Article

The Data-Constrained Generalized Maximum Entropy Estimator of the GLM: Asymptotic Theory and Inference

Ron Mittelhammer [1],*, Nicholos Scott Cardell [2] and Thomas L. Marsh [3]

[1] Economic Sciences and Statistics, Washington State University, Pullman, WA 99164, USA

[2] Salford Systems, San Diego, CA 92126, USA; E-Mail: scardell@gocougs.wsu.edu

[3] Economic Sciences and IMPACT Center, Washington State University, Pullman, WA 99164, USA; E-Mail: tl_marsh@wsu.edu

* Author to whom correspondence should be addressed; E-Mail: mittelha@wsu.edu; Tel.: +1-509-335-1706; Fax: +1-509-335-1173.

Abstract: Maximum entropy methods of parameter estimation are appealing because they impose no additional structure on the data, other than that explicitly assumed by the analyst. In this paper we prove that the data constrained GME estimator of the general linear model is consistent and asymptotically normal. The approach we take in establishing the asymptotic properties concomitantly identifies a new computationally efficient method for calculating GME estimates. Formulae are developed to compute asymptotic variances and to perform Wald, likelihood ratio, and Lagrangian multiplier statistical tests on model parameters. Monte Carlo simulations are provided to assess the performance of the GME estimator in both large and small sample situations. Furthermore, we extend our results to maximum cross-entropy estimators and indicate a variant of the GME estimator that is unbiased. Finally, we discuss the relationship of GME estimators to Bayesian estimators, pointing out the conditions under which an unbiased GME estimator would be efficient.

Keywords: generalized maximum entropy; generalized maximum cross-entropy; asymptotic Theory; GME computation; unbiased GME; GME as Bayesian estimation

MSC 2010 Codes: 62

1. Introduction

Information theoretic estimators have been receiving increasing attention in the econometric-statistics literature [1–7]. In other work, [3] proposed an information theoretic estimator based on minimization of the Kullback-Leibler Information Criterion as an alternative to optimally-weighted generalized method of moments estimation. This specific estimator handles weakly dependent data generating mechanisms and under reasonable regulatory assumptions it is consistent and asymptotically normally distributed. Subsequently, [1] proposed an information theoretic estimator based on minimization of the Cressie-Read discrepancy statistic as an alternative approach to inference in moment condition models. In [1] identified a special case of the Cressie-Read statistic—the Kullback-Leibler Information Criterion (e.g., maximum entropy)—as being preferred over other estimators (e.g., empirical likelihood) because of its efficiency and robustness properties. Special issues of the *Journal of Econometrics* (March 2002) and *Econometric Reviews* (May 2008) were devoted to this particular topic of information estimators.

Historically, information theoretic estimators have been motivated in several ways. The Cressie-Read statistic directly minimizes an information based concept of closeness between the estimated and empirical distribution [1]. Alternatively, the maximum entropy principle is based on an axiomatic approach that defines a unique objective function to measure uncertainty of a collection of events [8–10]. Interest in maximum entropy estimators stems from the prospect to recover and process information when the underlying sampling model is incompletely or incorrectly known and the data are limited, partial, or incomplete [10]. To date the principle of maximum entropy has been applied in an abundance of circumstances, including in the fields of econometrics and statistics [11–17], economic theory and applications [18–24], accounting and finance [25–27], and resources and agricultural economics [28–32]. Moreover, widely used econometric software packages are now incorporating procedures to calculate maximum entropy estimators in their latest releases (e.g., SAS, SHAZAM, and GAUSSX).

In most cases, rigorous investigation of small and large sample properties of information theoretic estimators have lagged far behind empirical applications [3]. Exceptions include [1–3] who examined information theoretic alternatives to generalized method of moments estimation; [14] who derived the statistical properties of the generalized maximum entropy estimator in the context of modeling multinomial response data; and, [10] who provided asymptotic properties for the moment-constrained generalized maximum entropy (GME) estimator for the general linear model (showing it is asymptotically equivalent to ordinary least squares). An alternative information theoretic estimator of the general linear model (GLM), yet to be rigorously investigated, but that has arisen in empirical applications (e.g., [24]), is the purely data-constrained formulation of the generalized maximum entropy estimator [10]. In a purely data-constrained formulation the regression model itself, as opposed to moment conditions of it, represents the constraining function to the entropy objective function. In the maximum entropy framework, unlike ordinary least square or maximum likelihood estimators of the GLM, moment constraints are not necessary to uniquely identify parameter estimates. Moreover, there exists distinct differences between the data and moment constrained versions of the GME for the GLM. For [10] have shown the data-constrained GME estimator to be mean square error superior to the moment-constrained GME estimator of the GLM in selected Monte Carlo experiments.

