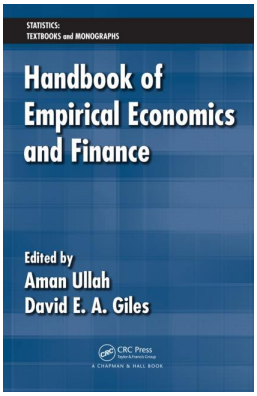

Handbook of Empirical Economics and Finance

Ullah Aman, E. A. Giles David

An Information Theoretic Estimator for the Mixed Discrete Choice Model

Publication details
https://test.routledgehandbooks.com/doi/10.1201/b10440-4
Golan Amos, H. Greene William
Published online on: 20 Dec 2010

3

An Information Theoretic Estimator for the Mixed Discrete Choice Model

Amos Golan and William H. Greene

CONTENTS

3.1 Introduction

There is much work in the social sciences on discrete choice models. Among those, the multinomial logit is the most common model used for analyzing survey data when the number of choices is greater than 2. In many cases, however, the underlying assumptions leading to the traditional maximum likelihood estimator (MLE) for the logit model are inconsistent with the perceived process that generated the observed data. One of these cases is the random parameter (RP) logit model (also known as "mixed logit" – see Revelt and Train 1998) that can be viewed as a variant of the multinomial choice model. In this chapter we formulate an Information-Theoretic (IT) estimator for the RP mixed logit model. Our estimator is easy to use and is computationally much less demanding than its competitors — the simulated likelihood class of estimators.

The objective of this work is to develop an estimation approach that is not simulation based, does not build on an underlying normal structure and is computationally efficient. This method is simple to use and apply and it works well for all sample sizes, though it is especially useful for smaller or ill-behaved data. The random parameters logit model is presented in Section 3.2. We discuss the information theoretic model and the motivation for constructing it in Section 3.3. In Section 3.4 we extend and generalize our basic model. In Section 3.5 we provide the necessary statistics for diagnostics and inference. We note, however, that the emphasis in this chapter is not on providing the large sample properties of our estimator, but rather to present the reader with convincing arguments that this model works well (relative to its competitors) for finite data and to provide the user with the correct set of tools to apply it. In Section 3.6 we provide simulated examples and contrast our IT estimator with competing simulated methods. We conclude in Section 3.7.

3.2 The Random Parameters Logit

The RP model is somewhat similar to the random coefficients model for linear regressions. (See, for example, Bhat 1996; Jain, Vilcassim, and Chintagunta 1994; Revelt and Train 1998; Train, Revelt, and Ruud, 1996; and Berry, Levinsohn, and Pakes 1995.) The core model formulation is a multinomial logit model, for individuals $i = 1, \ldots, N$ in choice setting t. Let y_{it} be the observed choice ($t = j$) of individual i and neglecting for the moment the error components aspect of the model, we begin with the basic form of the multinomial logit model, with alternative specific constants α_{ji} and a K-dimensional attributes vector $\mathbf{x}_{jit}$,

$$\text{Prob}(y_{it} = j) = \frac{\exp(\alpha_{ji} + \boldsymbol{\beta}_i' \mathbf{x}_{jit})}{\sum_{q=1}^{J_i} \exp(\alpha_{qi} + \boldsymbol{\beta}_i' \mathbf{x}_{qit})}. \tag{3.1}$$

The RP model emerges as the form of the individual specific parameter vector, $\boldsymbol{\beta}_i$ is developed. In the most familiar, simplest version of the model

$$\beta_{ki} = \beta_k + \sigma_k v_{ki},$$

and

$$\alpha_{ji} = \alpha j + \sigma_j v_{ji},$$

where β_k is the population mean, v_{ki} is the individual specific heterogeneity, with mean zero and standard deviation one, and σ_k is the standard deviation of the distribution of β_{ik}'s around β_k. The choice specific constants, α_{ji} and the elements of $\boldsymbol{\beta}_i$ are distributed randomly across individuals with fixed means. A refinement of the model is to allow the means of the parameter distributions to be heterogeneous with observed data, $\mathbf{z}_i$, (which does not include a constant term). This would be a set of choice invariant characteristics that

produce individual heterogeneity in the means of the randomly distributed coefficients, so that

$$\beta_{ki} = \beta_k + \boldsymbol{\delta}_k' \mathbf{z}_i + \sigma_k v_{ki},$$

and likewise for the constants. The basic model for heterogeneity is not limited to the normal distribution. We consider several alternatives below. One important variation is the lognormal model,

$$\beta_{ki} = \exp(\phi_k + \boldsymbol{\delta}_k' \mathbf{z}_i + \sigma_k v_{ki}).$$

The *v's* are individual unobserved random disturbances, the source of the heterogeneity. Thus, as stated above, in the population, if the random terms are normally distributed, then

$$\alpha_{ji} \text{ or } \beta_{ki} \sim \text{ Normal or Lognormal } [\phi_{j\,or\,k} + \boldsymbol{\delta}'_{j\,or\,k} \mathbf{z}_i, \sigma^2_{j\,or\,k}].$$

Other distributions may be specified in a similar fashion.

For the full vector of *K* random coefficients in the model, we may write the set of random parameters as

$$\boldsymbol{\phi}_i = \boldsymbol{\phi} + \boldsymbol{\Delta} z_i + \boldsymbol{\Gamma} \mathbf{v}_i. \tag{3.2}$$

where $\boldsymbol{\Gamma}$ is a diagonal matrix which contains σ_k on its diagonal. For convenience at this point, we will simply gather the parameters, choice specific or not, under the subscript '*k*.' (The notation is a bit more cumbersome for the lognormally distributed parameters.)

