Estimating Mixtures of Discrete Choice Model

Paul A. Ruud
University of California, Berkeley

September 20, 2007

In this note, we take up a computational problem observed with fitting such mixtures of discrete choice models as the mixed multinomial logit, the parameter values explode as a numerical optimization algorithm maximizes the logarithm of the simulated likelihood function. We describe two identification issues that can increase the probability of this phenomenon. First, the parameters of the variance-covariance matrix of differences in the latent utility indexes may not be identified. Second, that variance-covariance matrix may be singular.

1 The Mixed Multinomial Logit Model

A leading example of a mixture of discrete choice models is the mixed multinomial logit (MMNL). McFadden and Train define an MMNL model as a multinomial logit model with random coefficients. The probability that alternative j is chosen is

$$\begin{aligned}\Pr\{j \mid \mathbf{X}, \boldsymbol{\beta}\} &= \frac{\exp\left(\mathbf{x}_j'\boldsymbol{\beta}\right)}{\sum_{k=0}^{J}\exp\left(\mathbf{x}_k'\boldsymbol{\beta}\right)}, \qquad j = 0, 1, \ldots, J \\ \Pr\{\boldsymbol{\beta} \leq \mathbf{b} \mid \mathbf{X}\} &\sim G(\mathbf{b}; \boldsymbol{\theta}),\end{aligned}$$

for a choice set with $J+1$ alternatives indexed by j, observable variables $\mathbf{X} = \left[\mathbf{x}_j';\ j = 0, \ldots, J\right]$, and c.d.f. $G(\mathbf{b}; \boldsymbol{\theta})$. Alternatively,

$$\Pr\{j \mid \mathbf{X}, \boldsymbol{\theta}\} = \int \frac{\exp\left(\mathbf{x}_j'\mathbf{b}\right)}{\sum_{k=0}^{J}\exp\left(\mathbf{x}_k'\mathbf{b}\right)} dG(\mathbf{b}; \boldsymbol{\theta}).$$

Let K denote the dimension of the column vectors $\mathbf{x}_j$ and $\boldsymbol{\beta}$. Let $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^M$.

A leading specification sets $G(\mathbf{b}; \boldsymbol{\theta})$ to a multivariate normal distribution and implementation uses the method of maximum simulated likelihood (MSL). For the multivariate normal specification, the simulated probabilities are often

computed as

$$
\begin{aligned}
P(j, \boldsymbol{\theta}; \mathbf{X}, \omega) &= \mathrm{E}_R \left[\frac{\exp\left(\mathbf{x}_j' \boldsymbol{\beta}(\boldsymbol{\theta}; \omega_r)\right)}{\sum_{k=0}^{J} \exp\left(\mathbf{x}_k' \boldsymbol{\beta}(\boldsymbol{\theta}; \omega_r)\right)} \right] \\
&\equiv \sum_{r=1}^{R} \frac{\exp\left(\mathbf{x}_j' \boldsymbol{\beta}(\boldsymbol{\theta}; \omega_r)\right)}{\sum_{k=0}^{J} \exp\left(\mathbf{x}_k' \boldsymbol{\beta}(\boldsymbol{\theta}; \omega_r)\right)} \frac{1}{R}, \\
\boldsymbol{\beta}(\boldsymbol{\theta}; \omega_r) &= \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2} \omega_r
\end{aligned}
$$

where $\omega = [\omega_r;\ r = 1, \ldots, R]$, $\omega_r \overset{i.i.d.}{\sim} N(\mathbf{0},\ \mathbf{I}_K)$, and $\boldsymbol{\Sigma}^{1/2}$ is a matrix square root of a variance-covariance matrix $\boldsymbol{\Sigma}$. In this case, for $s \neq 1$ and $\mathbf{b} \neq \mathbf{0}$,

$$
\frac{\exp\left(s \cdot \mathbf{x}_j' \mathbf{b}\right)}{\sum_{k=0}^{J} \exp\left(s \cdot \mathbf{x}_k' \mathbf{b}\right)} \neq \frac{\exp\left(\mathbf{x}_j' \mathbf{b}\right)}{\sum_{k=0}^{J} \exp\left(\mathbf{x}_k' \mathbf{b}\right)}
$$

with nonzero probability and no scale normalization is required; both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are identified. The scale normalization appears implicitly in the multinomial logit part of the specification.

Occasionally, one finds references to computational problems implementing the MMNL. The fitted coefficients explode before numerical convergence of the computational algorithm. Simulations show that even a pure MNL data generating process, where $\boldsymbol{\Sigma} = \mathbf{0}$, generates such problems. Paradoxically, this phenomenon may become more probable as one increases R, the number of simulations for each observation. One might expect that as the variance in the likelihood simulator falls such problems would become less likely.

