A Maximum Entropy Approach to Recovering Information From Multinomial Response Data

Amos Golan; George Judge; Jeffrey M. Perloff

Journal of the American Statistical Association, Vol. 91, No. 434 (Jun., 1996), 841-853.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28199606%2991%3A434%3C841%3AAMEATR%3E2.0.CO%3B2-L

Journal of the American Statistical Association is currently published by American Statistical Association.

A Maximum Entropy Approach to Recovering Information From Multinomial Response Data

Amos GOLAN, George JUDGE, and Jeffrey M. PERLOFF

The classical maximum entropy (ME) approach to estimating the unknown parameters of a multinomial discrete choice problem, which is equivalent to the maximum likelihood multinomial logit (ML) estimator, is generalized. The generalized maximum entropy (GME) model includes noise terms in the multinomial information constraints. Each noise term is modeled as the mean of a finite set of a priori known points in the interval $[-1, 1]$ with unknown probabilities where no parametric assumptions about the error distribution are made. A GME model for the multinomial probabilities and for the distributions associated with the noise terms is derived by maximizing the joint entropy of multinomial and noise distributions, under the assumption of independence. The GME formulation reduces to the ME in the limit as the sample grows large or when no noise is included in the entropy maximization. Further, even though the GME and the logit estimators are conceptually different, the dual GME model is related to a generalized class of logit models. In finite samples, modeling the noise terms along with the multinomial probabilities results in a gain of efficiency of the GME over the ME and ML. In addition to analytical results, some extensions, sampling experiments, and an example based on real-world data are presented.

KEY WORDS: Choice-based sampling process; Discrete choice models; Generalized maximum entropy; Maximum entropy principle; Maximum likelihood logit; Nonlinear inversion procedure; Unordered multinomial process.

1. INTRODUCTION

This article generalizes the classical pure-moment maximum entropy (ME) formulation of the unordered, multinomial choice problem proposed by Denzau, Gibbons, and Greenberg (1989) and Soofi (1992) to include a noise component in the multinomial information-moment constraints. The resulting generalized maximum entropy (GME) estimator has the same desirable sampling properties for large samples as the classical ME-maximum likelihood (ML) estimators. In finite samples the GME estimator is superior to the ME-ML estimators.

The GME has several advantages over the conventional ML formulations (Greene 1993; Judge, Griffiths, Hill, Lütkepohl, and Lee 1985; Maddala 1983; Manski and McFadden 1981; McFadden 1974). The main advantages are that it is more efficient; avoids strong parametric assumptions; works well when the sample is small, covariates are highly correlated, or the design matrix is ill-conditioned; and has a dual objective that permits a choice between estimation precision and category prediction. It is as easy to compute and use as the ML. Moreover, both the ME and GME estimators permit incorporation of nonsample information on both the multinomial probabilities and the response parameters and it provides a basis for model diagnostics and information measures (Soofi 1992).

The unordered multinomial discrete choice problem that the ML, ME, and GME are designed to answer is as follows. Suppose, in an experiment consisting of N trials, that binary random variables $y_{1j}, y_{2j}, \ldots, y_{Nj}$ are observed, where y_{ij}, for $i = 1, 2, \ldots, N$, takes on one of J *unordered* categories for $j = 1, 2, \ldots, J$. Thus on trial i, one of the J alternatives is observed in the form of a binary variable, y_{ij}, that equals unity iff alternative j is observed on trial i and zero otherwise. Let the probability of alternative j on trial i be $p_{ij} = \Pr[y_{ij} = 1]$ and assume the p_{ij} are related to a set of explanatory variables through the model

$$p_{ij}(\boldsymbol{\beta}) \equiv \Pr(y_{ij} = 1 | \mathbf{x}_i, \boldsymbol{\beta}) = G(\mathbf{x}_i'\boldsymbol{\beta}_j) > 0$$

$$\text{for} \quad i = 1, 2, \ldots, N \quad \text{and} \quad j = 1, 2, \ldots, J \quad (1)$$

where $\boldsymbol{\beta}_j$ is a $(K \times 1)$ vector of unknown parameters, $\mathbf{x}_i' = (x_{i1}, x_{i2}, \ldots, x_{iK})$ is a $(1 \times K)$ vector of covariates, and $G(\cdot)$ is a function, linking the probabilities p_{ij} with the linear structure $\mathbf{x}_i'\boldsymbol{\beta}_j$, that maps the real line into $[0, 1]$ and $\sum_{j=1}^{J} G(\mathbf{x}_i'\boldsymbol{\beta}_j) = 1$. More generally, the covariates $\mathbf{x}_i'$ could vary by category, but for expositional simplicity we focus on a basic case.

Suppose that we are given noisy data y_{ij}, which we model as

$$y_{ij} = G(\mathbf{x}_i'\boldsymbol{\beta}_j) + e_{ij} = p_{ij} + e_{ij}, \quad (2)$$

where the p_{ij} are the unknown multinomial probabilities and the e_{ij} are noise components contained in the interval $[-1, 1]$. Equation (2) may be written in matrix notation as

$$\mathbf{y} = \mathbf{p} + \mathbf{e}, \quad (3)$$

where $\mathbf{y}$ and $\mathbf{x}$ are observed and $\mathbf{p}$ and $\mathbf{e}$ are unknown and unobserved.

If we use the traditional ML approach, with log-likelihood function $L(G, \boldsymbol{\beta}_i) = \sum_i \sum_j y_{ij} \log_e G(\mathbf{x}_i'\boldsymbol{\beta}_j)$, then the multinomial logit model and solution is obtained if $G(\cdot)$ is a logistic cumulative density function (cdf). Similarly, the multinomial probit model and solution is obtained if $G(\cdot)$ is a standard Gaussian cdf.

The format of this article is as follows. In Section 2 the equivalence of the ML and ME models is shown (Denzau et al. 1989; Soofi 1992), our new dual formulation is presented, and the information matrix is derived. In Section 3 a

Amos Golan is Visiting Professor, George Judge is Professor, and Jeffrey M. Perloff is Professor, Department of Agricultural and Resource Economics, University of California, Berkeley, CA 94720. The authors gratefully acknowledge helpful comments and suggestions from William Griffiths, Carter Hill, Douglas Miller, Dale Poirier, Paul Ruud, Ehsan Soofi, two anonymous referees, and an associate editor.

Journal of the American Statistical Association
June 1996, Vol. 91, No. 434, Theory and Methods

GME formulation is specified, a solution for the unknown parameters is given and analytical comparisons are made with ME logit. In Sections 4 and 5 the information measure proposed by Soofi (1992) and Golan (1988) are specified and the dual loss function resulting from the GME formulation is analytically evaluated. In Section 6 the results of limited sampling experiments, reflecting the analytical results, are reported. A real-world example is presented in Section 7, and conclusions and implications are provided in Section 8. The proofs of propositions in the text are given in the Appendix.

2. MODEL REFORMULATION AND A ME SOLUTION

The ME approach to the multinomial choice problem (Denzau et al. 1989; Soofi 1992) is presented in this section. In addition, an alternative, "dual" presentation is offered along with the information matrix. Before specifying the ME approach, it is convenient to transform the observables, in line with (3), into a generalized moment condition consistent with the GME formulation. This is done in the next subsection.

2.1 Model Transformation

If the *unknown and unobservable* $\mathbf{p}$ and $\mathbf{e}$ in (3) are to be recovered, then the indirect empirical measurements on the noisy, observable $\mathbf{y}$ and the known covariates $\mathbf{x}_i$ must be used in formulating the problem. In the GME approach to the multinomial choice problem, this information is written as an ill-posed inverse problem with noise that is linear in $\mathbf{p}$:

$$(\mathbf{I}_J \otimes X')\mathbf{y} = (\mathbf{I}_J \otimes X')\mathbf{p} + (\mathbf{I}_J \otimes X')\mathbf{e}, \tag{4}$$

where X is $(N \times K)$ and there are NJ data points, y_{ij}. The problem is ill-posed because there are KJ *moment relations* (data points) and $NJ(> KJ)$ unknown multinomial parameters.

To solve the ill-posed problem (4), it is conventional to follow sampling theory or Bayesian procedures and impose parametric restrictions on the functional form linking the multinomial probabilities and the covariates (Albert and Chib 1993; Koop and Poirier 1993; McFadden 1974; Zellner and Rossi 1984). Under this alternative, if the statistical model is not correct or the data are partial or incomplete, then the consequences of imposing the strong restrictions required for estimation and testing can be considerable. The GME formalism avoids some of the potential problems of traditional approaches.

2.2 The Maximum Entropy Formalism

The basis for the entropy formulations developed here is contained in the work of Shannon (1948), Jaynes (1957a; 1957b), Kullback (1959), Gokhale and Kullback (1978), Levine (1980), Jaynes (1984), and Csiszár (1991). Using Shannon's (1948) entropy, measure

$$H(p) \equiv -\sum_j p_j \log_e p_j, \tag{5}$$

where p_j is the probability of observing outcome j. Jaynes (1957a,b) proposed, as a basis for assigning or recovering the unknown probabilities, maximizing the entropy measure (5) subject to appropriate information-moment relations and adding up (normalization) constraints for the probabilities. Thus under Jaynes's ME (uncertainty) principle, out of all those probability distributions consistent with what we know (the data), we choose the one that maximizes (5). Keeping the ME principle in mind, we turn to a reformulation of the multinomial choice problem.

Denzau et al. (1989) and Soofi (1992) applied Jaynes's ME approach to the multinomial choice problem. They showed that if $\mathbf{p}$ is to preserve the observed sums of the attribute scores for the chosen categories, for each of the $k = 1, 2, \ldots, K$ covariates, then (4) reduces to the ill-posed pure inverse problem

$$(\mathbf{I}_J \otimes X')\mathbf{y} = (\mathbf{I}_J \otimes X')\mathbf{p}. \tag{6}$$

Maximization of (5) subject to (6), the adding up–normalization condition is equivalent to minimization of the Kullback–Leibler cross-entropy function (relative to a uniform distribution) and subject to the same constraints. This is known as the Gokhale and Kullback (1978) internal constraint problem. Under this specification, the multinomial probabilities $\mathbf{p}$ cannot be determined by direct inversion of (6). The data points may be consistent with a variety of different $\mathbf{p}$. When there are multiple solutions, under the principle of maximum entropy, *out of the possible multinomial probabilities consistent with the data,* the maximally noncommital choice is to select the $\mathbf{p}$ with minimum information content or, equivalently, the ME. The pure inverse ME formulation is

$$\max_{\mathbf{p}} \; -\mathbf{p}'\log_e \mathbf{p}, \tag{7}$$

subject to the information-moment constraints

$$(\mathbf{I}_J \otimes X')\mathbf{y} = (\mathbf{I}_J \otimes X')\mathbf{p} \tag{8}$$

and the normalization (adding up) conditions

$$[\mathbf{I}_{N1} \quad \mathbf{I}_{N2} \quad \cdots \quad \mathbf{I}_{NJ}]\mathbf{p} = \mathbf{1}, \tag{9}$$

where $\mathbf{1}$ is a vector of unit values.