We can go a step farther and allow the random parameters to be correlated. All that is needed to obtain this additional generality is to allow $\boldsymbol{\Gamma}$ to be a lower triangular matrix with nonzero elements below the main diagonal. Then, the full covariance matrix of the random coefficients is $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}'$. The standard case of uncorrelated coefficients has $\boldsymbol{\Gamma} = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_K)$. If the coefficients are freely correlated, $\boldsymbol{\Gamma}$ is an unrestricted lower triangular matrix and $\boldsymbol{\Sigma}$ will have nonzero off diagonal elements. It is convenient to aggregate this one step farther. We may gather the entire parameter vector for the model in this formulation simply by specifying that for the nonrandom parameters in the model, the corresponding rows in $\boldsymbol{\Delta}$ and $\boldsymbol{\Gamma}$ are zero. We also define the data and parameter vector so that any choice specific aspects are handled by appropriate placements of zeros in the applicable parameter vector. This is the approach we take in Section 3.3.

An additional extension of the model allows the distribution of the random parameters to be heteroscedastic. As stated above, the variance of v_{ik} is taken to be a constant. The model is made heteroscedastic by assuming, instead, that

$$\text{Var}[v_{ik}] = \sigma jk^2 [\exp(\boldsymbol{\omega}_k \,'\mathbf{hr}_i)]^2$$

where $\mathbf{hr}_i$ is a vector of covariates, and $\boldsymbol{\omega}_k$ is the associated set of parameters. A convenient way to parameterize this is to write the full model (Equation 3.2) as

$$\boldsymbol{\phi}_i = \boldsymbol{\phi} + \boldsymbol{\Delta} \mathbf{z}_i + \boldsymbol{\Gamma} \boldsymbol{\Omega}_i \, \mathbf{v}_i \tag{3.3}$$

where $\mathbf{\Omega}_i$ is a diagonal matrix of individual specific standard deviation terms: $\omega_{ik} = \exp(\boldsymbol{\omega}_k' \mathbf{hr}_i)$.

The list of variations above produces an extremely flexible, general model. Typically, depending on the problem at hand, we use only some of these variations, though in principle, all could appear in the model at once. The probabilities defined above (Equation 3.1) are conditioned on the random terms, $\mathbf{v}_i$. The unconditional probabilities are obtained by integrating v_{ik} out of the conditional probabilities: $P_j = E_{\mathbf{v}}[P(j|\mathbf{v}_i)]$. This is a multiple integral which does not exist in closed form. Therefore, in these types of problems, the integral is approximated by sampling R draws from the assumed populations and averaging. The parameters are estimated by maximizing the simulated log-likelihood,

$$\log L_s = \sum_{i=1}^{N} \log \frac{1}{R} \sum_{r=1}^{R} \prod_{t=1}^{T_i} \sum_{j=1}^{J_{it}} d_{ijt} \frac{\exp[\alpha_{ji} + \boldsymbol{\beta}_{ir}' \mathbf{x}_{jit}]}{\sum_{q=1}^{J_{it}} \exp[\alpha_{qi} + \boldsymbol{\beta}_{ir}' \mathbf{x}_{qit}]}, \tag{3.4}$$

with respect to ($\boldsymbol{\beta}$, $\boldsymbol{\Delta}$, $\boldsymbol{\Gamma}$,$\boldsymbol{\Omega}$), where

$d_{ijt} = 1$ if individual i makes choice j in period t, and zero otherwise,
$R =$ the number of replications,
$\boldsymbol{\beta}_{ir} = \boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i + \boldsymbol{\Gamma}\boldsymbol{\Omega}_i\mathbf{v}_{ir} =$ the rth draw on $\boldsymbol{\beta}_i$,
$\mathbf{v}_{ir} =$ the rth multivariate draw for individual i.

The heteroscedasticity is induced first by multiplying $\mathbf{v}_{ir}$ by $\mathbf{\Omega}_i$, then the correlation is induced by multiplying $\mathbf{\Omega}_i\mathbf{v}_{ir}$ by $\boldsymbol{\Gamma}$. See Bhat (1996), Revelt and Train (1998), Train (2003), Greene (2008), Hensher and Greene (2003), and Hensher, Greene, and Rose (2006) for further formulations, discussions and examples.

3.3 The Basic Information Theoretic Model

Like the basic logit models, the basic mixed logit model discussed above (Equation 3.1) is based on the utility functions of the individuals. However, in the mixed logit (or RP) models in Equation 3.1, there are many more parameters to estimate than there are data points in the sample. In fact, the construction of the simulated likelihood (Equation 3.4) is based on a set of restricting assumptions. Without these assumptions (on the parameters and on the underlying error structure), the number of unknowns is larger than the number of data points regardless of the sample size leading to an underdetermined problem. Rather than using a structural approach to overcome the identification problem, we resort here to the basics of information theory (IT) and the method of Maximum Entropy (ME) (see Shannon 1948; Jaynes 1957a, 1957b). Under that approach, we can maximize the total entropy of the system subject to the observed data. All the observed and known information enters as constraints within that optimization. Once the optimization is done, the problem is converted to its concentrated form (profile likelihood), allowing

us to identify the natural set of parameters of that model. We now formulate our IT model.

The model we develop here is a direct extension of the IT, generalized maximum entropy (GME) multinomial choice model of Golan, Judge, and Miller (1996) and Golan, Judge, and Perloff (1996). To simplify notations, in the formulation below we include all unknown signal parameters (the constants and choice specific covariates) within $\boldsymbol{\beta}$ so that the covariates $\mathbf{X}$ also include the choice specific constants. Specifically, and as we discussed in Section 3.2, we gather the entire parameter vector for the model by specifying that for the nonrandom parameters in the model, the corresponding rows in $\boldsymbol{\Delta}$ and $\boldsymbol{\Gamma}$ are zero. Further, we also define the data and parameter vector so that any choice specific aspects are handled by appropriate placements of zeros in the applicable parameter vector. This is the approach we take below.