One cause of this phenomenon is the mixture specification itself. If any component distribution in the mixture can have no weight, an estimator of the mixture generally has nonzero probability of assigning no weight to that component. Both the multivariate normal component and the multivariate logistic component can have zero probability without making the choice probabilities zero. In this parameterization, the normal component contributes nothing to the mixture if $\boldsymbol{\Sigma} = \mathbf{0}$. The logistic component contributes nothing if its scale relative to the normal component is zero, which occurs when $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}^{1/2}$ are made infinite while keeping the relative magnitudes of these parameters fixed. We will offer a reparameterization to address this computational problem below.

Note also that for some data sets the probability that MMNL computations will seek a parameter boundary increases with the number of simulations per observation. This occurs because at the boundary the MMNL fitted probabilities become crude frequency simulations. A finite number of simulations R permits a crude frequency probability simulator to place the lower bound $1/R$ on all fitted probabilities. Without some fitted probabilities approaching zero, poorly predicted observations do not penalize the quasi-log likelihood function so as to steer clear of the parameter boundary. The larger the number of simulations the greater the probability that this will occur.

2 The MMNL Approximation Property

McFadden and Train say "Under mild regularity conditions, any discrete choice model derived from random utility maximization has choice probabilities that can be approximated as closely as one pleases by a MMNL model." This involves approximating both a utility function and a distribution function. This approximation property is frequently cited as a motivation for applying the MMNL model.[1]

To focus on the approximation of choice probabilities and the role played by the multinomial logistic distribution, suppose that the functional form of the utility functions are known to be $\mathbf{X}\boldsymbol{\beta}_0$. The approximation property applies to this special case. Moreover, the McFadden/Train conditions imply an exact result: the choice probabilities may be identically *equal* to those of an MMNL model. Train (2003) makes a similar point. To see this, note that

$$\frac{\exp(\sigma a_j)}{\sum_{k=0}^{J}\exp(\sigma a_k)} \to \mathbf{1}\{a_j \geq a_k,\ k=0,\ldots,J\}$$

monotonically as $\sigma \to \infty$. Therefore, the monotone convergence theorem implies

$$\begin{aligned}\lim_{\sigma\to\infty} \mathrm{E}\left[\Pr\{y_j = 1 \,|\, \mathbf{X},\ \sigma\cdot\boldsymbol{\beta}\} \,|\, \mathbf{X}\right] &= \mathrm{E}\left[\lim_{\sigma\to\infty} \Pr\{y_j = 1 \,|\, \mathbf{X},\ \sigma\cdot\boldsymbol{\beta}\} \,|\, \mathbf{X}\right] \\ &= \Pr\left\{\mathbf{x}_j'\boldsymbol{\beta} \geq \mathbf{x}_k'\boldsymbol{\beta},\ k=0,\ldots,J \,|\, \mathbf{X}\right\},\end{aligned}$$

which is the choice probability function of the population RUM. In the limit, the scale of the multinomial logistic component vanishes relative to the rest of the MMNL utility specification, and an exact model is achieved by choosing the population distribution of the $\boldsymbol{\beta}$.

This approximation property of the MMNL is shared by many other tractable specifications. McFadden (1989) originally proposed several "smoothed frequency simulators" that could be treated as similar approximations. Generalizing from MMNL, an example is distributions with independent and identically distributed ε_j, $j=0,\ldots,J$, and tractable marginal c.d.f.s $F(a) = \Pr\{\varepsilon_j \leq a\}$. In this case,

$$\Pr\{y_j = 1 \,|\, \mathbf{X}, \boldsymbol{\beta}\} \neq \mathrm{E}\left[\prod_{k\neq j} F\left(0 \leq \mathbf{x}_j'\boldsymbol{\beta} - \mathbf{x}_k'\boldsymbol{\beta} + \varepsilon_j\right) \,|\, \mathbf{X}, \boldsymbol{\beta}\right]$$

[1] For example, Hensher and Greene (2003, fn. 7) say

> The proof in McFadden and Train (2001) that mixed logit can approximate any choice model, including any multinomial probit model is an important message. The reverse cannot be said: a multinomial probit model cannot approximate any mixed logit model, since multinomial probit relies critically on normal distributions. If a random term in utility is not normal, then mixed logit can handle it and multinomial probit cannot. Apart from this point, the difference between the models is a matter of which is easier to use in a given situation.

because

$$\begin{aligned}
&\lim_{\sigma\to\infty} \mathrm{E}\left[\prod_{k\neq j} F\left(0 \leq \sigma\cdot\left(\mathbf{x}_j'\boldsymbol{\beta}-\mathbf{x}_k'\boldsymbol{\beta}\right)+\varepsilon_j\right) \mid \mathbf{X},\boldsymbol{\beta}\right] \\
&\quad = \prod_{k\neq j} \mathbf{1}\left\{\mathbf{x}_j'\boldsymbol{\beta}\geq\mathbf{x}_k'\boldsymbol{\beta}\right\} \\
&\quad = \mathbf{1}\left\{\mathbf{x}_j'\boldsymbol{\beta}\geq\mathbf{x}_k'\boldsymbol{\beta},\ k=0,\ldots,J\right\}
\end{aligned}$$

The integration required over ε_j may not be tractable, but this is easily accommodated by including ε_j in the $\mathbf{x}_j'\boldsymbol{\beta}$ component that is simulated in the typical feasible implementation of MMNL. A potential drawback is that the simulated probabilities will not sum to one across alternatives.