First-order conditions for the Lagrangian for the optimization problem (7)–(9) form a basis for recovering the unknown $\mathbf{p}_j$ and the Lagrange parameters $\boldsymbol{\lambda}_j$. The solution to this optimization problem is

$$p^*_{ij} = \frac{\exp(-\mathbf{x}_i'\boldsymbol{\lambda}^*_j)}{\Omega_i(\boldsymbol{\lambda}^*)} = \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta}^*_j)}{\Omega_i(\boldsymbol{\beta}^*)} = \frac{\exp(\boldsymbol{\beta}^{*\prime}_j\mathbf{x}_i)}{1 + \sum_{j=2}^{J}\exp(\boldsymbol{\beta}^{*\prime}_j\mathbf{x}_i)}, \tag{10}$$

where

$$\Omega_i(\boldsymbol{\beta}^*) \equiv \mathbf{1}'\exp(-\mathbf{x}_i'\boldsymbol{\lambda}^*_j) = \mathbf{1}'\exp(\mathbf{x}_i'\boldsymbol{\beta}^*_j)$$

$$= \sum_j \exp\left(\sum_k \boldsymbol{\beta}^*_{jk}x_{ik}\right) \tag{11}$$

and $\beta_j^* = -\lambda_j^*$. The $\Omega_i(\beta^*)$, which are called partition functions, are normalization factors. The unknown β_j that link the p_{ij} to the $\mathbf{x}_i$ are the KJ Lagrange parameters that are determined so that the optimum solution p_{ij}^* satisfy the constraints (8) and (9). For completeness and because we will need it later for comparison purposes, we give the following results. Given the Lagrangian and the corresponding first-order conditions, the Hessian is

$$\frac{\partial^2 L}{\partial p_{ij}\partial p_{ij}} = -\frac{\mathbf{1}'\exp(\mathbf{x}_i'\beta_j)}{\exp(\mathbf{x}_i'\beta_j)} = -\frac{1}{p_{ij}}, \tag{12}$$

where all off-diagonal elements are zero. The result is a negative definite Hessian matrix that ensures a unique global solution for the p_{ij}'s. We now present some new results for the ME. First, we derive the information matrix for the ME estimator. Second, we show an alternative, unconstrained formulation of the problem. Summing over N and rearranging the Hessian (12) yields

$$\mathbf{I}(\mathbf{p}_j)_{\mathrm{ME}} = \sum_{i=1}^{N} \frac{\mathbf{1}'\exp(\mathbf{x}_i'\beta_j)}{\exp(\mathbf{x}_i'\beta_j)}\,\mathbf{1}\mathbf{1}' = \sum_i \frac{1}{p_{ij}}\,\mathbf{1}\mathbf{1}'. \tag{13}$$

In going from the p_{ij} to the β_{jk} space, we make use of a generalization of a correspondence result of Lehmann (1983, p. 118), which yields

$$\begin{aligned}
&\left(\frac{\partial \mathbf{p}}{\partial \beta_j}\right)\mathbf{I}(\mathbf{p}_j)_{\mathrm{ME}}\left(\frac{\partial \mathbf{p}_j}{\partial \beta_j}\right)' \\
&\quad = \mathbf{I}(\beta_j)_{\mathrm{ME}} \\
&\quad = \sum_i \frac{\exp(\mathbf{x}_i'\beta_j)}{[\Omega_i(\beta_j)]^2}\,\mathbf{x}_i\mathbf{x}_i' \\
&\quad = \sum_i p_{ij}\,\frac{1}{\Omega_i(\beta_j)}\,\mathbf{x}_i\mathbf{x}_i' = \mathbf{I}(\beta_j)_{\mathrm{ML}},
\end{aligned} \tag{14}$$

where (14) is the jth diagonal block of J^2 blocks of dimension $(K \times K)$ and is identical to the ML information matrix for β. The asymptotic covariance matrix can be estimated using the inverse of (14) evaluated at the classical ME logit estimates. Thus, although the conceptual bases of the traditional ML multinomial logit and the classical ME formulation are different, and under the classical ME formulation no particular functional form linking the p_{ij} and the $\mathbf{x}_i'\beta_j$ was specified, the resulting ML logit and classical ME solutions and information matrices are equivalent, and the usual asymptotic properties follow. An intuitive explanation of the correspondence between the classical ME and ML logit solutions is that (a) the information-moment data constraints in the ME formulation are the ML logit first-order conditions, and (b) the ME solution resulting from the optimization (7)–(9) has the same mathematical form as the logistic multinomial probabilities. We now show this correspondence explicitly.

The classic ME approach can be usefully reformulated as an unconstrained problem. There are two advantages. First, it reduces computational complexity substantially. Second, it simplifies derivation of the asymptotic results and comparison with the likelihood function. Building on the Lagrangian and the solution (10), we may rewrite the ME problem (7)–(9) in an unconstrained dual form as

$$M(\lambda_j) = \mathbf{y}'(\mathbf{I} \otimes X')\lambda_j + \sum_i \log_e[\Omega_i(\boldsymbol{\lambda})], \tag{15}$$

which is the same as the multinomial logit log-likelihood function (Maddala 1983, p. 36),

$$\begin{aligned}
\log_e L &= \sum_i \sum_j y_{ij} \log_e p_{ij} \\
&= \sum_i \sum_j y_{ij} \log_e \left(\frac{\exp(\mathbf{x}_i\beta_j)}{\sum_j \exp(\mathbf{x}_i\beta_j)}\right) \\
&= \sum_i \sum_j y_{ij}\mathbf{x}_{ik}\beta_{jk} - \sum_i \log_e[\Omega_i(\beta_i)],
\end{aligned} \tag{16}$$

where $\beta = -\lambda$. Consequently, the usual asymptotic properties follow. We use this dual approach in the following analysis.

3. A GENERALIZED MAXIMUM ENTROPY FORMULATION

In the ME-ML solutions of Section 2.2, strong assumptions were needed to ensure that the information-moment relations (6) hold. Consistent with the uncertainty about the appropriate statistical model that normally holds in practice, it would seem more realistic to avoid the strong requirement of (6) in the ME approach or the strong logistic cdf assumption in the ML approach, and to work with the more general noise inverse information-moment relations (4). Further, because in (4), both $\mathbf{p}$ and $\mathbf{e}$ appear directly as unknown and unobservable components in the inverse relation, it is natural to pursue a dual objective rather than the single criterion used in the ML logit formulations.

3.1 Reparameterizing the Inverse Relation

If we are to use the information in the inverse-moment relation (4) within the entropy formalism, then the unknown and unobservable $\mathbf{p}$ and $\mathbf{e}$ must have the properties of probabilities. The elements of $\mathbf{p}$ are already in the form of probabilities; however, the realizations of the e_{ij} noise components from statistical model (2) may range over the natural interval $[-1, 1]$. Because the noise components are not in a probability form, we use as a conceptual device a reparameterization proposed by Judge (1991) and Judge, Golan, and Miller (1993) to transform the e_{ij}, which may take on values over the natural boundaries $[-1, 1]$, to probabilities that may take on values over the interval [0, 1]. *In terms of the random variable e_{ij}, we define over the interval* $[-1, 1]$ bounded discrete random variable with finite possible realizations $\mathbf{v}_{ij} = (\nu_{ij1}, \nu_{ij2}, \ldots, \nu_{ijM})'$ of dimension $M \geq 2$, with corresponding unknown weights $\mathbf{w}_{ij} = (w_{ij1}, w_{ij2}, \ldots, w_{ijM})'$ that have the properties of probabilities, where $\sum_m w_{ijm} = 1$ and $e_{ij} = \sum_m \nu_{ijm} w_{ijm}$. For example, if we let $M = 3$, then $\mathbf{v} = (-1, 0, 1)'$ and there exist w_1, w_2, and w_3 such that with positive probability, each

noise component can be written as $e_{ij} = w_1(-1) + w_3(1)$. Other than the assumption of M, no subjective information on the distribution of probabilities is assumed. A further discussion of the effect of M is given at the end of this section. Theoretical examples and an empirical example where the choice of M is investigated are given in Sections 6 and 7. Finally, if prior information on the probabilities $\mathbf{p}$ exists, it can be formulated within the cross-entropy context (Good 1963; Kullback 1959).

Consistent with (3) and the inverse relation (4), if we let

$$\mathbf{e}_j = V_j\mathbf{w}_j = \begin{bmatrix} \mathbf{v}'_{j1} & & & \\ & \mathbf{v}'_{j2} & & \\ & & \ddots & \\ & & & \mathbf{v}'_{jN} \end{bmatrix} \begin{bmatrix} \mathbf{w}_{j1} \\ \mathbf{w}_{j2} \\ \vdots \\ \mathbf{w}_{jN} \end{bmatrix}, \quad (17)$$

then

$$\sum_m \nu_{ijm} w_{ijm} = e_{ij} \quad \text{for} \quad i = 1, 2, \ldots, N; \quad j = 1, 2, \ldots, J. \quad (18)$$

Under this reparameterization, we may rewrite (4) as

$$(\mathbf{I}_J \otimes X')\mathbf{y} = (\mathbf{I}_J \otimes X')\mathbf{p} + (\mathbf{I}_J \otimes X')V\mathbf{w}, \quad (19)$$

where both the unknown $\mathbf{p}$ and $\mathbf{w}$ are in the form of probabilities.