Instead of considering a specific (and usually unknown) $F(\cdot)$, or a likelihood function, we express the observed data and their relationship to the unobserved probabilities, P, as

$$y_{ij} = F(\mathrm{x}'_{ji}\boldsymbol{\beta}_j) + \varepsilon_{ij} = p_{ij} + \varepsilon_{ij},\ i = 1, \ldots, N,\ j = 1, \ldots, J,$$

where p_{ij} are the unknown multinomial probabilities and ε_{ij} are additive noise components for each individual. Since the observed Y's are either zero or one, the noise components are naturally contained in $[-1, 1]$ for each individual. Rather than choosing a specific $F(\cdot)$, we connect the observables and unobservables via the cross moments:

$$\sum_i y_{ij}x_{ijk} = \sum_i x_{ijk}p_{ij} + \sum_i x_{ijk}\varepsilon_{ij} \tag{3.5}$$

where there are $(N \times (J-1))$ unknown probabilities, but only $(K \times J)$ data points or moments. We call these moments "stochastic moments" as the last term is different from the traditional (pure) moment representation of $\sum_i y_{ij}x_{ijk} = \sum_i x_{ijk}p_{ij}$.

Next, we reformulate the model to be consistent with the mixed logit data generation process. Let each p_{ij} be expressed as the expected value of an M-dimensional discrete random variable $\mathbf{s}$ (or an equally spaced support) with underlying probabilities π_{ij}. Thus, $p_{ij} \equiv \sum_m^M s_m\pi_{ijm}$, $s_m \in [0, 1]$ and $m = 1, 2, \ldots, M$ with $M \geq 2$ and where $\sum_m^M \pi_{ijm} = 1$. (We consider an extension to a continuous version of the model in Section 3.4.) To formulate this model within the IT-GME approach, we need to attach each one of the unobserved disturbances ε_{ij} to a proper probability distribution. To do so, let ε_{ij} be the expected value of an H-dimensional support space (random variable) $\mathbf{u}$ with corresponding H-dimensional vector of weights, $\mathbf{w}$. Specifically, let $\mathbf{u} = (-1/\sqrt{N}, \ldots, 0, \ldots 1/\sqrt{N})'$, so $\varepsilon_{ij} \equiv \sum_{h=1}^H u_h w_{ijh}$ (or $\varepsilon_i = E[u_i]$) with $\sum_h w_{ijh} = 1$ for each ε_{ij}.

Thus, the H-dimensional vector of weights (proper probabilities) $\mathbf{w}$ converts the errors from the $[-1, 1]$ space into a set of $N \times H$ proper probability

distributions within **u**. We now reformulate Equation 3.5 as

$$\sum_i y_{ij}x_{ijk} = \sum_i x_{ijk}p_{ij} + \sum_i x_{ijk}\varepsilon_{ij} = \sum_{i,m} x_{ijk}s_m\pi_{ijm} + \sum_{i,h} x_{ijk}u_h w_{ijh}. \tag{3.6}$$

As we discussed previously, rather than using a simulated likelihood approach, our objective is to estimate, with minimal assumptions, the two sets of unknown $\boldsymbol{\pi}$ and **w** simultaneously. Since the problem is inherently underdetermined, we resort to the Maximum Entropy method (Jaynes 1957a, 1957b, 1978; Golan, Judge, and Miller, 1996; Golan, Judge, and Perloff, 1996). Under that approach, one uses an information criterion, called entropy (Shannon 1948), to choose one of the infinitely many probability distributions consistent with the observed data (Equation 3.6). Let $H(\boldsymbol{\pi}, \mathbf{w})$ be the joint entropies of $\boldsymbol{\pi}$ and **w**, defined below. (See Golan, 2008, for a recent review and formulations of that class of estimators.) Then, the full set of unknown $\{\boldsymbol{\pi}, \mathbf{w}\}$ is estimated by maximizing $H(\boldsymbol{\pi}, \mathbf{w})$ subject to the observed stochastic moments (Equation 3.6) and the requirement that $\{\boldsymbol{\pi}\}$, {**w**} and {P} are proper probabilities. Specifically,

$$\underset{\boldsymbol{\pi},\mathbf{w}}{\text{Max}} \left\{ H(\boldsymbol{\pi},\mathbf{w}) = -\sum_{ijm} \pi_{ijm}\log\pi_{ijm} - \sum_{ijh} w_{ijh}\log w_{ijh} \right\} \tag{3.7}$$

subject to

$$\begin{aligned}\sum_i y_{ij}x_{ijk} &= \sum_i x_{ijk}p_{ij} + \sum_i x_{ijk}\varepsilon_{ij} \\ &= \sum_{i,m} x_{ijk}s_m\pi_{ijm} + \sum_{i,h} x_{ijk}u_h w_{ijh}\end{aligned} \tag{3.8}$$

$$\sum_m \pi_{ijm} = 1, \ \sum_h w_{ijh} = 1 \tag{3.9a}$$

$$\sum_{j,m} s_m\pi_{ijm} = 1 \tag{3.9b}$$

with $\mathbf{s} \in [0, 1]$ and $\mathbf{u} \in (-1, 1)$.

