \phi(\theta_i|\mu, \Sigma) \, d\theta_i. \tag{4}
\]

Such models have been widely applied to microeconomic problems. Although the models themselves are identical, they are often referred to as hierarchical Bayes models (Allenby and Lenk 1994) when estimated with the Markov chain Monte Carlo (MCMC) methods, and as mixed logit (Revelt and Train 1998) when estimated by simulated likelihood. In both MCMC and simulated likelihood inference, the possibly high-dimensional integrals in the likelihood are approximated through sampling. These models have been extended to include other forms of the mixing distribution, such as the multivariate t-distribution or mixtures of normal distributions.

Although current empirical evidence favors neither the continuous nor the discrete representation of heterogeneity (Andrews, Ansari, and Curin (2002)), in this article we choose to focus on the FM MNL model. In applications of this model, next to estimates for the hyperparameters, those for the individual-level parameters and predictions of new choices are often of interest. However, the individual-level estimates of the FM MNL models are constrained to lie in the convex hull of the class-level estimates, \( \theta_k \). Supposedly, this negatively affects not only recovery of individual-level parameters, but also predictive accuracy on hold-out samples (Allenby and Rossi 1999). Although both the estimates of individual-level parameters and the prediction of hold-out choices are affected by the uncertainty in the maximum likelihood estimators (MLEs) of the hyperparameters, that estimation uncertainty has typically not been taken into account in computing these quantities of interest.

In the tradition of empirical Bayes’ methods (e.g., Carlin and Louis 1998), we show that the performance of finite mixtures in recovering individual-level estimates and prediction can be improved through a simple integration method, accommodating estimation error in the parameters. We investigate the robustness of that approach in situations where the FM model is misspecified in two Monte Carlo studies, and show that it improves the performance of the FM model in representing individual-level parameters and producing hold-out forecasts. Two empirical illustrations on stated and revealed preference data, reported in the final section, corroborate this finding.

2. THE PROPOSED APPROACH AND MONTE CARLO STUDIES

2.1 Proposed Approach

When using the maximum likelihood method to obtain estimates of the parameters of the finite mixture of MNLs, one typically computes the estimates of the individual-level parameters and individual-level predictions, with so-called “plug-in” estimators, as \( \hat{\theta} = \sum_i \hat{\theta}_i, \hat{\theta}_i \) and \( \pi(y_i|X_i; \hat{\Theta}) = \sum_i \hat{\omega}_i, \pi(y_i|X_i; \hat{\theta}_i) \). Here one uses the posterior probabilities computed using Bayes’ theorem (McLachlan and Peel 2000),

\[
\hat{\omega}_i, \pi = \frac{\hat{\omega}_i, \pi(y_i|X_i; \hat{\theta}_i)}{\sum \hat{\omega}_i, \pi(y_i|X_i; \hat{\theta}_i)}. \tag{5}
\]

These posterior probabilities and individual-level estimates are thus computed in the tradition of the parametric empirical Bayes approaches (Deely and Lindley 1981; Morris 1983), by plugging the MLEs of the hyperparameters, \( \hat{\Theta} \), into the expressions for the individual-level estimates and predictions provided earlier, which we denote by \( g(\hat{\Theta}) \).

The reason why the plug-in estimates work is that \( g(\hat{\Theta}) \) can be seen to provide a first-order approximation of the posterior mean of the function \( g(\Theta) \), using Laplace’s method (cf. Tierny, Kass, and Kadane 1989; Carlin and Louis 1998, p. 147). One can approximate the posterior distribution \( \pi(g(\Theta)|y) = \int g(\Theta) \pi(\Theta|y) \, d\Theta \), using the asymptotic normal distribution of the estimates as \( \pi(g(\Theta)|y) \approx \int g(\Theta) \phi(\Theta|\hat{\Theta}), H(\hat{\Theta})) \, d\Theta \), with \( H(\hat{\Theta}) \) the expected Fisher information matrix evaluated at the MLE. Asymptotically, this normal approximation converges to the Dirac delta function on the MLE \( \hat{\Theta} \), resulting in the plug-in estimator. (Note that in the models that we investigate, two types of asymptotic approximations hold, depending on whether \( N \) or \( T \) goes to infinity.) However, the posterior distribution \( \pi(g(\Theta)|y) \) converges faster to the normal than to the Dirac delta function, so that using the full asymptotic normal approximation is expected to be more accurate for finite samples than using the MLE point estimates.