3.2 The GME Formulation and Solution

Given the reparameterized inverse relation (19) involving the unknown and unobservable $\mathbf{p}$ and $\mathbf{w}$, and assuming *independence* between the two, the GME multinomial response problem may be stated as maximizing the dual objective function

$$\max_{\mathbf{p},\mathbf{w}} H(\mathbf{p}, \mathbf{w}) = \max_{\mathbf{p},\mathbf{w}} \{-\mathbf{p}'\log_e\mathbf{p} - \mathbf{w}'\log_e\mathbf{w}\}, \quad (20)$$

subject to the information-moment conditions (4),

$$(\mathbf{I}_J \otimes X')\mathbf{y} = (\mathbf{I}_J \otimes X')\mathbf{p} + (\mathbf{I}_J \otimes X')V\mathbf{w}, \quad (21)$$

and the normalization constraints

$$[\mathbf{I}_{N1} \quad \mathbf{I}_{N2} \quad \cdots \quad \mathbf{I}_{NJ}]\mathbf{p} = \mathbf{1} \quad \text{for} \quad i = 1, 2, \ldots, N \quad (22)$$

and

$$\mathbf{1}'\mathbf{w}_{ij} = 1 \quad \text{for} \quad i = 1, 2, \ldots, N \quad \text{and} \quad j = 1, 2, \ldots, J. \quad (23)$$

The corresponding Lagrangian is

$$\begin{aligned} L = {} & -\mathbf{p}'\log_e\mathbf{p} - \mathbf{w}'\log_e\mathbf{w} \\ & + \boldsymbol{\lambda}'[(\mathbf{I}_J \otimes X')\mathbf{y} - (\mathbf{I}_J \otimes X')\mathbf{p} - (\mathbf{I}_J \otimes X')V\mathbf{w}] \\ & + \boldsymbol{\rho}'\{\mathbf{1} - [\mathbf{I}_{N1}, \mathbf{I}_{N2}, \ldots, \mathbf{I}_{NJ}]\mathbf{p}\} + \boldsymbol{\delta}'(\mathbf{1} - \mathbf{1}'\mathbf{w}_{ij}). \end{aligned} \quad (24)$$

The first-order conditions for the Lagrangian form a basis for recovering the unknown $\mathbf{p}_j, \mathbf{e}_j$, and the Lagrange parameters $\boldsymbol{\beta}_j$. Solving the optimization problem, and noting that $\hat{\boldsymbol{\beta}}_j = -\hat{\boldsymbol{\lambda}}_j$, we obtain the solution

$$\hat{p}_{ij} = \frac{\exp(\mathbf{x}'_i\hat{\boldsymbol{\beta}}_j)}{\Omega_i(\hat{\boldsymbol{\beta}}_j)} = \frac{\exp(\mathbf{x}'_i\hat{\boldsymbol{\beta}}_j)}{\sum_j \exp(\mathbf{x}'_i\hat{\boldsymbol{\beta}}_j)} \quad (25)$$

and

$$\hat{w}_{ijm} = \frac{\exp(\mathbf{x}'_i\hat{\boldsymbol{\beta}}_j V_j)}{\Psi_{ij}(\hat{\boldsymbol{\beta}}_j)} = \frac{\exp(\mathbf{x}'_i\hat{\boldsymbol{\beta}}_j V_j)}{\sum_j \exp(\mathbf{x}'_i\hat{\boldsymbol{\beta}}_j V_j)}. \quad (26)$$

Finally, from (17),

$$\hat{\mathbf{e}}_j = V_j\hat{\mathbf{w}}_j. \quad (27)$$

The partition functions $\Omega_i(\hat{\boldsymbol{\beta}}_j)$ and $\Psi_{ij}(\hat{\boldsymbol{\beta}}_j)$ are normalization factors. The unknown $\boldsymbol{\beta}_j$, which link the p_{ij} to the $\mathbf{x}_i$, are the KJ Lagrange parameters, $-\hat{\boldsymbol{\lambda}}_j$, that are determined so that the optimum solution $\hat{p}_{ij}$ and $\hat{e}_{ij}$ satisfy the constraints (21)–(23). The solutions (25) and (26) are in exponential form, so $\hat{\mathbf{p}}_j$ and $\hat{\mathbf{w}}_j$ are always positive. The normalization factors ensure that $\hat{\mathbf{p}}_j$ and $\hat{\mathbf{w}}_j$ have the properties of probabilities and provide information on the *distribution of probabilities* for the unknown parameters.

As in Section 2, the recovered unknown parameters $\hat{p}_{ij}$ and $\hat{e}_{ij}$ are based only on the information in the information-moment relations, with no initial assumptions regarding the form of the function linking the p_{ij} and $\mathbf{x}'_i\boldsymbol{\beta}_j$. Because the $\hat{\boldsymbol{\lambda}}_j$ are not unique for both the ME and GME formulations, a common normalization procedure $\boldsymbol{\lambda}_1 = \boldsymbol{\beta}_1 = \mathbf{0}$ is imposed, where $\mathbf{0}$ is a $(K \times 1)$ zero vector.

In the GME formulation, the unknowns $\mathbf{p}$ and $\mathbf{w}$ are determined jointly so as to maximize (20) subject to constraints (21)–(23). Any prior knowledge concerning the multinomial probabilities p_{ij} can be added using a cross-entropy approach. *The GME moment-consistency relation (21) is much less restrictive than the ME and ML-multinomial logit moment relation (6) and solution (10).* By relaxing the ML-pure ME information-moment conditions (6) to the information-moment conditions (21), a larger set of solutions exist for the GME formulation that includes those consistent with (8). Consequently, the solution requirements for the $\mathbf{p}_j$ for GME are less binding. As a result, it is possible for the recovered $\mathbf{p}_j$ to exhibit more uniformity than for the ML logit-ME solution.

Just as the ME is related to the ML approach, Equations (15) and (16), the GME is related to a class of generalized logit (exponential family) formulations. Consider the generalized log-likelihood function, which is just the dual GME developed in the Appendix:

$$\begin{aligned} \log_e L = {} & \sum_i \sum_j y_{ij} x_{ik} \beta_{jk} \\ & - \sum_i \log_e \left[\sum_j \exp(\mathbf{x}'_i\boldsymbol{\beta}_j) \right] \\ & - \sum_i \sum_j \log_e \left[\sum_m \exp(\mathbf{x}'_i\boldsymbol{\beta}_j V_j) \right], \end{aligned} \quad (28)$$

where, consistent with the GME model (20)–(23), the right side term corresponds to the noise component. Maximizing this generalized log-likelihood function yields

$$\frac{\partial \log_e(L_{\mathrm{GL}})}{\partial \boldsymbol{\beta}_j} = \sum_i y_{ij} x_{ik} - \sum_i \left[\frac{\exp(\mathbf{x}_i'\boldsymbol{\beta}_j)}{\Omega_i(\boldsymbol{\beta}_j)} \mathbf{x}_i + \sum_m \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta}_j V_j)}{\psi_{ij}} V\mathbf{x}_i \right], \quad (29)$$

which are just the information-moment constraints (21). In other words, from (28) it is apparent that the dual GME formulation is a product of two logits, (25) for the p_{ij} and (26) for the w_{ijm} that share a common parameter $\boldsymbol{\beta}_j$. As such, the GME gains its advantage over the ME because of its duality with two prespecified logit models that are joined and based on the independence assumption. *However, as is discussed and shown in the sampling experiments and the empirical Sections 6 and 7, this complication does not affect the computational burden. On the contrary, in most cases the dual GME converges to its optimal solution faster than the ME or ML logit.*

To obtain a GME estimate, it is sufficient to have two points, $M = 2$, in the support of $\mathbf{v}_{ij}$, which converts the errors from $[-1, 1]$ into [0, 1] space. But we have found that using a larger M results in superior estimates in terms of all standard criteria. The larger the M, the more moments captured in the optimization estimation process. If, for example, we use $M = 2$, then we capture only the first moment of the unknown errors. But if we use $M = 7$, we recover the first six moments of the errors, so the estimates of the unknown probabilities and coefficients improve. As a practical matter, we have found a substantial improvement in the estimates gained by increasing M from 2 to 3. For larger M, the improvement diminishes. Typically, little improvement is realized from increasing M beyond about 7 or 9.

3.3 Analytical Results

Given the Lagrangian (24), the distribution implied by (25) and (26) and the information-moment relations (25), the elements of the Hessian for the GME are

$$\begin{aligned} \frac{\partial^2 L}{\partial p_{ij}\partial p_{ij}} + \sum_m \frac{\partial^2 L}{\partial w_{ijm}\partial w_{ijm}} &= -\frac{\mathbf{1}'\exp(\mathbf{x}_i'\boldsymbol{\beta}_j)}{\exp(\mathbf{x}_i'\boldsymbol{\beta}_j)} - \sum_m \frac{\mathbf{1}'\exp(\mathbf{x}_i'\boldsymbol{\beta}_j V_j)}{\exp(\mathbf{x}_i'\boldsymbol{\beta}_j V_j)} \\ &= -\frac{1}{p_{ij}} - \sum_m \frac{1}{w_{ijm}}, \end{aligned} \quad (30)$$

where all off-diagonal elements are zero.

Because we have a negative definite Hessian matrix, there is a unique global solution for the p_{ij}'s. Summing over N and rearranging (30), we obtain the matrix

$$\begin{aligned} \mathbf{I}(\mathbf{p}, \mathbf{w})_{\mathrm{GME}} &= \sum_i \frac{\mathbf{1}'\exp(\mathbf{x}_i'\boldsymbol{\beta}_j)}{\exp(\mathbf{x}_i'\boldsymbol{\beta}_j)} \mathbf{1}\mathbf{1}' + \sum_i \sum_m \frac{\mathbf{1}'\exp(\mathbf{x}_i'\boldsymbol{\beta}_j V_j)}{\exp(\mathbf{x}_i'\boldsymbol{\beta}_j V_j)} \mathbf{1}\mathbf{1}' \\ &= \sum_i \frac{1}{p_{ij}} \mathbf{1}\mathbf{1}' + \sum_i \sum_m \frac{1}{w_{ijm}} \mathbf{1}\mathbf{1}'. \end{aligned} \quad (31)$$

Again making use of a generalization of a correspondence result of Lehmann (1983),

$$\begin{aligned} \frac{\partial(\mathbf{p}, \mathbf{w})}{\partial \boldsymbol{\beta}} \mathbf{I}(\mathbf{p}, \mathbf{w})_{\mathrm{GME}} \left[\frac{\partial(\mathbf{p}, \mathbf{w})}{\partial \boldsymbol{\beta}_j}\right]' &= \mathbf{I}(\boldsymbol{\beta}_j)_{\mathrm{GME}} \\ &= \sum_i \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta}_j)\mathbf{x}_i\mathbf{x}_i'}{[\Omega_i(\boldsymbol{\beta}_j)]^2} + \sum_i \sum_m \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta}_j V_j)}{[\Psi_{ij}(\boldsymbol{\beta}_j)]^2} (\mathbf{x}_i V_j)(\mathbf{x}_i V_j)' \\ &= \sum_i p_{ij} \frac{1}{\Omega_i(\boldsymbol{\beta}_j)} \mathbf{x}_i\mathbf{x}_i' + \sum_i \sum_m w_{ijm} \frac{1}{\Psi_{ij}(\boldsymbol{\beta}_j)} (\mathbf{x}_i V_j)(\mathbf{x}_i V_j)', \end{aligned} \quad (32)$$

where (32) is the jth diagonal block of J^2 blocks of dimension $(K \times K)$. Differentiating twice the generalized logit log-likelihood function (28) with respect to $\boldsymbol{\beta}$, we obtain an information matrix $\mathbf{I}(\boldsymbol{\beta}_j)$ that is equivalent to (32).