Forming the Lagrangean and solving yields the IT estimators for $\boldsymbol{\pi}$

$$\hat{\pi}_{ijm} = \frac{\exp\left[s_m\left(-\sum_k \hat{\lambda}_{kj}x_{ijk} - \hat{\rho}_i\right)\right]}{\sum_{m=1}^{M}\exp\left[s_m\left(-\sum_k \hat{\lambda}_{kj}x_{ijk} - \hat{\rho}_i\right)\right]} \equiv \frac{\exp\left[s_m\left(-\sum_k \hat{\lambda}_{kj}x_{ijk} - \hat{\rho}_i\right)\right]}{\Omega_{ij}(\hat{\lambda}, \hat{\rho})}, \tag{3.10}$$

and for **w**

$$\hat{w}_{ijh} = \frac{\exp\left(-u_h\sum_k x_{ijk}\hat{\lambda}_{jk}\right)}{\sum_h \exp\left(-u_h\sum_k x_{ijk}\hat{\lambda}_{jk}\right)} \equiv \frac{\exp\left(-u_h\sum_k x_{ijk}\hat{\lambda}_{jk}\right)}{\Psi_{ij}(\hat{\lambda})} \tag{3.11}$$

where $\boldsymbol{\lambda}$ is the set of $K \times (J-1)$ Lagrange multiplier (estimated coefficients) associated with (Equation 3.8) and $\boldsymbol{\rho}$ is the N-dimensional vector of Lagrange

multipliers associated with Equation 3.9a). Finally, $\hat{p}_{ij} = \sum_m s_m \hat{\pi}_{ijm}$ and $\hat{\varepsilon}_{ij} = \sum_h u_h \hat{w}_{ijh}$. These $\boldsymbol{\lambda}$'s are the $\boldsymbol{\alpha}$'s and $\boldsymbol{\beta}$'s defined and discussed in Section 3.1: $\boldsymbol{\lambda}' = (\boldsymbol{\alpha}', \boldsymbol{\beta}')$. We now can construct the concentrated entropy (profile likelihood) model which is just the dual version of the above constrained optimization model. This allows us to concentrate the model on the lower dimensional, real parameters of interest ($\boldsymbol{\lambda}$ and $\boldsymbol{\rho}$). That is, we move from the {P, W} space to the {$\boldsymbol{\lambda}$, $\boldsymbol{\rho}$} space.

The concentrated entropy (likelihood) model is

$$\underset{\lambda,\rho}{\text{Min}} \left\{ -\sum_{ijk} y_{ij} x_{ijk} \lambda_{kj} + \sum_i \rho_i + \sum_{ij} \ln\Omega_{ij}(\lambda, \rho) + \sum_{ij} \ln\Psi_{ij}(\lambda) \right\}. \tag{3.12}$$

Solving with respect to $\boldsymbol{\lambda}$ and $\boldsymbol{\rho}$, we use Equation 3.10 and Equation 3.11 to get $\hat{\boldsymbol{\pi}}$ and $\hat{\mathbf{w}}$ that are then transformed to $\hat{\mathbf{p}}$ and $\hat{\varepsilon}$.

Returning to the mixed logit (Mlogit) model discussed earlier, the set of parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\rho}$ are the parameters in the individual utility functions (Equation 3.2 or 3.3) and represent both the population means and the random (individual) parameters. But unlike the simulated likelihood approach, no simulations are done here. Under this general criterion function, the objective is to minimize the joint entropy distance between the data and the state of complete ignorance (the uniform distribution or the uninformed empirical distribution). It is a dual-loss criterion that assigns equal weights to prediction (P) and precision (W). It is a shrinkage estimator that simultaneously shrinks the data and the noise to the center of their pre-specified supports. Further, looking at the basic primal (constrained) model, it is clear that the estimated parameters reflect not only the unknown parameters of the distribution, but also the amount of information in each one of the stochastic moments (Equation 3.8). Thus, λ_{kj} reflects the informational contribution of moment kj. It is the reduction in entropy (increase in information) as a result of incorporating that moment in the estimation. The $\boldsymbol{\rho}$'s reflect the individual effects.

As common to these class of models, the analyst is not (usually) interested in the parameters, but rather in the marginal effects. In the model developed here, the marginal effects (for the continuous covariates) are

$$\frac{\partial p_{ij}}{\partial x_{ijk}} = \sum_m s_m \frac{\partial \pi_{ijm}}{\partial x_{ijk}}$$

with

$$\frac{\partial \pi_{ijm}}{\partial x_{ijk}} = \pi_{ijm} \left(s_m \lambda_{kj} - \sum_m \pi_{ijm} s_m \lambda_{kj} \right)$$

and finally

$$\frac{\partial p_{ij}}{\partial x_{ijk}} = \sum_m s_m \left[\pi_{ijm} \left(s_m \lambda_{kj} - \sum_m \pi_{ijm} s_m \lambda_{kj} \right) \right].$$

3.4 Extensions and Discussion

So far in our basic model (Equation 3.12) we used discrete probability distributions (or similarly discrete spaces) and uniform (uninformed) priors. We now extend our basic model to allow for continuous spaces and for nonuniform priors. We concentrate here on the noise distributions.

3.4.1 Triangular Priors

Under the model formulated above, we maximize the joint entropies subject to our constraints. This model can be reconstructed as a minimization of the entropy distance between the (yet) unknown posteriors and some priors (subject to the same constraints). This class of methods is also known as "cross entropy" models (e.g., Kullback 1959; Golan, Judge, and Miller, 1996). Let, w^0_{ijh} be a set of prior (proper) probability distributions on $\mathbf{u}$. The normalization factors (partition functions) for the errors are now

$$\Psi_{ij} = \sum_h w^0_{ijh} \exp\left(u_h \sum_k x_{ijk}\lambda_{jk}\right)$$

and the concentrated IT criterion (Equation 3.12) becomes

$$\underset{\lambda,\rho}{\text{Max}} \left\{ \sum_{ijk} y_{ij} x_{ijk} \lambda_{kj} - \sum_i \rho_i - \sum_{ij} \ln\Omega_{ij}(\lambda, \rho) - \sum_{ij} \ln\Psi_{ij}(\lambda) \right\}.$$