Our paper contributes to the econometric literature in several ways. First, regularity conditions are identified that provide a solid foundation from which to develop statistical properties of the data constrained GME estimator of the GLM and hypothesis tests on model parameters. Given the regularity

conditions, we define a conditional maximum entropy function to rigorously prove consistency and asymptotic normality. As demonstrated in this paper the data-constrained GME estimator is not asymptotically equivalent to the moment-constrained GME estimator or ordinary least squares estimator. However, the GME estimator is shown to be nearly asymptotically efficient. Moreover, we derive formulae to compute the asymptotic variance of the proposed estimator. This allows us to define classical Wald, Likelihood Ratio, and Lagrange Multiplier tests for testing hypothesis about model parameters.

Second, theoretical extensions to unbiased, cross entropy, and Bayesian estimation are also identified. Further, we demonstrate that the GME specification can be extended from finite-discrete parameter and error spaces to infinite-continuous parameter and error spaces. Alternative formulations of the data constrained GME estimator of the GLM under selected regularity conditions, and the implications to properties of the estimator, are also discussed.

Third, to compliment the theoretical results, Monte Carlo experiments are used in comparing the performance of the data-constrained GME estimates to least squares estimates for small and medium size samples. The performance of the GME estimator is tested relative to selected distributions of the errors, to the user supplied supports of the parameters and errors, and to its robustness to model misspecification. Monte Carlo experiments are also performed to examine the size and power of the Wald, Likelihood Ratio, and Lagrange Multiplier test statistics.

Fourth, insight into computational efficiency and guidelines for setting boundaries of parameters and error support spaces are discussed. The conditional maximum entropy formulation utilized in proof of asymptotic properties provides a basis for new computationally efficient method of calculating GME estimates. The approach involves a nonlinear search over a *K*-vector of coefficient parameters, which is much more efficient than numerical approaches proposed elsewhere in the literature. Finally, practical guidelines for setting boundaries of parameters and error support spaces are analyzed and discussed.

2. The Data-Constrained GME Formulation

Let $Y = X\beta + \varepsilon$ represent the general linear model with Y being an $N \times 1$ dependent variable vector, X being a fixed $N \times K$ matrix of explanatory variables, β being a $K \times 1$ vector of parameters, and ε being an $N \times 1$ vector of disturbance terms (All of our results can be extended to stochastic X. For example, if $X_{i\bullet}$ is iid with $Var(X_{i\bullet}) = \Omega$, a positive definite matrix, then the asymptotic properties are identical to those developed below). The GME rule for defining the estimator of the unknown β in the general linear model formulation is given by $\hat{\beta} = Z\hat{p}$ with $\hat{p} = (\hat{p}'_1, \ldots, \hat{p}'_K)'$ derived from the following constrained maximum entropy problem: $p = (p'_1, \ldots, p'_K)'$

$$\max_{p_k, w_i : \forall k, i} \left(-\sum_{k=1}^{K} p'_k \ln(p_k) - \sum_{i=1}^{N} w'_i \ln(w_i) \right)$$

subject to:

$$Y = XZp + Vw$$

$$\mathbf{1}' p_k = 1 \ \forall k$$

$$\mathbf{1}' w_i = 1 \ \forall i$$

$$p_k > [0], w_i > [0], \forall i, k.$$

In the preceding formulation, the matrices Z and V are $K \times KM$ and $N \times NJ$ matrices of support points for the β and ε vectors, respectively, as:

$$Z = \begin{pmatrix} z_1' & 0 & \cdots & 0 \\ 0 & z_2' & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & z_K' \end{pmatrix} \text{ and } V = \begin{pmatrix} v_1' & 0 & \cdots & 0 \\ 0 & v_2' & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & v_N' \end{pmatrix},$$

where $z_k = (z_{k1}, \ldots, z_{kM})'$ is a $M \times 1$ vector such that $z_{k1} \le z_{k2} \le \cdots \le z_{kM}$ and $\beta_k \in (z_{k1}, z_{kM}) \forall k = 1, \ldots, K$, and similarly $v_i = (v_{i1}, \cdots, v_{iJ})'$ is a $J \times 1$ vector such that $v_{i1} \le v_{i2} \cdots \le v_{iJ}$ and $\varepsilon_i \in (v_{i1}, v_{iJ}) \forall i = 1, \ldots, N$ (in their original formulation, [10] required ε_i to be contained in a fixed interval with arbitrarily high probability. Here we assume such an event occurs with probability). The $M \times 1$ p_k vectors and the $J \times 1$ w_i vectors are weight vectors having nonnegative elements that sum to unity and are used to represent the β and ε vectors as $\beta = Zp$, for $p = (p_1', \ldots, p_K')'$, and $\varepsilon = Vw$, for $w = (w_1', \ldots, w_J')'$.

The basic principle underlying the estimator $\hat{\beta} = Z\hat{p}$ for β is to choose an estimate that contains only the information available. In this way the maximum entropy estimator is not constrained by any extraneous assumptions. The information used is the observed information contained in the data, the information contained in the constraints on the admissible values of β, and the information inherent in the structure of the model, including the choice of the supports for the β_k's. In effect, the information set used in estimation is shrunk to the boundary of the observed data and the parameter constraint information. Because the objective function value increases as the weights in p_i and w_i are more uniformly distributed, any deviation from uniformity represents the effect of the data constraints on the weighting of the support points used for representing β and ε. This fact also motivates the interpretation of the GME as a shrinkage-type estimator that in the absence of constraints on β will shrink $\hat{\beta}$ to the centers of the supports defined in the specification of Z. We next establish consistency and asymptotic normality results for the GME estimator under general regularity conditions on the specification of the estimation problem.

3. Consistency and Asymptotic Normality of the GME Estimator

Regularity Conditions. To establish asymptotic results for the GME estimator, we utilize the following regularity conditions for the problem of estimating β in $Y = X\beta + \varepsilon$.

R1. The ε_i's are iid with $c_1 + \delta \le \varepsilon_i \le c_J - \delta$ for some $\delta > 0$ and large enough finite positive $c_J = -c_1$.

R2. The pdf of ε_i, $f(\varepsilon_i)$, is symmetric around 0 with variance σ^2.

R3. $\beta_k \in (\beta_{kL}, \beta_{kH})$, for finite β_{kL} and $\beta_{kH}, \forall k = 1, \ldots, K$.

R4. X has full column rank.

R5. $\frac{1}{N}(X'X)$ is O(1) and the smallest eigenvalue of $\frac{1}{N}(X'X) > \varepsilon$ for some $\varepsilon > 0$, and $\forall N > N^*$, where N^* is some positive integer.

R6. $\frac{1}{N}(X'X) \to \Omega$, a finite positive definite symmetric matrix.

Note that condition R1 states that the support of ε_i is contained in the interior of some large enough closed finite interval $[c_1, c_J]$. Condition R3 states that the true value of parameter β_k can be enclosed

within some open interval (β_{kL}, β_{kH}). The conditions R4-R6 on X are familiar analogues to typical assumptions made in the least squares context for establishing asymptotic properties of the least squares estimator of β. We utilize condition R6 to simplify the demonstration of asymptotic normality, but the result can be established under weaker conditions, as alluded to in the proof. Finally, our proof of the asymptotic results will utilize symmetry of the disturbance distribution, which is the content of condition R2.