Given the information matrices (13) and (31) or, equivalently, (14) and (32), it is now possible to make an inferential comparison of the GME and ML logit-ME estimation rules. For expository purposes and without lack of generality, we consider the case of an orthonormal X where $X'X = \mathbf{I}_K$. (When $X'X \neq \mathbf{I}$, the characteristic roots of $X'X$ enter into the inequality comparisons.) We noted in Section 2.3 that $\mathbf{I}(\boldsymbol{\beta}_j)_{\mathrm{ME}} = \mathbf{I}(\boldsymbol{\beta}_j)_{\mathrm{ML}}$. We now focus on a comparison of the diagonal blocks of the ME and GME information matrices (14) and (32) and the corresponding covariance matrices.

Proposition 1. At the limit, $\hat{\mathbf{p}}$ is a consistent estimator of $\mathbf{p}$ or $\hat{\boldsymbol{\beta}} = f(\hat{\boldsymbol{\lambda}})$ is a consistent estimator of $\boldsymbol{\beta}$ and the ME-ML logit solution vector is equivalent to the GME solution vector and the asymptotic $\mathrm{var}(\hat{\boldsymbol{\beta}}_j)_{\mathrm{GME}}$ = asymptotic $\mathrm{var}(\boldsymbol{\beta}_j^*)_{\mathrm{ME\text{-}ML}}$.

Consequently, the usual, traditional asymptotic properties for the ML logit estimator follow for the GME estimator. As a consequence of Proposition 1, there is an inverse relationship between the boundaries of V_j and N. This inverse relationship follows from (A.3)–(A.5), where a choice for the boundaries of V_j is $\pm 1/\sqrt{N}$. For example, for $N = 100$ and $M = 3, \nu_{ij} = (-.1, 0, .1)'$. In terms of (32), as $N \to \infty$, the second element of the right side approaches zero and the first element on the right side approaches $\mathbf{I}(\boldsymbol{\beta}_j)_{\mathrm{ME}}$.

Proposition 2. For a given X and N, and for all $0 < |\beta_{jk}| \ll \infty$, the $\text{var}(\hat{\boldsymbol{\beta}}_j)_{\text{GME}} \leq \text{var}(\boldsymbol{\beta}_j^*)_{\text{ME}} = \text{var}(\boldsymbol{\beta}_j^*)_{\text{ML logit}}$.

The following proposition identifies the different rates of convergence for GME and ME-ML logit estimates.

Proposition 3. For a given $X, 0 < N < \infty$ and $M < \infty, \mathbf{I}(\beta_{jk})_{\text{GME}} \geq (1 + J^2/M)\mathbf{I}(\beta_{jk})_{\text{ME-ML logit}}$ for all j, k or, equivalently, $(1 + J^2/M)\text{var}(\hat{\beta}_{jk})_{\text{GME}} \leq \text{var}(\beta_{jk}^*)_{\text{ME-ML logit}}$, for all j, k, where J is the number of categories and N is the number of observations.

Proposition 3 shows that for any finite-sample size N, the GME estimator converges at least $(1 + J^2/M)$ faster than the ME-ML logit estimator. As J increases, the difference in the rate of convergence increases by a factor of J^2/M. For example, in the case of an orthonormal X, $J = 3$ and $M = 2$, the variance of the GME estimator with $N = 30$, is at least as small as the ME-ML logit estimator with $N = 165$.

3.4 An Extension of the GME Estimator

We can reformulate the GME model to obtain a discrete probability distribution to be specified for each of the p_{ij}. We specify for each p_{ij} a discrete probability distribution defined over the parameter space $[0, 1]$ by a set of discrete points $\mathbf{z} = (z_1, z_2, \ldots, z_D)'$ with corresponding probabilities π_{ijd} for $d = 1, 2, \ldots, D$. Under this formulation, the GME problem may be posed as

$$\max_{\pi_{ijd}, w_{ijm}} -\sum_i \sum_j \sum_d \pi_{ijd} \log_e(\pi_{ijd}) - \sum_i \sum_j \sum_m w_{ijm} \log_e(w_{ijm}) \quad (33)$$

subject to (21)–(23) and

$$\sum_d \pi_{ijd} = 1 \quad \forall\, i, j \quad (34)$$

and

$$\sum_j \sum_d \pi_{ijd} z_{ijd} = 1 \quad \forall\, i, j \quad (35)$$

Under this extended formulation, the solution $\tilde{\pi}_{ijd}$ provides a probability measure over the D-dimensional parameter space $\mathbf{z}$ for each p_{ij}, where $\tilde{p}_{ij} = \sum_d \tilde{\pi}_{ijd} z_{ijd}$.

4. INFORMATION MEASURES

Golan (1988) and Soofi (1992, 1994) have suggested a basis for measuring the importance of the contribution of each piece of data or constraint in the reduction of uncertainty. In terms of the GME formulation, the maximum possible entropy of the multinomial choice probabilities results when the information-moment relations (21) are not enforced and the distribution of probabilities over each choice set is uniform. As we add each piece of effective data (information-moment constraints), a departure from the uniform distributions results and a reduction in uncertainty occurs.

Following Golan (1988) and Soofi (1992, 1994), the proportion of the remaining total uncertainty may be measured in the case of multinomial probabilities, p_{ij}, by the normalized entropy,

$$S(\hat{\mathbf{p}}) = \left[-\sum_i \sum_j \hat{p}_{ij} \log_e \hat{p}_{ij}\right] \Big/ (\log_e(J) \cdot N) = -\mathbf{p}' \log_e \mathbf{p} / (\log_e(J) \cdot N), \quad (36)$$

where $S(\hat{\mathbf{p}}) \in [0, 1]$, or the *reduction in uncertainty information index* (Soofi 1992) $\mathbf{I}(\hat{\mathbf{p}}) = 1 - S(\hat{\mathbf{p}})$, where $\log_e(J) \cdot N$ represents maximum uncertainty and provides a basis for gauging the informational content. An $S(\hat{\mathbf{p}}) = 0$ implies no uncertainty, whereas $S(\hat{\mathbf{p}}) = 1$, which means that p_{ij} is uniform for all i and j, implies perfect uncertainty. Because $S(\hat{\mathbf{p}})$ is a relative measure of uncertainty, it can be used to compare different cases or scenarios. For example, an attribute (covariate) k can be eliminated and $S(K)$ can be compared to $S(K - 1)$. If both are equal, then we can conclude, based on the data, that attribute k introduces no additional information and does not help reduce the level of uncertainty concerning the unknown p_{ij}. A similar measure of normalized entropy for $\hat{\mathbf{w}}$, the other element in the objective function (20) is

$$S(\hat{\mathbf{w}}) = -\hat{\mathbf{w}}' \log_e \hat{\mathbf{w}} / (\log_e(M) \cdot N \cdot J). \quad (37)$$

Given the information measures (36) and (37), and the conclusion of Section 3.3, we can say the following. As the constraints (8) are relaxed to be consistent with (21), $|\beta_{jk}|$ approaches zero and the p_{ij} and w_{ijm} approach uniformity. Correspondingly, $S(\hat{\mathbf{p}})$ and $S(\hat{\mathbf{w}})$ approach 1. Consequently, $|\beta_{jk}|_{\text{GME}} \leq |\beta_{jk}|_{\text{ME-ML logit}}$ for all j, k, and $S(\hat{\mathbf{p}})_{\text{GME}} \geq S(\hat{\mathbf{p}})_{\text{ME-ML logit}}$.

Finally, within the extended formulation of Section 3.4, the normalized entropy is

$$S(\tilde{\pi}) = -\sum_i \sum_j \sum_d \tilde{\pi}_{ijd} \log_e \tilde{\pi}_{ijd} \Big/ (J \cdot N \cdot \log_e(D)), \quad (38)$$

whereas the normalized entropy for each p_{ij} is

$$S(\tilde{p}_{ij}) = -\sum_d \tilde{\pi}_{ijd} \log_e \tilde{\pi}_{ijd} \Big/ (\log_e(D)). \quad (39)$$

5. A DUAL CRITERION

The dual objective function (20) identifies the unknown signal and noise components in (3). As a result, in the GME approach estimates of the unknown $\mathbf{p}_j$ and $\mathbf{e}_j$ are jointly determined. In (20) the prediction and estimation objectives are both identified and, in this case, balanced or equally weighted.

The idea and use of a dual balanced loss function has been discussed by Zellner (1993) within a Bayesian context. In the case of ill-posed statistical models such as (4),

it is, within a sampling theory context, traditional to use the method of regularization to provide a solution based on both the data and a penalty function (O'Sullivan 1986; Tikhonov and Arsenin 1977). A regularization or smoothing parameter is then chosen to make use of the data and provide a stable solution. Thus in both sampling theory and Bayesian inference, there is a basis for using formulations that involve dual loss functions to depict the prediction-estimation precision trade-off.