Thus we propose to integrate over the asymptotic posterior distribution of the hyperparameters in computing individual-level parameters and predictions in likelihood estimation for the FM MNL model. The procedure that we propose is a simple empirical Bayes approach (Carlin and Louis 1998). We propose using a sampling importance resampling (SIR) procedure (Rubin 1988) to approximate the integrals over the asymptotic distribution of the hyperparameters, in computing individual-level parameter estimates and predictions. That is, we use \( R \) samples, \( \theta^i, r = 1, \ldots, R \), from the approximate asymptotic distribution of the estimates, and approximate the integrals by weighted sums across the draws. We compare that method in two Monte Carlo studies to the standard FM (SFM) model.

2.2 Design of the First Monte Carlo Study

We investigate the performance of our approach (SFM) relative to the plug-in estimator (FM) under conditions that do not match the model assumptions, so that we obtain insights in their robustness under various conditions that may reflect those in empirical applications. We first present a Monte Carlo experiment in which we manipulate

- The degree of parametric heterogeneity within classes (H)
- The degree of stochasticity in choices (A)
- The number of observed choices per subject used for estimation (T)
- The sample size or number of subjects making choice decisions (N).

We consider a situation in which a sample of 100 or 500 simulated subjects make choices among sets of four choice alternatives, each defined by two attributes drawn randomly from independent uniform distributions in the \([-A, A]\) range. We assume a linear utility function for the two attributes, with coefficients \( \theta_1 \) and \( \theta_2 \) distributed as a mixture of bivariate
normals with five components \( f(\beta) = \sum_{k=1}^{5} \omega_k \phi(\theta_k | \mu_k, \Sigma_k) \), where \( \Sigma_k = \sigma^2 I \). Note that the assumption that the parameters are distributed as a mixture of normal distributions renders their distribution continuous and multimodal. The mass (\( \omega_k \)) and location (\( \mu_k \)) parameters are fixed to the following values:

\[
\begin{array}{cccc}
K & \mu_{1s} & \mu_{2s} & \omega_s \\
1 & -2.0 & 0 & .10 \\
2 & -1.0 & 1.0 & .15 \\
3 & 0 & 2.0 & .20 \\
4 & 1.0 & -1.0 & .25 \\
5 & 0 & -2.0 & .30 \\
\end{array}
\]

We manipulate the degree of preference heterogeneity (\( H \)) within each class by varying the standard deviation \( \sigma \) of the mixture components above at three levels (\( H = \sigma = 0.3, 1.0 \)). The first value corresponds to the assumptions of the finite mixture model. The degree of stochasticity in choice behavior (\( A \)) is manipulated by varying the range of the uniform distribution of the attribute values for the four choice alternatives at two levels (\( A = 1.0, 5.0, \) and \( 10.0 \)). Lower values produce more stochastic choices; higher values produce more deterministic choices. The amount of choice data available from each of the synthetic subjects is manipulated by generating \( T + 4 \) observed choices for each, but only using the first \( T \) (\( T = 5, 10, \) or \( 20 \)) observations for parameter estimation. The remaining four observations are used for hold-out tests of predictive fit. We manipulate these four factors in a \( 2 \times 3^3 \) orthogonal design. In addition, we generate four replications per cell, which gives us a reasonable number of observations from which to estimate residual variance, yet keeps the computational requirements of the Monte Carlo study within limits. To compare the performance of the approaches in recovering "true" individual-level preference parameters in the two Monte Carlo studies, we compute the mean squared error, MSE(\( \theta \)). To compare predictive accuracy, we use the individual-level estimates of the parameters to make probabilistic predictions for the four choices held out for each consumer and define the predictive fit as \( (1 - p_{ij}) \) if \( i \) alternative \( j \) was actually chosen or \( -p_{ij} \) if it was not, where \( p_{ij} \) denotes the predicted probability. We compute the predictive MSE (\( p \)), as a measure of predictive accuracy.