Reformulated GME Rule. The asymptotic results are derived within the context of the following representation of the GME model, represented in scalar notation to facilitate exposition of the proof. The GME representation described below is completely consistent with the formulation in Section 2 under the condition that the support points represented by the vector v_i are chosen to be symmetrically dispersed around 0. We use the same vector of support points for each of the ε_i's, consistent with the iid nature of the disturbances, and so henceforth v_ℓ refers to the common ℓ^{th} scalar support point in the development below. The representation is also more general than the representation in Section II in the sense that different numbers of support points can be used for the representation of different β_k parameters. The constrained maximum entropy problem is as follows:

$$\underset{b,\,p,\,w}{\text{Max}} \left(-\sum_{k=1}^{K} \sum_{\ell=1}^{J_k} p_{k\ell} \ln(p_{k\ell}) - \sum_{i=1}^{N} \sum_{\ell=1}^{J} w_{i\ell} \ln(w_{i\ell}) \right) \tag{1}$$

subject to:

C1. $\quad \sum_{\ell=1}^{J_k} z_{k\ell} p_{k\ell} = b_k, \beta_{kL} = z_{k1} \leq z_{k2} \leq \cdots \leq z_{kJ_k} = \beta_{kH}, \; k = 1, \ldots, K$

C2. $\quad \sum_{\ell=1}^{J} v_\ell w_{i\ell} = e_i = y_i - X_{i\bullet}b = e_i(b), \; i = 1, \ldots, N$

C3. $\quad c_1 = v_1 \leq v_2 \leq \cdots \leq v_J = c_J$

C4. $\quad -v_\ell = v_{J+1-\ell} (\text{thus for } J \text{ odd } v_{\frac{J+1}{2}} \equiv 0)$

C5. $\quad \sum_{\ell=1}^{J_k} p_{k\ell} = 1, k = 1, \ldots, K$

C6. $\quad \sum_{\ell=1}^{J} w_{i\ell} = 1, i = 1, \ldots, N$

As will become apparent, the nonnegativity restrictions on $p_{k\ell}$ and $w_{i\ell}$ are inherently enforced by the structure of the optimization problem itself, and thus need not be explicitly incorporated into the constraint set.

Asymptotic Properties. The following theorem establishes the consistency and asymptotic normality of the GME estimator of β in the GLM.

Theorem. *Under regularity conditions R1-R5, the GME estimator* $\hat{\beta} = Z\hat{p}$ *is a consistent estimator of β. With the addition of regularity condition R6, the GME estimator is asymptotically normally distributed as*

$$\hat{\beta} \overset{a}{\sim} N\left(\beta, \frac{\sigma_\gamma^2}{N\xi^2} \Omega^{-1} \right)$$

for appropriate definitions of $\sigma_\gamma^2, \xi,$ and Ω.

Proof. Define the maximized entropy function, conditional on $b = \tau$, as:

$$F(\tau) = \underset{\substack{p, w : b = \tau \\ (C1)-(C6)}}{\text{Max}} \left(-\sum_{k=1}^{K} \sum_{\ell=1}^{J_k} p_{k\ell} \ln(p_{k\ell}) - \sum_{i=1}^{N} \sum_{\ell=1}^{J} w_{i\ell} \ln(w_{i\ell}) \right) \tag{2}$$

The optimal value of $w_i = (w_{i1}', \ldots, w_{iJ}')'$ in the conditionally-maximized entropy function is given by:

$$w_i(\tau) = \underset{w_i : C6, \sum_{\ell=1}^{J} v_\ell w_{i\ell} = e_i(\tau)}{\arg\max} \left(-\sum_{\ell=1}^{J} w_{i\ell} \ln(w_{i\ell}) \right),$$

which is the maximizing solution to the Lagrangian:

$$L_{w_i} = -\sum_{\ell=1}^{J} w_{i\ell} \ln(w_{i\ell}) + \lambda_i^w \left(\sum_{\ell=1}^{J} w_{i\ell} - 1 \right) + \gamma_i \left(\sum_{\ell=1}^{J} v_\ell w_{i\ell} - e_i(\tau) \right).$$