The GME approach does not require a balanced dual-criterion function (20). If the prediction-estimation precision objectives for a particular problem warrants, then a weighted criterion function with $\gamma \in (0,1)$ may be specified. For example, for $\gamma \in (0,1)$, we can rewrite the problem as

$$\max_{\mathbf{p},\mathbf{w}} \; -(1-\gamma)\mathbf{p}'\log_e\mathbf{p} - \gamma\mathbf{w}'\log_e\mathbf{w} \tag{40}$$

subject to constraints (21)–(23) and call this GME(γ). Changing γ changes the prediction or estimation emphasis, and the choice of γ determines this trade-off. The recovered λ_{jk} are, under the dual objective (20), the recovered Lagrange parameters, and for $\gamma \in (0,1)$, $\hat{\beta}_{jk} = -\hat{\lambda}_{jk}(1/(1-\gamma))$. Similarly, the dual unconstrained GME problem is reformulated as

$$\begin{aligned} M(\boldsymbol{\lambda}_j,\gamma) &= y'(\mathbf{I}_j \otimes X)\boldsymbol{\lambda} \\ &\quad + (1-\gamma)\sum_i \log_e\left[\sum_j \exp\left(-\frac{1}{1-\gamma}\,\mathbf{x}_i'\boldsymbol{\lambda}_j\right)\right] \\ &\quad + \gamma\sum_i\sum_j \log_e\left[\sum_m \exp\left(-\frac{1}{\gamma}\,\mathbf{x}_i'\boldsymbol{\lambda}_j V_j\right)\right]. \end{aligned} \tag{41}$$

Given the conclusions of Section 3.3 and the information measures (36) and (37), four conclusions about the impact of the choice of γ can be made:

1. The $\text{var}(\hat{\boldsymbol{\beta}}_j)_{\text{GME}(\gamma)} \leq \text{var}(\boldsymbol{\beta}_j^*)_{\text{ME,ML logit}}$ for all $\gamma \in (0,1)$. This conclusion follows because $|\hat{\beta}_{jk}|_{\text{GME}(\gamma)} \leq |\beta_{jk}^*|_{\text{ME,ML logit}}$ for all j,k and $\gamma \in (0,1)$. Consequently, the first element on the right side of (32) is larger than the right side of (14) and the second element of the right side of (32) is nonnegative. Therefore, Proposition 1 holds for all $\gamma \in (0,1)$.
2. $S(\hat{\mathbf{p}})_{\text{GME}(\gamma)} \geq S(\hat{\mathbf{p}})_{\text{ME-ML logit}}$ for $\gamma \in (0,1)$.
3. As $\gamma \in (0,1)$ increases, the information measure $S(\hat{\mathbf{p}})$ decreases and the information measure $S(\hat{\mathbf{w}})$ increases. This negative relationship occurs because as γ increases, the relative weight on the $\mathbf{p}_j$'s decreases and the relative weight on the $\mathbf{w}_j$'s increases. Consequently, the information-moment relations (21) become more restricted and, simultaneously, more restricted consistency relations imply that the $\mathbf{p}_j$'s must be relatively less uniform, meaning $S(\hat{\mathbf{p}}(\gamma))$ must decrease, and as the information-moment relations are more restricted, the w_{ijm} moves toward uniformity ($e_{ij} \to 0$), and this means that $S(\hat{\mathbf{w}}(\gamma))$ must increase. Intuitively, the process is as follows: As $\gamma \to 1$, the information-moment relations become more restricted and the GME(γ) $\to$ ME-ML logit. Consequently, the $|\beta_{jk}|$ increase and the w_{ijm} are more uniform (high $S(\hat{\mathbf{w}})$).
4. For any sample of data, there is a unique $\gamma \in (0,1)$ that corresponds simultaneously to the $\min\{\sum_j\sum_k \text{var}(\hat{\beta}_{jk})\}$ and $\max\{S = S(\hat{\mathbf{p}}) + S(\hat{\mathbf{w}})\}$.

The proof for statement (4) is as follows: In statement (3) we showed that as $\gamma \in (0,1)$ increases, $S(\hat{\mathbf{p}})$ decreases and $S(\hat{\mathbf{w}})$ increases. The result that $\max\{S(\gamma)\}$ corresponds to $\min\{\sum_j\sum_k \text{var}(\hat{\beta}_{jk}(\gamma))\}$ is a direct consequence of (31) or (32) along with (36)–(37). As $\mathbf{p}_j$ goes to uniformity, or $|\beta_{jk}| \to 0$, the first component on the right side of (32) approaches its upper bound. Similarly, as $\mathbf{w}_{ij}$ approaches uniformity, the second component of the right side of (32) approaches its upper bound. An upper bound of (32) implies a lower bound for $\text{var}(\hat{\boldsymbol{\beta}}_j)$. Further, as p_{ij} and w_{ijm} approach uniformity, $S \to 2$, its upper bound. Consequently, the lowest $\sum_j\sum_k \text{var}(\hat{\beta}_{jk}(\gamma))$ must correspond to the $\max\{S(\gamma)\}$.

Combining this result with statement 3 implies that, given the negative relationship between $S(\hat{\mathbf{p}}(\gamma))$ and $S(\hat{\mathbf{w}}(\gamma))$ along with the negative relationship between $S(\gamma)$ and $\sum_j\sum_k \text{var}(\hat{\beta}_{jk}(\gamma))$, there must exist a unique $\gamma \in (0,1)$ that simultaneously maximizes $S(\gamma)$ and minimizes $\sum_j\sum_k \text{var}(\hat{\beta}_{jk}(\gamma))$.

In other words, from statement 3, there is a unique γ that maximizes $S(\gamma)$. From the first part of statement 4, the $\max\{S(\gamma)\}$ corresponds to the $\min\{\sum_j\sum_k \text{var}(\hat{\beta}_{jk}(\gamma))\}$; consequently, there is a unique γ that simultaneously yields a $\max\{S(\gamma)\}$ and $\min\{\sum_j\sum_k \text{var}(\hat{\beta}_{jk}(\gamma))\}$.

For a given data sample, if precision in the estimation of $\boldsymbol{\beta}_j$ is the objective, then a data-based choice of γ is provided by $\max S(\gamma \in (0,1))$.

6. SAMPLING EXPERIMENTS

The analytical results of Sections 3.3, 4, and 5 provide conditions under which the GME estimators will be well behaved in large samples and give a basis for evaluating and comparing the performance of GME(γ) relative to ME-ML logit. To indicate the nature of the output of the ME and GME(γ) formulation and provide an empirical basis for comparing the small-sample performance of the GME(γ) and ME or traditional ML multinomial logit procedures, we present the results of Monte Carlo (MC) sampling experiments.

6.1 Design of the Sampling Experiment

Consider a multinomial response problem involving three choice categories, $j = 0,1,2$, and four covariates, x_{ik}, $k = 1,2,3,4$. The choice response coefficients β_{jk} in the linkage function $p_{ij} = G(\mathbf{x}_i'\boldsymbol{\beta}_j)$ are $\boldsymbol{\beta}_{1k} = (0,0,0,0)'$, $\boldsymbol{\beta}_{2k} = (-1,1,2,-1)'$, and $\boldsymbol{\beta}_{3k} = (1,-1,-2,1)'$. The x_{ik}, except for the constant x_{i1}, were generated from a multivariate normal $(\mathbf{0},\mathbf{I}_N)$.

Given the $\mathbf{x}_i$ and $\boldsymbol{\beta}_j$, the relationship between the multinomial choice probabilities and the attribute vectors is specified to be of the *logistic* form $G(\mathbf{x}_i'\boldsymbol{\beta}_j)$,

where $p_{ij} = G(\mathbf{x}_i'\boldsymbol{\beta}_j) > 0, \sum_j G(\mathbf{x}_i'\boldsymbol{\beta}_j) = 1$. The p_{ij} are used to calculate the probability that $y_{ij} = 1$ for $i = 1, 2, \ldots, N$ and $j = 0, 1, 2$. The value of y_{ij} is determined, following Griffiths, Hill, and Pope (1987), by drawing a uniform random number on the unit interval to assign an observation to a category. For example, if for $\mathbf{x}_i'\boldsymbol{\beta}_j$, the generated proportions are $p_{i0} = .5$, $p_{i1} = .3$ and $p_{i2} = .2$, then a random draw between $[0, .5)$ is assigned to category zero, a random draw between $[.5, .8)$ is assigned to category 1, and a random draw between $[.8, 1.0]$ is assigned to category 2.

Using this sample design, 500 samples each of size $N = 30$ and $N = 100$ were generated. Under a logistic cdf specification, the traditional ML procedure or the ME procedure (7)–(9) was used to recover the β_{jk} and to predict the choice outcomes. Alternatively, the more general GME procedure using formulation (40) and (21)–(23), or the dual (41), was used to recover the unknown p_{ij}, e_{ij}, and β_{jk} and to predict the choice outcomes. In the GME formulation, for the $N = 100$ case and consistent with (17), to reparameterize the e_{ij}, we used the discrete set of points $v_{ijm} = [-.1, 0, .1]$ that is consistent with the convergence rate $1/\sqrt{N}$ discussed in Section 3.3.

6.2 Results of a Sampling Experiment

The empirical mean squared error $\text{MSE}(\hat{\boldsymbol{\beta}}) = \sum_j \sum_k (\hat{\beta}_{jk} - \beta_{jk})^2/(J-1)K$ and the mean of the normalized entropy measures $S(\hat{\mathbf{p}})$ and $S(\hat{\mathbf{w}})$ from 500 replications of a sampling experiment, using the design discussed in Section 6.1, and $N = 100$, are presented in Table 1. For comparison purposes, five choices of γ are reported. Soofi (1992) recommended using $\mathbf{I}(\hat{\mathbf{p}}) = 1 - S(\hat{\mathbf{p}})$ as the appropriate information measure. To be consistent with him, we report $\mathbf{I}(\hat{\mathbf{p}})$ and $\mathbf{I}(\hat{\mathbf{w}}) = 1 - S(\hat{\mathbf{w}})$ in our tables.

Consistent with the analytical propositions in Section 3.3, the $\text{MSE}(\hat{\boldsymbol{\beta}})$ results from the sampling experiment reflect in general the superior finite-sample performance characteristics of the GME(γ) estimator relative to the pure ME-ML estimators for all $\gamma \in (0, 1)$. In this experiment the $\text{MSE}(\boldsymbol{\beta}^*)$ is approximately three times larger than $\text{MSE}(\hat{\boldsymbol{\beta}})$ for $\gamma = .5$. Relative to ML logit, a balanced loss objective ($\gamma = .5$) *loses very little in prediction* (the average number of misses is virtually the same and the maximum number of misses is slightly lower) and *gains substantially in estimation precision.* In terms of the information measure $S(\hat{\mathbf{p}})$, and consistent with Sections 3.3 and 4, the empirical $S(\hat{\mathbf{p}})$ for $\gamma \in (.1, .9)$ are larger than $S(\mathbf{p}^*)$, so $\mathbf{I}(\hat{\mathbf{p}}) < \mathbf{I}(\mathbf{p}*)$.

Table 1. Results of a Sampling Experiment, 500 Replications, N = 100

Estimator	Misses			MSE($\hat{\beta}$)	I($\hat{\mathbf{p}}$)	I($\hat{\mathbf{w}}$)
	Min	Average	Max			
GME ($\gamma = .1$)	12	21.3	31	.410	.330	.180
GME ($\gamma = .3$)	11	21.0	29	.218	.430	.035
GME ($\gamma = .5$)	10	21.3	32	.183	.469	.011
GME ($\gamma = .7$)	11	20.9	30	.205	.508	.003
GME ($\gamma = .9$)	13	20.8	32	.357	.533	.001
ME-ML logit	10	21.1	32	.555	.542	

Further, in line with Section 5, as γ increases, the empirical $S(\hat{\mathbf{p}})$ decreases and $S(\hat{\mathbf{w}})$ increases or, similarly, $\mathbf{I}(\hat{\mathbf{p}})$ increases and $\mathbf{I}(\hat{\mathbf{w}})$ decreases. In the experiment, the maximum $S(\gamma) = S(\hat{\mathbf{p}}) + S(\hat{\mathbf{w}})$ and the minimum empirical variance occur when $\gamma = .3$.