Alternatives to the extreme value distribution posited by MMNL include the univariate Cauchy, logistic, and normal distributions. The normal distribution is a leading member because it also produces the multinomial probit model. Like the MMNL, the normal does not require simulation with respect to ε_j. Conditional on ε_j and $\boldsymbol{\beta}$,

$$\Pr\left\{y_j=1 \mid \mathbf{X},\boldsymbol{\beta},\varepsilon_j\right\} = \prod_{k\neq j}\Phi\left(\left(\mathbf{x}_k-\mathbf{x}_j\right)'\boldsymbol{\beta}-\varepsilon_j\right)$$

Using Gauss-Hermite quadrature, one can approximate the equicorrelated MNP integral,

$$\Pr\left\{y_j=1 \mid \mathbf{X},\boldsymbol{\beta}\right\} = \mathrm{E}\left[\prod_{k\neq j}\Phi\left(\left(\mathbf{x}_k-\mathbf{x}_j\right)'\boldsymbol{\beta}-\varepsilon_j\right) \mid \mathbf{X},\boldsymbol{\beta}\right]. \tag{1}$$

Such approximation has become the basis for routine estimation (in Stata) of equicorrelated probit models for panel data. These simulated probabilities will sum to one.

3 Multinomial Probit

The multinomial probit model is one case that motivated the MMNL approximation. Focusing on identification for this special case is helpful for the application of MMNL. Let $\mathbf{X}=[\mathbf{x}_k']$ denote the $(J+1)\times K$ matrix of alternative characteristics. Let the $\varepsilon_j \sim N(0,1)$ independently. For convenience, we isolate the conditionally homoscedastic part of the $\mathbf{X}\boldsymbol{\beta}$ by decomposing it into $\mathbf{X}_1\boldsymbol{\beta}_1+\mathbf{X}_2\boldsymbol{\beta}_2$ where $\mathbf{X}_2$ spans the space of alternative specific dummy variables. Let the matrix $\boldsymbol{\Omega}$ denote

$$\boldsymbol{\Omega} = \mathrm{Var}\left[\mathbf{X}_2\boldsymbol{\beta}_2 \mid \mathbf{X}_1\boldsymbol{\beta}_1\right].$$

An identifiable and unrestricted parameterization of $\boldsymbol{\Omega}$ is to set the first row and column to zeros and to scale the remaining $J\times J$ submatrix, denoted $\boldsymbol{\Omega}_2=[\omega_{ij};i,j=1,\ldots,J]$. We specify $\Sigma_{j=1}^{J}\omega_{jj}=1$.

With this specification, the probit kernel above corresponds to decomposing

$$\mathbf{\Omega}_2 = (\mathbf{\Omega}_2 - \alpha \cdot (\mathbf{I}_J + \mathbf{J}_J)) + \alpha \cdot (\mathbf{I}_J + \mathbf{J}_J)$$

for some scalar $\alpha > 0$, where $\mathbf{I}_J$ is a $J \times J$ identity matrix and $\mathbf{J}_J$ is a $J \times J$ matrix of ones. The matrix $\alpha \cdot (\mathbf{I}_J + \mathbf{J}_J)$ is the familiar equicorrelated variance-covariance matrix and (1) corresponds to integrating over $\boldsymbol{\beta}_2$ conditional on $\boldsymbol{\beta}_1$. The α rescales the ε distribution so that the $\mathbf{\Omega}_2 - \alpha \cdot (\mathbf{I}_J + \mathbf{J}_J)$ component remains positive semi-definite. If the population $\mathbf{\Omega}$ were a scalar matrix, then $\alpha \cdot (\mathbf{I}_J + \mathbf{J}_J)$ would be the conditional variance-covariance matrix of the utility differences that determine a choice probability.