The optimal value of $w_{i\ell}$ is then:

$$w_{i\ell}\left(\gamma(e_i(\tau))\right) = w_\ell\left(\gamma(e_i(\tau))\right) = \frac{e^{\gamma(e_i(\tau))v_\ell}}{\sum_{m=1}^{J} e^{\gamma(e_i(\tau))v_m}}, \quad \ell = 1, \ldots, J, \tag{3}$$

where $\gamma(e_i(\tau))$ is the optimal value of the Lagrangian multiplier γ_i under the condition $b = \tau$, and $w_\ell(\gamma) \equiv \dfrac{e^{\gamma v_\ell}}{\sum\limits_{m=1}^{J} e^{\gamma v_m}}$. It follows from the symmetry of the v_i's around zero that:

$$\sum_{\ell=1}^{J} v_\ell w_\ell(-\gamma(e_i(\tau))) = -\sum_{\ell=1}^{J} v_\ell w_\ell(\gamma(e_i(\tau))) \tag{4}$$

Similarly, the optimal value of $p_k = (p_{k1}, \ldots, p_{kJ_k})'$ in the conditionally-maximized entropy function is given by:

$$p_k(\tau_k) = \underset{p_k : C5, \sum_{\ell=1}^{J_k} z_{k\ell} p_{k\ell} = \tau_k}{\arg\max} \left(-\sum_{\ell=1}^{J_k} p_{k\ell} \ln(p_{k\ell}) \right),$$

which is the maximizing solution to the Lagrangian:

$$L_{p_k} = -\sum_{\ell=1}^{J_k} p_{k\ell} \ln(p_{k\ell}) + \lambda_k^p \left(\sum_{\ell=1}^{J_k} p_{k\ell} - 1 \right) + \eta_k \left(\sum_{\ell=1}^{J_k} z_{k\ell} p_{k\ell} - \tau_k \right).$$

The optimal value of $p_{k\ell}$ is then:

$$p_{k\ell}(\tau_k) = \frac{e^{\eta_k(\tau_k) z_{k\ell}}}{\sum_{m=1}^{J_k} e^{\eta_k(\tau_k) z_{km}}}, k = 1, \ldots, K, \tag{5}$$

where $\eta_k(\tau_k)$ is the optimal value of the Lagrangian multiplier η_k under the condition $b_k = \tau_k$.

Substituting the optimal solutions for the $p_{k\ell}$'s and $w_{i\ell}$'s into (2) obtains the conditional maximum value function:

$$F(\tau) = -\sum_{k=1}^{K}\left(\eta_k(\tau_k)\tau_k - \ln\left(\sum_{m=1}^{J_k} e^{\eta_k(\tau_k)z_{km}}\right)\right)$$

$$-\sum_{i=1}^{N}\left(\gamma(e_i(\tau))e_i(\tau) - \ln\left(\sum_{m=1}^{J} e^{\gamma(e_i(\tau))v_m}\right)\right).$$

Define the gradient vector of $F(\tau)$ as $G(\tau) = \dfrac{\partial F(\tau)}{\partial \tau}$ so that:

$$G_k(\tau) = \frac{\partial F(\tau)}{\partial \tau_k} = -\eta_k(\tau_k) + \sum_{i=1}^{N} \gamma(e_i(\tau))X_{ik}, k = 1,\ldots,K,$$

and thus $G(\tau) = -\eta(\tau) + X'\gamma(e(\tau))$, where $\eta(\tau)$ and $\gamma(e(\tau))$ are $K \times 1$ and $N \times 1$ vectors of Lagrangian multipliers. It follows that the Hessian matrix of $F(\tau)$ is given by:

$$H(\tau) = \frac{\partial^2 F(\tau)}{\partial \tau \partial \tau'} = \frac{\partial G(\tau)}{\partial \tau'}$$

$$= -\begin{pmatrix} \dfrac{\partial \eta_1(\tau_1)}{\partial \tau_1} & 0 & \cdots & 0 \\ 0 & \dfrac{\partial \eta_2(\tau_2)}{\partial \tau_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \dfrac{\partial \eta_K(\tau_k)}{\partial \tau_K} \end{pmatrix} + X'\frac{\partial \gamma(e(\tau))}{\partial \tau'}.$$