In much applied work, the primary emphasis is on recovering and using estimates of the response coefficients β_{jk}. Consequently, the possibility of using the GME(γ) estimator to gain precision for the $\hat{\beta}_{jk}$ without sacrificing category prediction is attractive. Other replicated sampling experiments, which are not reported, yielded essentially the same performance characteristics for the ME, ML logit, and GME(γ) estimators as those reported in Table 1.

In Table 1 and our other simulation experiments, we used $M = 3$. The advantages of using a relatively small M are that obtaining the estimates take less time, which is useful when running a large number of simulations, and it shows a worst-case scenario for GME. How much better would GME have performed had we used a larger M? To illustrate the importance of the size of M, we ran 500 replications of the GME for $\gamma = .5$ using $M = 7$ for the problem in Table 1. The $\text{MSE}(\hat{\boldsymbol{\beta}})$ dropped from .183 for $M = 3$ to .176, and the average misses fell to 20.85. The minimum misses rose to 11; the maximum misses, to 33. The new average $\mathbf{I}(\hat{\mathbf{p}})$ is .494, and the $\mathbf{I}(\hat{\mathbf{w}})$ is .005. In short, when we use a larger M, the advantages of the GME in our simulation experiments are more pronounced. Finally, we note that the GME with $M = 7$ outperforms the ME-ML estimator in terms of both precision (MSE) and prediction (number of misses) in these simulations.

To reflect the sampling consequences of a large sample of observations, an experiment identical to the one reported in Table 1 was completed with $N = 500$. The results for ME-ML are $\text{MSE}(\boldsymbol{\beta}^*) = .060$, $\text{var}(\boldsymbol{\beta}^*) = .059$, average misses $= 116.5$, and $\mathbf{I}(\mathbf{p}^*) = .499$; for GME ($\gamma = .5$), $\text{MSE}(\hat{\boldsymbol{\beta}}) = .050$, $\text{var}(\hat{\boldsymbol{\beta}}) = .042$, average misses $= 116.5$, and $\mathbf{I}(\hat{\mathbf{p}}) = .480$. Thus, even in large samples, the GME performs better.

6.3 Results with Partial and Ill-conditioned Data

Much of the experience with ME formulations (e.g., Donoho, Johnstone, Hoch, and Stern 1992; Judge et al. 1993) suggests that in the case of ill-posed inverse problems, the performance of the ME approach is superior to that of traditional sampling theoretic methods. For a range of statistical models, in the case of a small number of observations and a large number of covariates and/or an ill-conditioned covariate design matrix, when using traditional estimation procedures (O'Sullivan 1986), it is expected that the parameter estimates will be highly unstable, giving rise to high variance. For example, Le Cessie and van Houwelingen (1992) and Griffiths et al. (1987) discussed and investigated design matrix problems in the case of multinomial response data. Hastie, Buja, and Tibshirani (1993) discussed the case of penalized discriminant analysis in the case of many highly correlated predictors. The inferential impact of the attribute-design matrix is reflected in the information matrices for the ME-ML logit and GME estimators

Table 2. Results of a Sampling Experiment, 500 Replications and N = 30

Estimator	Misses: Min	Misses: Average	Misses: Max	MSE($\hat{\beta}$)	I($\hat{\mathbf{p}}$)
ME-ML logit	0	4.57	9	288,402	.745
GME ($\gamma = .5$)	0	5.27	13	.466	.462

developed in Sections 2.2 and 3.2. To support the analytical results, Monte Carlo experiments are reported in Sections 6.3.1 and 6.3.2 that illustrate the sampling consequences in finite samples of using, in the case of multinomial response data, a small number of observations and/or an ill-conditioned covariate matrix.

6.3.1 Results for $N = 30$. To reflect the sampling consequences of a small number of observations, we use the three-category, four-covariate problem and the data-generation scheme of Section 6.1. The results from 500 replications of a sampling experiment with samples of size $N = 30$ are reported in Table 2. When $N = 30$, the very large empirical MSE(β^*) for the ME-ML logit estimator reflects the unstable nature of the solution and the computational difficulties of the ML approach for some of the samples. As in the linear statistical model case, the ML logit predicts well (4.57 misses) compared to the GME (5.27 misses). Alternatively, consistent with the results of Section 3.2, the MSE($\hat{\beta}$) for GME ($\gamma = .5$) was relatively small, reflecting results that were stable from sample to sample. The MSE($\hat{\beta}$) results for the GME ($\gamma = .5$) estimator are consistent with those of Table 1 for $N = 100$ and the analytical results of Sections 3.3 and 5. A limited number of data points appear to have had a limited effect on sampling performance in the case of GME ($\gamma = .5$).

6.3.2 Results for an Ill-Conditioned Design Matrix. Using the three-category, four-covariate design of Section 6.1, to measure the degree of collinearity in the design matrix we use the condition number $\kappa(X'X) = \delta_D/\delta_S$ which is the ratio of the largest and smallest roots (Belsley 1991). As the degree of collinearity increases, $\delta_S \to 0$ and $\kappa(X'X) \to \infty$. To reflect an ill-designed experiment consistent with significant collinearity in the design matrix and the nature of much nonexperimental data, we use the condition number $\kappa(X'X) = 90$. In this experiment $\beta_{1K} = (0, 0, 0, 0)'$, $\beta_{2K} = (-1, 1, 2, -1)'$, and $\beta_{3K} = (-1.4, 1.5, 1, -1)'$. The results from 500 replications of samples of size $N = 100$ and GME ($\gamma = .1, .5, .9$) are presented in Table 3. Despite the high degree of collinearity and an unfavorable signal-to-noise ratio, the GME ($\gamma \in (.1, .9)$) estimator performs well based on a squared error loss measure relative to β. In contrast, the traditional multinomial logit estimator is highly unstable, and the empirical MSE($\hat{\beta}^*$) exceeded the empirical MSE($\hat{\beta}$) for GME ($\gamma \leq .5$) by a factor of 20 or more, with a slightly superior predictive ability.

Finally we present in Table 4 the results of a sampling experiment involving a *small* number of data points, $N = 30$, and a design matrix with a *high condition number* of $\kappa(X'X) = 90$. Despite an ill-conditioned design matrix

Table 3. Results of a Sampling Experiment With 500 Replications, N = 100 and $\kappa(X'X) = 90$

Estimator	Misses: Min	Misses: Average	Misses: Max	MSE($\hat{\beta}$)	I($\hat{\mathbf{p}}$)
GME ($\gamma = .1$)	48	57.26	70	1.661	.001
GME ($\gamma = .5$)	47	56.99	72	4.373	.003
GME ($\gamma = .9$)	48	56.22	70	33.387	.021
ME-ML logit	44	56.10	70	85.240	.044

and a small number of data points, in general the empirical MSE($\hat{\beta}$) and prediction results for GME ($\gamma \in (.1, .9)$) tell the same story relative to ML logit, as in the $\kappa(X'X) = 90$ and $N = 100$ cases. For multinomial response data from an ill-conditioned design, the GME ($\gamma < .5$) estimator appears to be a winner if the focus is on a squared error measure for the response coefficients. As expected, as the number of data points decreases, the empirical MSE(β^*) for the traditional multinomial logit estimator increases; however, as in the well-posed case, reducing the number of observations does not appear to significantly affect the GME(γ) estimator performance.

Given the results of Sections 2.3 and 3.3, to illustrate the difference in the ML logit and GME estimated covariance matrices we use a single sample from the $N = 30, \kappa(X'X) = 90$ sampling experiment. For category 2, the ML logit and GME ($\gamma = .5$) estimated covariance matrices are given in Table 5. As suggested by the analytical results, the difference in the estimated precision of the two estimators is quite striking. Comparing the trace of the two estimated covariance matrices, the GME ($\gamma = .5$) estimator is superior to ML-logit-ME by a factor of 8, which exceeds the theoretically predicted convergence factor in Proposition 3.

7. EMPIRICAL EXAMPLE

We illustrate the application of GME(γ) using the data on occupational choice that Long (1983) and Koop and Poirier (1993) examined using multinomial logit. The data used by Long (1983) are included with the GAUSS computer package and thus are available to researchers for replication. In this problem there are $J = 5$ occupational categories, where $j = 1$ is professional, $j = 2$ is menial, $j = 3$ is blue collar, $j = 4$ is craft, and $j = 5$ is white collar. Occupation status is viewed as a function of $K - 1 = 3$ individual characteristics with $\mathbf{x}_n' = [1, x_{n2}, x_{n3}, x_{n4}]$, where x_{n2} is experience, x_{n3} is education, and x_{n4} is a discrete variable that equals 1 if individual n is white. The sample includes

Table 4. Results of a Sampling Experiment with 500 Replications, N = 30 and $\kappa(X'X) = 90$

Estimator	Misses: Min	Misses: Average	Misses: Max	MSE($\hat{\beta}$)	I($\hat{\mathbf{p}}$)
GME ($\gamma = .1$)	7	16.32	22	1.700	.001
GME ($\gamma = .5$)	6	14.18	20	4.468	.009
GME ($\gamma = .9$)	6	16.29	22	18.755	.070
ME-ML logit	6	13.60	20	145.235	.154

Table 5. Estimated Covariance Matrices for ML Logit and GME ($\gamma = .5$), When $N = 30$ and $\kappa(X'X) = 90$

Estimated covariance matrices, category 2, $K = 4$			
	ML Logit		
5.97	−4.00	−5.86	−2.40
	3.06	3.58	1.70
		6.18	2.50
			1.37
	GME ($\gamma = .5$)		
.75	−.50	−.74	−.30
	.38	.50	.21
		.78	.31
			.17

$N = 337$ employed males from the 1982 General Social Survey. Based on these data, we report in Table 6 the ML multinomial logit and GME ($\gamma = .5$) estimates for β_{jk}. In line with Section 3, we used the noise component with $M = 7, \nu_{ij} = (-.1, -.0666, \ldots, .1)$.