The estimated $\mathbf{w}$'s are:

$$\tilde{w}_{ijh} = \frac{w^0_{ijh} \exp\left(u_h \sum_k x_{ijk}\tilde{\lambda}_{jk}\right)}{\sum_h w^0_{ijh} \exp\left(u_h \sum_k x_{ijk}\tilde{\lambda}_{jk}\right)} \equiv \frac{w^0_{ijh} \exp\left(u_h \sum_k x_{ijk}\tilde{\lambda}_{jk}\right)}{\Psi_{ij}(\tilde{\lambda})}$$

and $\tilde{\varepsilon}_{ij} = \sum_h u_h \tilde{w}_{ijh}$. If the priors are all uniform ($w^0_{ijh} = 1/H$ for all i and j) this estimator is similar to Equation 3.12. In our model, the most reasonable prior is the triangular prior with higher weights on the center (zero) of the support $\mathbf{u}$. For example, if $H = 3$ one can specify $w^0_{ij1} = 0.25$, $w^0_{ij2} = 0.5$ and $w^0_{ij3} = 0.25$ or for $H = 5$, $\mathbf{w}^0 = (0.05, 0.1, 0.7, 0.1, 0.05)'$ or any other triangular prior the user believes to be consistent with the data generating process. Note that like the uniform prior, the a priori mean (for each ε_{ij}) is zero. Similarly, if such information exists, one can incorporate the priors for the signal. However, unlike the noise priors just formulated, we cannot provide here a natural source for such priors.

3.4.2 Bernoulli

A special case of our basic model is the Bernoulli priors. Assuming equal weights on the two support bounds, and letting $\eta_{ij} = \sum_k x_{ijk}\lambda_{jk}$ and u_1 is the

support bound such that $\mathbf{u} \in [-u_1, u_1]$, then the errors' partition function is

$$\begin{aligned}\Psi(\lambda) &= \prod_{ij} \frac{1}{2}\left(e^{\sum_k x_{ijk}\lambda_{jk}u_1} + e^{-\sum_k x_{ijk}\lambda_{jk}u_1}\right) \\ &= \prod_{ij} \frac{1}{2}\left(e^{\eta_{ij}u_1} + e^{-\eta_{ij}u_1}\right) = \prod_{ij} \cosh(\eta_{ij}u_1).\end{aligned}$$

Then Equation 3.12 becomes

$$\max_{\lambda,\rho}\left\{\sum_{ijk} y_{ij}x_{ijk}\lambda_{kj} - \sum_i \rho_i - \sum_{ij} \ln\Omega_{ij}(\lambda, \rho) - \sum_{ij} \ln\Psi_{ij}(\lambda)\right\}$$

where

$$\sum_{ij} \ln\Psi_{ij}(\lambda) = \sum_{ij} \ln\left[\frac{1}{2}\left(e^{\eta_{ij}u_1} + e^{-\eta_{ij}u_1}\right)\right] = \sum_{ij} \ln\cosh(\eta_{ij}u_1).$$

Next, consider the case of Bernoulli model for the signal $\boldsymbol{\pi}$. Recall that $s_m \in [0, 1]$ and let the priors weights be q_1 and q_2 on zero (s_1) and one (s_2), respectively. The signal partition function is

$$\begin{aligned}\Omega(\lambda,\rho) &= \prod_{ij}\left(q_1 e^{s_1\left(\sum_k x_{ijk}\lambda_{jk}+\rho_i\right)} + q_2 e^{s_2\left(\sum_k x_{ijk}\lambda_{jk}+\rho_i\right)}\right) \\ &= \prod_{ij}\left(q_1 + q_2 e^{\sum_k x_{ijk}\lambda_{jk}+\rho_i}\right) = \prod_{ij}\left(q_1 + q_2 e^{\eta_{ij}+\rho_i}\right)\end{aligned}$$

and Equation 3.12 is now

$$\max_{\lambda,\rho}\left\{\sum_{ijk} y_{ij}x_{ijk}\lambda_{kj} - \sum_i \rho_i - \sum_{ij} \ln\Omega_{ij}(\lambda, \rho) - \sum_{ij} \ln\Psi_{ij}(\lambda)\right\}$$

where

$$\sum_{ij} \ln\Omega_{ij}(\lambda, \rho) = \sum_{ij} \ln\left[q_1 + q_2 e^{\eta_{ij}+\rho_i}\right].$$

Traditionally, one would expect to set uniform priors ($q_1 = q_2 = 0.5$).

3.4.3 Continuous Uniform

Using the same notations as above and recalling that $\mathbf{u} \in [-u_1, u_1]$, the errors' partition functions for continuous uniform priors are

$$\Psi_{ij}(\lambda) = \frac{e^{\eta_{ij}u_1} - e^{-\eta_{ij}u_1}}{2u_1\eta_{ij}} = \frac{\sinh(u_1\eta_{ij})}{u_1\eta_{ij}}.$$

The right-hand side term of Equation 3.12 becomes

$$\sum_{ij} \ln\Psi_{ij}(\lambda) = \sum_{ij}\left[\ln\left(\frac{1}{2}\left(e^{\eta_{ij}u_1} + e^{-\eta_{ij}u_1}\right)\right) - \ln\left(\eta_{ij}u_1\right)\right]$$
$$= \sum_{ij}\left[\ln\left(\sinh\left(\eta_{ij}u_1\right)\right) - \ln\left(\eta_{ij}u_1\right)\right].$$

Similarly, and in general notations, for any uniform prior $[a, b]$, the signal partition function for each i and j is

$$\Omega_{ij}(\lambda,\rho) = \frac{e^{a(-\eta_{ij}-\rho_i)} - e^{b(-\eta_{ij}-\rho_i)}}{(b-a)\eta_{ij}}.$$