Regarding the functional form of the derivatives of the Lagrangian multipliers appearing in the definition of $H(\tau)$, it follows from (C2) that:

$$\frac{\partial \sum_{\ell=1}^{J} v_\ell w_\ell(\gamma(e_i(\tau)))}{\partial \gamma(e_i(\tau))} \frac{\partial \gamma(e_i(\tau))}{\partial e_i(\tau)} = 1,$$

so that from (3):

$$\frac{\partial \gamma(e_i(\tau))}{\partial e_i(\tau)} = \left(\sum_{\ell=1}^{J} v_\ell^2 w_\ell(\gamma(e_i(\tau))) - e_i^2(\tau)\right)^{-1}.$$

Then, from (C2) $\dfrac{\partial e_i(\tau)}{\partial \tau_k} = -X_{ik}$, and thus:

$$H_{k\ell}(\tau) = -\sum_{i=1}^{N} \frac{X_{ik}X_{i\ell}}{\left(\sum_{\ell=1}^{J} v_\ell^2 w_\ell(\gamma(e_i(\tau))) - e_i^2(\tau)\right)} \text{ for } k \neq \ell.$$

Also, based on (C1):

$$\frac{\partial \eta_k(\tau_k)}{\partial \tau_k} = \left(\sum_{\ell=1}^{J_k} z_{k\ell}^2 p_{k\ell} - \tau_k^2\right)^{-1},$$

so that:

$$H_{kk}(\tau) = -\left(\sum_{i=1}^{N} \frac{X_{ik}^2}{\left(\sum_{\ell=1}^{J} v_\ell^2 w_\ell(\gamma(e_i(\tau))) - e_i^2(\tau)\right)}\right) - \frac{1}{\left(\sum_{\ell=1}^{J_k} z_{k\ell}^2 p_{k\ell} - \tau_k^2\right)}.$$

Because the denominators of the terms in the definition of the $H_{k\ell}$'s are positive valued, it follows that $H(\tau)$ is a negative definite matrix, because $X'X$ is positive definite.

Now consider the case where $\tau = \beta$, so that:

$$e_i(\beta) = y_i - X_{i\bullet}\beta = \varepsilon_i = \sum_{\ell=1}^{J} v_\ell w_\ell(\gamma(e_i(\beta))), i = 1, \ldots, N$$

are iid with mean zero, and thus:

$$\varepsilon_i = \sum_{\ell=1}^{J} v_\ell \frac{e^{\gamma(e_i(\beta))v_\ell}}{\sum_{m=1}^{J} e^{\gamma(e_i(\beta))v_m}}$$

are iid with mean zero. Because ε_i is bounded in the interior of $[v_1, v_J]$, the range of $\gamma(e_i(\beta)) \equiv \gamma(\varepsilon_i)$ is bounded as well. In addition, $\gamma(e_i(\beta))$ is symmetrically distributed around zero because the ε_i's are so distributed, and, from (4):

$$\varepsilon_i = \zeta = \sum_{\ell=1}^{J} v_\ell \frac{e^{\gamma(e_i(\beta))v_\ell}}{\sum_{m=1}^{J} e^{\gamma(e_i(\beta))v_m}} \Rightarrow \sum_{\ell=1}^{J} v_\ell \frac{e^{-\gamma(e_i(\beta))v_\ell}}{\sum_{m=1}^{J} e^{-\gamma(e_i(\beta))v_m}} = -\zeta = -\varepsilon_i \tag{6}$$

It follows that $E(\gamma(e_i(\beta))) = 0$, the $\gamma(\varepsilon_i) \equiv \gamma(e_i(\beta))$'s are iid, and $\gamma(\varepsilon_i)$ has finite variance, say $Var(\gamma(\varepsilon_i)) = \sigma_\gamma^2$. Then, using a multivariate version of Liapounov's central limit theorem, and given condition R6 (asymptotic normality can be established without regularity condition R6. In fact, the boundedness properties on the *X*-matrix stated in R5 would be sufficient. See [33] for a related proof under the weaker regularity conditions).

$$\frac{1}{\sqrt{N}} G(\beta) = \frac{1}{\sqrt{N}} (-\eta(\beta) + X'\gamma(e_i(\beta))) \xrightarrow{d} N([0], \sigma_\gamma^2 \Omega)$$