Both models were estimated using GAMS. Estimating the ME-ML or the GME using the dual (unconstrained) approach takes under 2 seconds in real time. The GME actually takes fewer iterations than the ME-ML approach (62 compared to 66). In contrast, the primal (constrained) approach to estimating the GME takes 4,112 iterations, and the primal approach to the ME takes 4,200 iterations. In short, using the dual approach places little, if any, computational burden on either the ME or GME estimators.

The ME-ML procedure has 168 category misses (49.9% of the observations are correctly predicted), whereas the GME ($\gamma = .5$) estimator has 169 category misses (49.6% correct). The $\mathbf{I}(\hat{\mathbf{p}})$ is lower for the GME (.195) than for the ME-ML estimates (.213).

Once M gets large (7 or above), the GME ($\gamma = .5$) estimates are not very sensitive to the choice of M. If instead of using $M = 7$ as in Table 6, we used $M = 9$ (63 iterations), then the number of category misses and $\mathbf{I}(\hat{\mathbf{p}})$ would remain the same and the coefficients of both estimates would be close, in most cases differing in the second through fourth places.

At first glance, the coefficients from the ME-ML and GME estimators in Table 6 look substantially different. But this difference is misleading. The coefficients depend on arbitrary normalization conventions. A more informative comparison is to look at the effect of a change in a variable on the probability. In Table 7, we show the marginal effects of a change in x_k on the probabilities evaluating at the mean $\mathbf{x}_n = (1, 20.5, 13.10, .92)'$. the first two rows of the table show the change in the probabilities given a small change in one of the continuous variables, when the other variables are fixed at their mean values. The last row shows the difference in the probabilities when the discrete variable x_4 (race) equals 1 and when it equals zero, holding the other variables at their mean values. By inspection, the marginal effects of the GME are quite close to those of the ME, though the GME effects are slightly closer to zero.

Table 6. ML Multinomial Logit and GME ($\gamma = .5$) Results Based on $N = 337$, $J = 5$, $K = 4$, Under the Normalization $\beta_{1k} = (0, 0, 0, 0)'$

Parameter vector	Constant coefficient	Experience coefficient	Education coefficient	Race coefficient
		ML logit		
β_{2k}	.7411	.0047	−.0994	1.2367
	(2.525)	(.030)	(.181)	(1.133)
β_{3k}	−1.0916	.0277	.0938	.4724
	(2.301)	(.026)	(.163)	(.807)
β_{4k}	−6.2390	.0346	.3532	1.5717
	(4.714)	(.041)	(.312)	(1.870)
β_{5k}	−11.5190	.0357	.7788	1.7746
	(3.850)	(.036)	(.234)	(1.354)
		GME ($\gamma = .5$)		
β_{2k}	1.8858	−.0031	−.1782	1.0348
	(2.446)	(.028)	(.171)	(1.169)
β_{3k}	.1871	.0193	−.0040	.2699
	(2.195)	(.024)	(.154)	(.809)
β_{4k}	−4.4770	.0255	.2359	1.2561
	(4.004)	(.036)	(.258)	(1.706)
β_{5k}	−9.5628	.0267	.6502	1.4445
	(3.133)	(.030)	(.186)	(1.185)

NOTE: Asymptotic standard derivations are shown in parentheses.

At the mean, the probability that occupation j is chosen is $p_1 = .09, p_2 = .18, p_3 = .29, p_4 = .16$, and $p_5 = .27$, based on the ML estimates. The GME ($\gamma = .5$) estimates have the slightly more uniform probabilities: $p_1 = .12$, $p_2 = .18, p_3 = .28, p_4 = .156$, and $p_5 = .27$.

Using the normalized entropy measure, we follow Golan (1988, 1994) and Soofi (1992, 1994) in analyzing the im-

Table 7. Marginal Effects on the Probability of Changes in x_k for ML and GME When $\mathbf{x}_m' = (1, 20.50, 13.10, .92)$

j	1	2	3	4	5
k					
		ML logit-ME			
2	−.0023	−.0036	.0011	.0017	.0031
3	−.0258	−.0687	−.0529	.0128	.1345
4 ($(p(x_4 = 0), (p(x_4 = 1))$	.217, .086	.136, .186	.439, .279	.088, .168	.120, .281
		GME ($\gamma = .5$)			
2	−.0019	−.0034	.0010	.0015	.0029
3	−.0208	−.0644	−.0491	.0090	.1254
4 ($(p(x_4 = 0), (p(x_4 = 1))$	.222, .117	.133, .181	.418, .282	.093, .155	.134, .266

Table 8. Impact of Covariate Choice

Model	Misses	$\Delta S(\hat{\mathbf{p}})$
x_1, x_2, x_3	172	.005
x_1, x_2, x_4	220	.131
x_1, x_3, x_4	174	.006

pact of not including one of the covariates. The results are presented in Table 8. The first row shows the effect of excluding the x_4 (race) covariate. Other rows show the effects of excluding other covariates. Excluding x_3 (education) substantially increases the number of misses and has a great effect in terms of the $S(\hat{\mathbf{p}})$ measure. In general, excluding any covariate increases $S(\hat{\mathbf{p}})$. For example, excluding x_2 (experience) increases $S(\hat{\mathbf{p}})$ by only .006, whereas excluding x_3 (education) increases $S(\hat{\mathbf{p}})$ by .131. Thus these information measures provide one basis for identifying how much each covariate contributes relative to the information measures $S(\hat{\mathbf{p}})$ or $\mathbf{I}(\hat{\mathbf{p}})$. In this data set the covariate x_3 (education) appears to be the most important in its contribution to the reduction of the uncertainty measure. In this example, the covariate x_4 (race) is of the smallest information value.

8. SUMMARY AND COMMENTS

We use a generalized ME formalism, which we call GME(γ), as a basis for recovering the unknown multinomial probabilities p_{ij} and covariate parameters β_j from a sample of multinomial response data. In our formulation, the information contained in the data is captured in the form of inverse information-moment relations with noise. Under the GME(γ) formulation, the criterion function contains a dual objective function involving the unknown multinomial probabilities p_{ij} and the random noise component e_{ij}. Under both the classical ME (Denzau et al. 1989; Soofi 1992) and our GME(γ) formulations, the solution values of the Lagrange multipliers, $\boldsymbol{\lambda}_j$, associated with the information-moment relations for the data may, with a change of sign, be interpreted as the unknown response coefficients β_j. Under traditional approaches to the multinomial response problem, a likelihood function and a cdf linking the p_{ij} and $\mathbf{x}_i'\beta_j$ are specified, and KJ likelihood equations are used to obtain estimates of the β_j and thus the p_{ij} and e_{ij}. Because the NJ unknown p_{ij} exceeds the KJ number of likelihood equations, only the assumption of a specific cdf specification makes the ML formulation a well-posed problem that can be solved for β_j and thus p_{ij}.

As noted in Section 1, both the ME and GME approaches to recovering estimates of unknown parameters from multinomial response data provide advantages over the traditional ML approach. The advantages of both ME and GME are that they provide a simple basis for introducing and evaluating the value of prior information in recovering the multinomial probabilities and the response parameters and they provide a basis for model diagnostics through the use of information measures.

In addition, the GME estimator has several advantages over the classical ME and traditional ML multinomial logit estimators, in that it (a) avoids strict parametric assumptions; (b) makes use of a dual criterion function that permits, by the use of weights, a choice between the estimation precision and prediction objectives; (c) converges more rapidly that ME-ML; (d) yields more precise estimates in finite samples of the response parameters β_j; (e) works well when the sample is small, covariates are highly correlated, or the design matrix is ill-conditioned; and (f) contains the classical ME and traditional ML as special cases in the limit as the sample becomes large.

Solution algorithms for solving the ME (Agmon, Alhassid, and Levine 1979) and GME (Miller 1994) nonlinear inversion problems are easy to implement using nonlinear optimization software package such as GAMS or MATLAB. Soon, GME will be available in SHAZAM. Our simulation and real-world data experiment indicate that the GME estimates require no more computer time than the ME or ML estimates.

The framework developed in Sections 3, 4, and 5 also forms the basis for specifying a GME(γ) formulation for the conditional logit problem (McFadden 1974; Soofi 1992) and a basis for recovering the K-dimensional β vector and the corresponding p_{ij} and e_{ij}. In another work (Golan, Judge, and Perloff 1995), we have also been able to apply the GME(γ) formulation of the multinomial response problem to various forms of model misspecifications and to determine its usefulness in analyzing censored and ordered multinomial response data.

APPENDIX: PROOFS OF PROPOSITIONS

Proof of Proposition 1

Given the GME problem (20)–(23), the corresponding Lagrangian (24), and the notation in (25) and (26), we may, as in Section 2, rewrite the dual *unconstrained* GME problem as

$$\begin{aligned} M(\lambda_j) &= -\mathbf{p}(\lambda)'\log_e\mathbf{p}(\lambda) - \mathbf{w}(\lambda)'\log_e\mathbf{w}(\lambda) \\ &\quad + \lambda'[(\mathbf{I}_J \otimes X')y - (\mathbf{I}_J \otimes X')\mathbf{p} - (\mathbf{I}_J \otimes X')V\mathbf{w}] \\ &= -\mathbf{p}(\lambda)'[-\mathbf{x}_i'\lambda_j - \log_e\Omega_i] - \mathbf{w}(\lambda)'[-\mathbf{x}_i'\lambda_j V_j - \log_e\psi_{ij}] \\ &\quad + \lambda'[(\mathbf{I}_J \otimes X')y - (\mathbf{I}_J \otimes X')\mathbf{p} - (\mathbf{I}_J \otimes X')V\mathbf{w}] \\ &= \mathbf{y}'(\mathbf{I}_J \otimes X)\lambda + \sum_i \log_e[\Omega_i(\lambda_j)] \\ &\quad + \sum_i\sum_j \log_e[\psi_{ij}(\lambda_j)] \end{aligned} \tag{A.1}$$

with gradient

$$\Delta M(\lambda_j) = (\mathbf{I} \otimes X')\mathbf{y} - (\mathbf{I} \otimes X')\mathbf{p} - (\mathbf{I} \otimes X')V\mathbf{w}, \tag{A.2}$$

which is the information-moment conditions (21) and the gradient of the generalized log-likelihood (29). Consistent with (15), the first two components on the right side of (A.1) are identical to the components of the log-likelihood of the logistic distribution. The Hessian is positive definite with unique solution λ_j^0. If we rewrite (A.1) in a form normed by N, then we have

$$\begin{aligned} M_N(\lambda_j) = N^{-1}\mathbf{y}'(\mathbf{I}_J \otimes X')\lambda_j - N^{-1}\sum_i \log_e[\mathbf{1}'\exp(-\mathbf{x}_i'\lambda_j)] \\ - N^{-1}\sum_i\sum_j \log_e[\mathbf{1}'\exp(-\mathbf{x}_i'\lambda_j V_j)]. \end{aligned} \tag{A.3}$$