This reduces to

$$\Omega_{ij}(\lambda,\rho) = \frac{1 - e^{-\eta_{ij}-\rho_i}}{\eta_{ij}}$$

for the base case $[a, b] = [0, 1]$ which is the natural support for the signal in our model. The basic model is then

$$\underset{\lambda,\rho}{\text{Min}}\left\{\sum_{ijk} y_{ij}x_{ijk}\lambda_{kj} - \sum_i \rho_i - \sum_{ij}\left[\ln\left(1 - e^{-\eta_{ij}-\rho_i}\right) - \ln\left(\eta_{ij}\right)\right]\right.$$
$$\left. - \sum_{ij}\left[\ln\left(\sinh\left(\eta_{ij}u_1\right)\right) - \ln\left(2\eta_{ij}u_1\right)\right]\right\}$$
$$= \underset{\lambda,\rho}{\text{Min}}\left\{\sum_{ijk} y_{ij}x_{ijk}\lambda_{kj} - \sum_i \rho_i - \sum_{ij}\ln\Omega_{ij}(\lambda,\rho) - \sum_{ij}\ln\Psi_{ij}(\lambda)\right\}.$$

Finally, the estimator for P (the individuals' choices) is

$$\widehat{p}_{ij} = \frac{1}{(b-a)}\left\{\frac{ae^{a(-\widehat{\eta}_{ij}-\widehat{\rho}_i)} - be^{b(-\widehat{\eta}_{ij}-\widehat{\rho}_i)}}{\widehat{\eta}_{ij}} + \frac{e^{a(-\widehat{\eta}_{ij}-\widehat{\rho}_i)} - e^{b(-\widehat{\eta}_{ij}-\widehat{\rho}_i)}}{\widehat{\eta}_{ij}^2}\right\}$$

for any $[a, b]$ and

$$\widehat{p}_{ij} = \frac{-e^{-\widehat{\eta}_{ij}-\widehat{\rho}_i}}{\widehat{\eta}_{ij}} + \frac{1 - e^{-\widehat{\eta}_{ij}-\widehat{\rho}_i}}{\widehat{\eta}_{ij}^2}$$

for our problem of $[a, b] = [0, 1]$.

In this section we provided further detailed derivations and background for our proposed IT estimator. We concentrated here on prior distributions that seem to be consistent with the data generating process. Nonetheless, in some very special cases, the researcher may be interested in specifying other structures that we did not discuss here. Examples include normally

distributed errors or possibly truncated normal with truncation points at -1 and 1. These imply normally distributed $\mathbf{w}_i$s within their supports. Though, mathematically, we can provide these derivations, we do not do it here as it does not seem to be in full agreement with our proposed model.

3.5 Inference and Diagnostics

In this section we provide some basic statistics that allow the user to evaluate the results. We do not develop here large sample properties of our estimator. There are two basic reasons for that. First, and most important, using the error supports $\mathbf{v}$ as formulated above, it is trivial to show that this model converges to the ML Logit. (See Golan, Judge, and Perloff, 1996, for the proof of the simpler IT-GME model.) Therefore, basic statistics developed for the ML logit are easily modified for our model. The second reason is simply that our objective here is to provide the user with the necessary tools for diagnostics and inference when analyzing finite samples.

Following Golan, Judge, and Miller (1996) and Golan (2008) we start by defining the information measures, or normalized entropies

$$S_1(\hat{\boldsymbol{\pi}}) \equiv \frac{-\sum_{ijm} \hat{\boldsymbol{\pi}}_{ijm} \ln \hat{\boldsymbol{\pi}}_{ijm}}{(N \times J) \ln(M)}$$

and

$$S_2(\hat{\boldsymbol{\pi}}_{ij}) \equiv \frac{-\sum_{m} \hat{\boldsymbol{\pi}}_{ijm} \ln \hat{\boldsymbol{\pi}}_{ijm}}{\ln(M)},$$

where both sets of measures are between zero and one, with one reflecting uniformity (complete ignorance: $\boldsymbol{\lambda} = \mathbf{0}$) of the estimates, and zero reflecting perfect knowledge. The first measure reflects the (signal) information in the whole system, while the second one reflects the information in each i and j. Similar information measures of the form $I(\hat{\boldsymbol{\pi}}) = 1 - S_j(\hat{\boldsymbol{\pi}})$ are also used (e.g., Soofi, 1994).

Following the traditional derivation of the (empirical) likelihood ratio test (within the likelihood literature), the empirical likelihood literature (Owen 1988, 1990, 2001; Qin and Lawless 1994), and the IT literature, we can construct an entropy ratio test. (For additional background on IT see also Mittelhammer, Judge, and Miller, 2000.) Let ℓ_Ω be the unconstrained entropy model Equation 3.12, and ℓ_ω be the constrained one where, say $\boldsymbol{\gamma}' = (\boldsymbol{\lambda}', \boldsymbol{\rho}') = \mathbf{0}$, or similarly $\boldsymbol{\beta} = \boldsymbol{\alpha} = \mathbf{0}$ (in Section 3.2). Then, the entropy ratio statistic is $2(\ell_\omega - \ell_\Omega)$. The value of the unconstrained problem ℓ_Ω is just the value of $\text{Max}\{H(\boldsymbol{\pi}, \mathbf{w})\}$, or similarly the maximal value of Equation 3.12, while $\ell_\omega = (N \times J)\ln(M)$ for uniform $\boldsymbol{\pi}$'s. Thus, the entropy-ratio statistic is just

$$W(IT) = 2(\ell_\omega - \ell_\Omega) = 2(N \times J)\ln(M)[1 - S_1(\hat{\boldsymbol{\pi}})].$$

Under the null hypothesis, W(IT) converges in distribution to $\chi^2_{(n)}$ where "n" reflects the number constraints (or hypotheses). Finally, we can derive the Pseudo-R^2 (McFadden 1974) which gives the proportion of variation in the data that is explained by the model (a measure of model fit):

$$\text{Pseudo-}R^2 \equiv 1 - \frac{\ell_\Omega}{\ell_\omega} = 1 - S_1(\hat{\pi}).$$