2.3 Number of Classes in Finite Mixture Models

For the FM models, the number of unobserved classes is unknown. Although the determination of the appropriate number of classes has not seen a completely satisfactory statistical solution, we use the Schwarz (1978) Bayesian information criterion (BIC) to select the optimal number of classes. For all datasets, we estimate the model from \( S = 1 \) to \( S = 18 \), and select that value that yields the lowest value of BIC. The results show that as within-class heterogeneity increases, more classes are needed beyond the five classes we used to generate the data, to capture the excess heterogeneity. This effect is exacerbated if choices become more deterministic and if the number of observations on a subject is larger. Figure 1 illustrates these findings.

![](image)

Figure 1. Effect of Choice Stochasticity (a), Preference Heterogeneity (b), and Number of Observations per Subject (top/bottom panel) on the Number of Latent Classes.

2.4 Results of the First Study

To compare the performance of the two approaches, we used a MANOVA with models (FM and SFM), degree of stochasticity (\( A = 1.0, 5.0, \) and \( 10.0 \)), within-class heterogeneity (\( H = 0.3, 1.0 \)), number of observations per subject (\( T = 5, 10, \) and \( 20 \)) and sample size (\( N = 100 \) and \( 500 \)) as factors and the MSFs of the coefficients as the dependent variables. For our purpose, only the model–factor interactions are relevant. It appears that the differences in parameter recovery across the methods depend on the other factors and their interactions. (A MANOVA on the log-transformed MSFs yields similar results.) As one would expect, the SFM estimates are more accurate than those obtained with the FM approach, because they take estimation error into account. As shown in Table 1, the accuracy of the proposed empirical Bayes approach over the standard one increases as the choice behavior becomes more deterministic (larger values of \( A \)), and when there is a smaller number of observations per subject available for estimation (\( T = 5 \)). These results show that taking hyperparameter uncertainty into account with our proposed SFM procedure leads to better individual-level estimates than the FM model approach, in all conditions investigated in this study.

Predictive fit has been the most popular criterion in the choice modeling literature for comparing the performance of choice models, probably because predictive fit gives an intuitive sense of the advantages of one model over others, in explaining and forecasting choice behavior. In the ANOVA using the
same treatment factors as before, and predictive fit as the dependent variable, one finds statistically significant ($p < .001$) differences in predictive performance among the approaches depending on heterogeneity (H) and choice stochasticity (A). ANOVA on the log-transformed MSES yields similar results. A look at the predictive MSES in Table 2 clearly shows that the SFM approach outperforms the FM in predictive fit, even though the differences in MSE($p$) are modest.

Overall, the results from this Monte Carlo experiment show that by considering estimation errors when computing the individual-level estimates and predictions, we can obtain more accurate individual-level estimates and predictions than with the SFM, which holds under all conditions investigated in this study.

2.5 Second Monte Carlo Study

To investigate the performance of these models in situations that more strongly violate their assumptions, we conducted another Monte Carlo experiment similar to the first one. Here the $\theta_1$ and $\theta_2$ parameters for $N = 500$ and $N = 100$ synthetic choice makers were drawn from independent uniform, exponential, and Cauchy distributions across three levels of choice stochasticity ($A = 1, 5, \text{or } 10$) and two different numbers of estimation observations for each individual ($T = 5 \text{ or } 10$). Parameter recovery for the Cauchy distribution appears to be extremely poor, due to the long “tails” of this distribution, which generates extreme values that cannot be accommodated very well by the FM model or its SIR extension. Using a MANOVA, we found no statistically significant differences between the models both in terms of parameter recovery across all treatment factors and the three distributions. A possible explanation is that severe misspecification of the heterogeneity distribution in RUM models biases the estimates to such an extent that all individual-level estimates are far off.