The last component on the right side of (A.3) may be expressed as

$$N^{-1}\sum_i\sum_j \log_e[\mathbf{1}'\exp(-\mathbf{x}_i'\boldsymbol{\lambda}_j V_j)]$$

$$\equiv N^{-1}\sum_j\sum_k \log_e[\mathbf{1}'\exp(-V_j^*\boldsymbol{\lambda}_j)], \quad \text{(A.4)}$$

where

$$\sum_i X_{ik}e_{ij} = \sum_i\sum_m X_{ik}V_{ijm}w_{ijm} = \sum_m V_{kjm}^* w_{kjm} = \delta_{kj}$$

for all k and j. The $V_{kjm}^* \in [-\max|x_{ik}|, \max|x_{ik}|]$ for all $k = 1, 2, \ldots, K$. As the $\{\mathbf{x}_i\}$ are uniformly bounded, the right side of (A.4) vanishes as $N \to \infty$, because $\sum_j\sum_k \log_e(\cdot)$ is bounded and $N^{-1}\sum_j\sum_k \log_e(\cdot) \to 0$. At the limit

$$M(\lambda_j)\mathbf{p} \to \mathbf{y}'(\mathbf{I}_j \otimes X')\boldsymbol{\lambda}_j^0$$

$$- \sum_i \log_e[\mathbf{1}'\exp(-\mathbf{x}_i'\boldsymbol{\lambda}_j^0)] - O_p(N^{-1}), \quad \text{(A.5)}$$

where $\boldsymbol{\lambda}_j^0$ is the unique solution to $\Delta M^*(\hat{\boldsymbol{\lambda}}_j) = 0$ and $\text{plim}[\tilde{\boldsymbol{\lambda}}_j(N)] = \boldsymbol{\lambda}_j^0$ by theorem 4.1.1 of Amemiya (1985, pp. 106–107). Note that (A.5) is just the ME-ML logit log-likelihood function. This result means the first-order conditions are identical and implies that at the limit $\hat{\mathbf{p}} \to \mathbf{p}$ or $\hat{\boldsymbol{\beta}} \to \boldsymbol{\beta}$, meaning the ME-ML logit and GME solution vectors are equivalent, and the asymptotic $\text{var}(\hat{\boldsymbol{\beta}}_j)_{\text{GME}} = \text{asymptotic var}(\boldsymbol{\beta}_j^*)_{\text{ME-ML logit}}$.

Proof of Proposition 2

a. The upper bound of (32) is reached when $\boldsymbol{\beta}_j = \mathbf{0}$, implying that $p_{ij} = 1/J$ for all i, j and $w_{ijm} = 1/m$ for all i, j, m.
b. Constraint (21) in the GME formulation is a relaxed version of the ME information-moment constraints (8). Consequently, the GME constraints (21) are less binding than the ME constraints (8). Because all constraints are binding ($0 < |\beta_{jk}| \ll \infty$), and given the negative definite Hessian (30) and thus the concavity of (20), this implies that $|\lambda_{jk}|_{\text{GME}} \leq |\lambda_{jk}|_{\text{ME}}$ or $|\beta_{jk}|_{\text{GME}} \leq |\beta_{jk}|_{\text{ME}}$ for all j, k.
c. Given (a) and (b), it follows that the first right side element of (32) is always equal to or larger than the right side of (14).
d. As the second right side element of (32) is always nonnegative, the elements on the diagonal of $\mathbf{I}(\boldsymbol{\beta}_j)_{\text{GME}}$ must be *greater than or equal to* the corresponding element of $\mathbf{I}(\boldsymbol{\beta}_j)_{\text{ME}}$.
e. Because the variance of β_{jk} are the diagonal elements of $[\mathbf{I}(\boldsymbol{\beta}_j)]^{-1}$, it follows that the $\text{var}(\boldsymbol{\beta}_j)_{\text{GME}} \leq \text{var}(\boldsymbol{\beta}_j)_{\text{ME-ML logit}}$ for any $N \ll \infty$.
f. Finally, from Proposition 1, the ME-ML logit and GME solution vectors and information matrices are equivalent and the $\text{var}(\boldsymbol{\beta}_j)_{\text{GME}} = \text{var}(\boldsymbol{\beta}_j)_{\text{ME-ML logit}}$ as $N \to \infty$.

Proof of Proposition 3

a. As $|\beta_{jk}| \to 0$, for all j, k, the $\text{var}(\boldsymbol{\beta}_j)_{\text{ME}}$ and $\text{var}(\boldsymbol{\beta}_j)_{\text{GME}}$ attain their lower bound.
b. Taking these bounds for (14) and (32) means that each element of $\mathbf{I}(\boldsymbol{\beta}_j)_{\text{ME}} \to N/J^2$ and each element of $\mathbf{I}(\boldsymbol{\beta}_j)_{\text{GME}} \to N/J^2 + N/M$.
c. Rewriting (b) in terms of the variances yields Proposition 3.

[Received January 1994. Revised August 1995.]

REFERENCES

Agmon, N., Alhassid, Y., and Levine, R. D. (1979), "An Algorithm for Finding the Distribution of Maximal Entropy," *Journal of Computational Physics*, 30, 250–259.

Albert, J. H., and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679.

Amemiya, T. (1985), *Advanced Econometrics*, Cambridge, MA: Harvard University Press.

Behara, M. (1990), *Additive and Nonadditive Measures of Entropy*, New York: John Wiley.

Belsley, D. A. (1991), *Conditioning Diagnostics*, New York: John Wiley.

Csiszár, I. (1991), "Why Least Squares and Maximum Entropy? An Axiomatic Approach to Inference for Linear Inverse Problems," *The Annals of Statistics*, 19, 2032–2066.

Denzau, A. T., Gibbons, P. C., and Greenberg, E. (1989), "Bayesian Estimation of Proportions With a Cross-Entropy Prior," *Communications in Statistics, Part A—Theory And Methods*, 18, 1843–1861.

Donoho, D. L., Johnstone, I. M., Hoch, J. C., and Stern, A. S. (1992), "Maximum Entropy and the Nearly Black Object," *Journal of the Royal Statistical Society*, Ser. B, 54, 41–81.

Gokhale, D. V., and Kullback, S. (1978), *The Information in Contingency Tables*, New York: Marcel Dekker.

Golan, A. (1988), "A Discrete Stochastic Model of Economic Production and a Model of Fluctuations in Production—Theory and Empirical Evidence," unpublished doctoral thesis, University of California, Berkeley.

——— (1994), "A Multivariable Stochastic Theory of Size Distribution of Firms With Empirical Evidence," *Advances in Econometrics*, 10, 1–46.

Golan, A., Judge, G., and Perloff, J. (1995), "Estimation and Inference With Censored and Ordered Multinomial Response Data," unpublished paper, University of California, Berkeley.

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

Greene, W. H. (1993), *Econometric Analysis* (2nd ed.), New York: Macmillan.

Griffiths, W. E., Hill, R. C., and Pope, P. J. (1987), "Small-Sample Properties of Probit Model Estimators," *Journal of the American Statistical Association*, 82, 929–937.

Hastie, T., Buja, A., and Tibshirani, R. (1993), "Penalized Discriminant Analysis," unpublished paper, Bell Laboratories.

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

——— (1957b), "Information Theory and Statistical Mechanics II," *Physics Review*, 108, 171–190.

——— (1984), "Prior Information and Ambiguity in Inverse Problems," *Inverse Problems*, ed. D. W. McLaughlin, Providence RI: American Mathematical Society (pp. 151–166).

Judge, G. G. (1991), "A Reformulation of Ill-Posed Inverse Problems With Noise," unpublished manuscript, University of California, Berkeley.

Judge, G. G., Golan, A., and Miller, D. (1993), "Recovering Information in the Case of Ill-Posed Inverse Problems With Noise," unpublished manuscript, University of California, Berkeley.

Judge, G. G., Griffiths, W. E., Hill, R. C., Lütkepohl, H., and Lee, T.-C. (1985), *The Theory and Practice of Econometrics*, (2nd ed.), New York: John Wiley.

Koop, G., and Poirier, D. J. (1993), "Bayesian Analysis of Logit Models Using Natural Conjugate Priors," *Journal of Econometrics*, 56, 323–340.

Kullback, J. (1959), *Information Theory and Statistics*, New York: John Wiley.

Le Cessie, S., and van Houwelingen, J. C. (1992), "Ridge Estimators in Logistic Regression," *Journal of Applied Statistics*, 41, 191–201.

Lehmann, E. (1983), *Theory of Point Estimation*, New York: John Wiley.

Levine, R. D. (1980), "An Information Theoretical Approach to Inversion Problems," *Journal of Physics*, Ser. A, 13, 91–108.

Long, J. S. (1983), "A Graphical Method for the Interpretation of Multinomial Logit Analysis," *Sociological Methods and Research*, 15, 420–446.

Maddala, G. S. (1983), *Limited Dependent and Qualitative Variables in Econometrics*, Cambridge, U.K.: Cambridge University Press.

Mallows, C. L. (1973), "Some Comments on C_p," *Technometrics*, 15, 661–665.

Manski, C., and D. McFadden (Eds.) (1981), *Structural Analysis of Discrete Data with Econometric Applications*, Cambridge, MA: MIT Press.

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

Miller, D. J. (1994), "Entropy and Information Recovery in Linear Economic Models," unpublished doctoral thesis, University of California, Berkeley.

O'Sullivan, F. (1986), "A Statistical Perspective on Ill-Posed Inverse Problems," *Statistical Science*, 1, 502–527.

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

Soofi, E. S. (1992), "A Generalizable Formulation of Conditional Logit With Diagnostics," *Journal of the American Statistical Association*, 87, 812–816.

——— (1994), "Capturing the Intangible Concept of Information," *Journal of the American Statistical Association*, 89, 1243–1254.

Tikhonov, A. N., and Arsenin, V. Y. (1977), *Solutions of Ill-Posed Problems*, Washington, D.C.: Winston.

Zellner, A. (1993), "Bayesian and Non-Bayesian Estimation Using Balanced Loss Functions," in *Statistical Decision Theory: Theory and Related Topics*, eds. S. S. Gupta and J. O. Berger, New York: Springer-Verlag, pp. 377–390.

Zellner, A., and Rossi, P. E. (1984), "Bayesian Analysis of Dichotomous Quantal Response Models," *Journal of Econometrics*, 25, 365–393.