To make it somewhat clearer, the relationship of the entropy criterion and the χ^2 statistic can be easily shown. Consider, for example the cross entropy criterion discussed in Section 3.4. This criterion reflects the entropy distance between two proper distributions such as a prior and post-data (posterior) distributions. Let $I(\pi||\pi^0)$ be the entropy distance between some distribution $\boldsymbol{\pi}$ and its prior π^0. Now, with a slight abuse of notations, to simplify the explanation, let $\{\boldsymbol{\pi}\}$ be of dimension M. Let the null hypothesis be $H_0 : \pi = \pi^0$. Then,

$$\chi^2_{(M-1)} = \sum_m \frac{1}{\pi^0_m}(\pi_m - \pi^0_m)^2.$$

Looking at the entropy distance (cross entropy) measure $I\left(\pi||\pi^0\right)$ and formulating a second order approximation yields

$$I(\boldsymbol{\pi}||\boldsymbol{\pi}^0) \equiv \sum_m \pi_m \log(\pi_m/\pi^0_m) \cong \frac{1}{2}\sum_m \frac{1}{\pi^0_m}(\pi_m - \pi^0_m)^2$$

which is just the entropy (log-likelihood) ratio statistic of this estimator. Since 2 times the log-likelihood ratio statistic corresponds approximately to χ^2, the relationship is clear. Finally, though we used here a certain prior π^0, the derivation holds for all priors, including the uniform (uninformed) priors (e.g., $\pi_m = 1/M$) used in Section 3.3.

In conclusion, we stress the following: Under our IT-GME approach, one investigates how "far" the data pull the estimates away from a state of complete ignorance (uniform distribution). Thus, a high value of χ^2 implies the data tell us something about the estimates, or similarly, there is valuable information in the data. If, however, one introduces some priors (Section 3.4), the question becomes how far the data take us from our initial (a priori) beliefs — the priors. A high value of χ^2 implies that our prior beliefs are rejected by the data. For more discussion and background on goodness of fit statistics for multinomial type problems see Greene (2008). Further discussion of diagnostics and testing for ME-ML model (under zero moment conditions) appears in Soofi (1994). He provides measures related to the normalized entropy measures discussed above and provides a detailed formulation of decomposition of these information concepts. For detailed derivations of statistics for a whole class of IT models, including discrete choice models, see Golan (2008) as well as Good (1963). All of these statistics can be used in the model developed here.

3.6 Simulated Examples

Sections 3.3 and 3.4 have developed our proposed IT model and some extensions. We also discussed some of the motivations for using our proposed model, namely that it is semiparametric, and that it is not dependent on simulated likelihood approaches. It remains to investigate and contrast the IT model with its competitors. We provide a number of simulated examples for different sample sizes and different level of randomness. Among the appeals of the Mixed Logit, (RP) models is its ability to predict the individual choices. The results below include the in-sample and out-of-sample prediction tables for the IT models as well.

The out-of-sample predictions for the simulated logit is trivial and is easily done using NLOGIT (discussed below). For the IT estimator, the out-of-sample prediction involves estimating the $\boldsymbol{\rho}$'s as well. Using the first sample and the estimated $\boldsymbol{\rho}$'s from the IT model (as the dependent variables), we run a Least Squares model and then use these estimates to predict the out-of-sample $\boldsymbol{\rho}$'s. We then use these predicted $\boldsymbol{\rho}$'s and the estimated $\boldsymbol{\lambda}$'s from the first sample to predict out-of-sample.

3.6.1 The Data Generating Process

The simulated model is a five-choice setting with three independent variables. The utility functions are based on random parameters on the attributes, and five nonrandom choice specific intercepts (the last of which is constrained to equal zero). The random errors in the utility functions (for each individual) are iid extreme value in accordance with the multinomial logit specification. Specifically, $\mathbf{x}_1$ is a randomly assigned discrete (integer) uniform in [1, 5], $\mathbf{x}_2$ is from the uniform (0, 1) population and $\mathbf{x}_3$ is normal (0, 1). The values for the $\boldsymbol{\beta}$'s are: $\beta_{1i} = 0.3 + 0.2u_1$, $\beta_{2i} = -0.3 + 0.1u_2$, and $\beta_{3i} = 0.0 + 0.4u_3$, where u_1, u_2 and u_3 are iid normal (0, 1). The values for the choice specific intercept ($\boldsymbol{\alpha}$) are 0.4, 0.6, -0.5, 0.7 and 0.0 respectively for choices $j = 1, \ldots, 5$. In the second set of experiments, $\boldsymbol{\alpha}$'s are also random. Specifically, $\alpha_{ij} = \alpha_j + 0.5u_{ij}$, where u_j is iid normal(0,1) and $j = 1, 2, \ldots, 5$.

3.6.2 The Simulated Results

Using the software NLOGIT (Nlogit) for the MLogit model, we created 100 samples for the simulated log-likelihood model. We used GAMS for the IT-GME models – the estimator in NLOGIT was developed during this writing. For a fair comparison of the two different estimators, we use the correct model for the simulated likelihood (Case A) and a model where all parameters are taken to be random (Case B). In both cases we used the correct likelihood. For the IT estimator, we take all parameters to be random and there is no need for incorporating distributional assumptions. This means that if the IT dominates when it's not the correct model, it is more robust for the underlying

TABLE 3.1

In and Out-of-Sample Predictions for Simulated Experiments. All Values Are the Percent of Correctly Predicted

	N = 100 In/Out	N = 200 In/Out	N = 500 In/Out	N = 1000 In/Out	N = 1500 In/Out	N = 3000 In
Case 1: Random $\boldsymbol{\beta}$						
MLogit - A	29/28	34/38.5	34.4/33.6	35.5/33.3	34.6/34.0	33.8
MLogit - B	29/28	32.5/28.5	31.4/26.8	29.9/28.9	28.5/29	29.4
IT-GME*	41/23	35/34	33.6/35.6	36.4/34.6	34.4/33.9	34.8
Case 2: Random $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$						
MLogit	31/22	31/27	34.2/26.8	32/28.9	30.3/31.9	31
IT-GME*	45/29	40.5/29.5	38.4/32.4	37/34.2	37.1/34.9	36.3