We did find statistically significant differences in predictive fit between the two approaches; these are shown in Table 3. The table shows that there are no very large differences in terms of predictive fit when the true distribution is continuous, but violates the assumptions of the FM model, although our SFM is slightly better. (These results still hold when the results for the Cauchy distribution are omitted from the analyses.)

### 3. COMPARISONS ON EMPIRICAL DATA

To further illustrate our findings in empirical choice model applications, we provide two empirical comparisons, for stated and revealed preference data. We compare the FM and SFM approaches on predictive fit.

#### 3.1 Stated Preference Data

For the first comparison we use data from a commercial choice-based conjoint experiment, involving a sample of 535 consumers. In a questionnaire, each consumer was asked to choose one high-tech electronic product prototype, blocked in 20 sets of four choice alternatives. These 20 choice sets were designed by combining six features of the product according to an optimal orthogonal array. (For confidentiality reasons, neither the product nor its features can be disclosed.) Two of the product features were continuous, whereas the other four were nominal and were effects-type coded relative to the first attribute level. Based on BIC model selection, we fit a six-class FM logit model to the first 19 choices made by each of the 535 consumers. We compare the FM and SFM approaches in terms of fit to the 19 choices per subject used during estimation and predictive fit to the one observed choice held out for out-of-sample predictions. Using the same MSE($p$) measure defined earlier for the calibration sample, we find that the SFM approach provides better in sample fit [MSE($p$) = .126] than the FM [MSE($p$) = .130] approach. The out of sample predictive performance of SFM [MSE($p$) = .143], is also better than FM [MSE($p$) = .148]. Thus we find that the proposed empirical Bayes procedure outperforms the standard procedure in terms of prediction in this application.

#### 3.2 Revealed Choice Data

In this second comparison, we apply the approaches to revealed choice data in a classic scanner panel dataset on purchases of ketchup observed in a sample of 600 households belonging to A. C. Nielsen’s single-source scanner panel. Each of
these 600 households have made between 3 and 36 purchases of ketchup during the sampling period. Information on the net prices and store display (0/1) status for each of eight brand-size combinations is available for each purchase occasion. We estimate the FM model using all of the observed choices except the final one, which is held out for out-of-sample predictive testing. Based on the BIC, we fit a 10-class FM logit model. We use both the plug-in and SIR empirical Bayes procedures to compute individual-level predictions. In terms of fit to the observed choices, SFM [MSE(\(\hat{p}\)) = .064], does substantially better than FM [MSE(\(\hat{p}\)) = .077], whereas similar results are found in terms of predictive fit, with the SFM approach producing better predictions [MSE(\(\hat{p}\)) = .081] than the standard FM approach [MSE(\(\hat{p}\)) = .084].

4. CONCLUSION

Heterogeneity of consumer preferences is of significant economic concern, and RUMs that accommodate distributions of preference parameters enable one to investigate differences in those parameters across subjects in a sample, as well as their effects on prediction of stated or revealed choice behavior. Because the true form of the underlying preference distribution is unknown in most applications, a researcher needs to assume a distribution of the preference coefficients, and which distribution most accurately represents the data becomes an empirical issue. In this article we have shown that when estimating the FM model in a likelihood framework, the estimation of individual-level coefficients, as well as the prediction of individual choices, is much improved by integrating out the asymptotic distribution of the estimates, using an SIR approach, in the tradition of empirical Bayes procedures. In addition, both of our Monte Carlo studies have shown that when the mixture model is misspecified, the SIR augmentation of the FM model improves its predictive performance. The SIR approach is easy to implement, at hardly any additional computational cost. All computations are closed form, and no iterations are required. The ease of implementation makes this approach even more attractive for practical applications of FM logit models, as we have illustrated in two examples.

ACKNOWLEDGMENT

The authors are very grateful to the associate editor and two anonymous reviewers for many helpful suggestions.

REFERENCES