Given this decomposition, motivated by a mixed MNP specification, identification requires a normalization for α. A convenient choice is to take α as the smallest eigenvalue of the generalized eigenvalue problem

$$|\mathbf{\Omega}_2 - \alpha \cdot (\mathbf{I}_J + \mathbf{J}_J)| = 0.$$

This eigenvalue is always positive, because $\mathbf{\Omega}_2$ is positive semi-definite and $\mathbf{I}_J + \mathbf{J}_J$ is positive definite; α is strictly greater than zero if $\mathbf{\Omega}_2$ is nonsingular. The remainder $\mathbf{\Omega}_2 - \alpha \cdot (\mathbf{I}_J + \mathbf{J}_J)$ is a positive semi-definite matrix with rank less than or equal $J-1$. It is common to parameterize this in terms of a lower-triangular matrix square root. However, this specification requires that this square root is singular.

It is not necessary to solve the eigenvalue problem in implementation. Simply solve

$$1 = \sum_{j,k} c_{jk}^2 + J \; \alpha(\mathbf{C}) \qquad \Longleftrightarrow \qquad \alpha(\mathbf{C}) = \frac{1 - \sum_{j,k} c_{jk}^2}{J}$$

so that the normalization of the variance-covariance matrix corresponds to setting a diagonal element of a Cholesky factor equal to zero. $\alpha > 0$ is enforced by the functional form

$$\begin{aligned} &\Pr\{y_j = 1 \,|\, \mathbf{X}, \boldsymbol{\beta}\} \\ &\quad = \mathrm{E}\left[\prod_{k \neq j} \Phi\left(\frac{(\mathbf{x}_{1k} - \mathbf{x}_{1j})' \boldsymbol{\beta}_1 + \Delta_{kj}\mathbf{C}\boldsymbol{\omega} - \sqrt{\alpha(\mathbf{C})}\varepsilon_j}{\sqrt{\alpha(\mathbf{C})}}\right) \,|\, \mathbf{X}_1\boldsymbol{\beta}_1\right]. \end{aligned}$$

where $\mathbf{C} = [c_{ij}]$ is a $J \times J$ lower triangular matrix with $c_{JJ} = 0$ and

$$\mathbf{C}\mathbf{C}' = \mathbf{\Omega}_2 - \alpha(\mathbf{C}) \cdot (\mathbf{I}_J + \mathbf{J}_J)\,.$$

Given that a MMNL with multivariate normal $\boldsymbol{\beta}$ is observationally similar to the MMNP, a similar reparameterization would be sensible there. This would prevent parameters from running to the parameter boundary when the estimates put no probability on the logit component because in that case $\alpha(\mathbf{C})$ approaches zero instead. Because the logistic and normal probability functions are similar, we expect the differences between MMNP and MMNL estimates

to be inconsequential for statistical inference. If, however, the logit component is considered structural, and not merely convenient, then restricting $c_{JJ} = 0$ is inappropriate and c_{JJ} should be estimated along with the other unknown parameters.

4 Singular Variance-Covariance Matrix

There is another reason why estimated MMNL parameters may land on a boundary of the parameter space: the estimator for $\mathbf{\Omega}$ may be singular. That this is a possibility is apparently not widely known. Though it may not be expected *a priori*, if $\mathbf{\Omega}$ is singular the choice probabilities are still well-defined. Because $\mathbf{I}_J + \mathbf{J}_J$ is nonsingular, the parameterization above restricts $\mathbf{\Omega}$ to be definite except when $\alpha(\mathbf{C}) = 0$ or, under the original parameterization, the coefficients in the latent conditional regression functions go to infinity.

Note that a small number of replications will often mask this problem as well, so that as the number of simulations is increased exploding MMNL parameter estimators become more likely. The mixed logit model will yield estimates of unidentified parameters. One way to see or understand that this is possible is to consider the mixed logit model estimator when there is only one simulation per observation. In this case, the quasi-MLE of an over-parameterized mixing distribution is easily computed because the model is the familiar conditional logit model.

A singular multivariate distribution for the latent utilities of the discrete choice model presents new problems for statistical inference. Such alternative approaches to inference with simulation as the Gibbs sampler and the GHK simulator cannot accommodate this possibility in their current forms. The mixture model, appropriately parameterized, can but discontinuities in the estimator objective function or its derivatives repose old computational issues.

5 Bibliography

Ben-Akiva, Moshe, Denis Bolduc, and Joan Walker (2001), "Specification, Identification, & Estimation of the Logit Kernel (or Continuous Mixed Logit) Model," working paper.

Hensher, David A. and William H. Greene (2003), "The Mixed Logit model: The state of practice," *Transportation* 30(2), 133–176.

McFadden, Daniel (1989), "A Method of Simulated Moments for Estimation of Discrete Response Models," *Econometrica* 57, 995–1026.

McFadden, Daniel and Kenneth Train (2000), "Mixed MNL Models of Discrete Response," *Journal of Applied Econometrics* 15, 447–470.

Train, Kenneth (2003), *Discrete Choice Methods with Simulation,* Cambridge Univ. Press.