Note: A: The correct model.
B: The incorrect model (both $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ random).
*All IT-GME models are for both $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ random.

structure of the parameters. The results are presented in Table 3.1. We note a number of observations regarding these experiments. First, the IT-GME model converges far faster than the simulated likelihood approach–since no simulation is needed, all expressions are in closed form. Second, in the first set of experiments (only the $\boldsymbol{\beta}$'s are random) and using the correct simulated likelihood model (Case 1A), both models provide very similar (on average) predictions, though the IT model is slightly superior. In the more realistic case, when the user does not know the exact model and uses RP for all parameters (Case 1B), the IT method is always superior. Third, for the more complicated data (generated with RP for both $\boldsymbol{\beta}$'s and $\boldsymbol{\alpha}$'s) – Case 2 – the IT estimator dominates for all sample sizes.

In summary, though the IT estimator seems to dominate for all samples and structures presented, it is clear that its relative advantage increases as the sample size decreases and as the complexity (number of random parameters) increases. From the analyst's point of view, it seems that for data with many choices and with much uncertainty about the underlying structure of the model, the IT is an attractive method to use. For the less complicated models and relatively large data sets, the simulated likelihood methods are proper (but are computationally more demanding and are based on a stricter set of assumptions).

3.7 Concluding Remarks

In this chapter we formulate and discuss an IT estimator for the mixed discrete choice model. This model is semiparametric and performs well relative to the class of simulated likelihood methods. Further, the IT estimator is computationally more efficient and is easy to use. This chapter is written in a way that

makes it possible for the potential user to easily use this estimator. A detailed formulation of different potential priors and frameworks, consistent with the way we visualize the data generating process, is provided as well. We also provide the concentrated model that can be easily coded in some software.

References

Berry, S., J. Levinsohn and A. Pakes. 1995. Automobile Prices in Market Equilibrium. *Econometrica* 63(4): 841–890.

Bhat, C. 1996. *Accommodating Variations in Responsiveness to Level-of-Service Measures in Travel Mode Choice Modeling.* Department of Civil Engineering, University of Massachusetts, Amherst.

Golan, A. 2008. Information and Entropy Econometrics – A Review and Synthesis. *Foundations and Trends® in Econometrics* 2(1–2): 1–145.

Golan, A., G. Judge, and D. Miller. 1996. *Maximum Entropy Econometrics: Robust Estimation with Limited Data.* New York: John Wiley & Sons.

Golan, A., G. Judge, and J. Perloff. 1996. A Generalized Maximum Entropy Approach to Recovering Information from Multinomial Response Data. *Journal of the American Statistical Association* 91: 841–853.

Good, I.J. 1963. Maximum Entropy for Hypothesis Formulation, Especially for Multidimensional Contingency Tables. *Annals of Mathematical Statistics* 34: 911–934.

Greene, W.H. 2008. *Econometric Analysis*, 6th ed. Upper Saddle River, NJ: Prentice Hall.

Hensher D.A., and W.H., Greene. 2003. The Mixed Logit Model: The State of Practice. *Transportation* 30(2): 133–176.

Hensher, D.A., J. M. Rose, and W.H. Greene. 2006. *Applied Choice Analysis*. Cambridge, U.K.: Cambridge University Press.

Jain, D., N. Vilcassim, and P. Chintagunta. 1994. A Random-Coefficients Logit Brand Choice Model Applied to Panel Data. *Journal of Business and Economic Statistic* 12(3): 317–328.

Jaynes, E.T. 1957a. Information Theory and Statistical Mechanics. *Physics Review* 106: 620–630.

Jaynes, E.T. 1957b. Information Theory and Statistical Mechanics II. *Physics Review* 108: 171–190.

Jaynes, E.T. 1978. Where Do We Stand on Maximum Entropy. In *The Maximum Entropy Formalis*, eds. R.D. Levine and M. Tribus, pp. 15–118. Cambridge, MA: MIT Press.

Kullback, S. 1959. *Information Theory and Statistics.* New York: John Wiley & Sons.

McFadden, D. 1974. Conditional Logit Analysis of Qualitative Choice Behavior. In *Frontiers of Econometrics*, ed. P. Zarembka. New York: Academic Press, pp. 105–142.

Mittelhammer, R.C., G. Judge, and D. M. Miller. 2000. *Econometric Foundations*. Cambridge, U.K.: Cambridge University Press.

Owen, A. 1988. Empirical Likelihood Ratio Confidence Intervals for a Single Functional. *Biometrika* 75(2): 237–249.

Owen, A. 1990. Empirical Likelihood Ratio Confidence Regions. *The Annals of Statistics* 18(1): 90–120.

Owen, A. 2001. *Empirical Likelihood*. Boca Raton, FL: Chapman & Hall/CRC.

Qin, J., and J. Lawless. 1994. Empirical Likelihood and General Estimating Equations. *The Annals of Statistics* 22: 300–325.

Revelt, D., and K. Train. 1998. Mixed Logit with Repeated Choices of Appliance Efficiency Levels. *Review of Economics and Statistics* LXXX (4): 647–657.

Shannon, C.E. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal* 27: 379–423.

Soofi, E.S. 1994. Capturing the Intangible Concept of Information. *Journal of the American Statistical Association* 89(428): 1243–1254.

Train, K.E. 2003. *Discrete Choice Methods with Simulation*. New York: Cambridge University Press.

Train, K., D. Revelt, and P. Ruud. 1996. *Mixed Logit Estimation Routine for Cross-Sectional Data*. UC Berkeley, http://elsa.berkeley.edu/Software/abstracts/train0196.html.