3.1. Consistency

For any τ, represent the conditional maximum value function, $F(\tau)$, by a second order Taylor series around β as:

$$F(\tau) = F(\beta) + G(\beta)'(\tau - \beta) + \frac{1}{2}(\tau - \beta)' H(\beta^*)(\tau - \beta) \tag{7}$$

where β^* lies between τ and β. The value of the quadratic term in the expansion can be bounded by:

$$\frac{1}{2}(\tau - \beta)' H(\beta^*)(\tau - \beta) \le -\frac{1}{2}\lambda_s(-\frac{1}{N}H(\beta^*)) \cdot N \cdot \|\tau - \beta\|^2 \tag{8}$$

where $\lambda_s(-\frac{1}{N}H(\beta^*))$ denotes the smallest eigenvalue of $-\frac{1}{N}H(\beta^*)$ and $\|a\| \equiv \left(\sum_{k=1}^{K} a_k^2\right)^{\frac{1}{2}}$ [34]. The

smallest eigenvalue exhibits a positive lower bound given by $\left(\dfrac{1}{C_J^2}\right)\lambda_s(\dfrac{1}{N}X'X)$ whatever the value of β^*.

The value of the linear term in the expansion is bounded in probability; that is, $\forall\,\alpha>0$ and for $N>N(\alpha)$, there exists a finite $A(\alpha)$ such that:

$$P\left(\left|G(\beta)'(\tau-\beta)\right|<\sqrt{N}A(\alpha)\left|\tau-\beta\right|,\forall\,\tau\right)>1-\alpha \tag{9}$$

because $\dfrac{1}{\sqrt{N}}G(\beta)\xrightarrow{d}N([0],\sigma_\gamma^2\Omega)$. It follows from Equations (7)–(9) that, for all $\delta>0$, $P(\underset{\tau:|\beta-\tau|>\delta}{\text{Max}}(F(\tau))<F(\beta))\to 1$ as $N\to\infty$. Thus $\hat{\beta}=\arg\underset{\tau}{\max}(F(\tau))\xrightarrow{p}\beta$, and the GME estimator of β is consistent.

3.2. Asymptotic Normality

Expand $G(b)$ in a Taylor series around β, where $\hat{\beta}=\arg\underset{\tau}{\max}F(\tau)$ is the GME estimator of β, to obtain:

$$G(\hat{\beta})=G(\beta)+H(\beta^*)(\hat{\beta}-\beta) \tag{10}$$

where β^* is between $\hat{\beta}$ and β. In general, different β^* points will be required to represent the different coordinate functions in $G(\hat{\beta})$. At the optimum, $G(\hat{\beta})=[0]$ and $\hat{\beta}$ is a consistent estimator of β; therefore $\beta^*\xrightarrow{p}\beta$, and:

$$\sqrt{N}(\hat{\beta}-\beta)\overset{d}{=}-\left(\dfrac{1}{N}H(\beta)\right)^{-1}\dfrac{1}{\sqrt{N}}G(\beta),$$

where $\overset{d}{=}$ denotes equivalence of limiting distributions. Using $e_i(\beta)\equiv\varepsilon_i$, note that:

$$\dfrac{1}{N}H(\beta)=\dfrac{1}{N}\sum_{i=1}^{N}\dfrac{-X_{i\bullet}'X_{i\bullet}}{\displaystyle\sum_{\ell=1}^{J}v_\ell^2 w_\ell(\gamma(\varepsilon_i))-\varepsilon_i^2}+0_P\left(\dfrac{1}{N}\right),$$

where $\displaystyle\sum_{\ell=1}^{J}v_\ell^2 w_\ell(\gamma(\varepsilon_i))-\varepsilon_i^2, i=1,\ldots,N$ are iid. It follows from R6 that $\dfrac{1}{N}H(\beta)\xrightarrow{p}-\xi\Omega$ with $\xi=E\left(\left(\displaystyle\sum_{\ell=1}^{J}v_\ell^2 w_\ell(\gamma(\varepsilon_i))-\varepsilon_i^2\right)^{-1}\right)$. Recalling that $\dfrac{1}{\sqrt{N}}G(\beta)\xrightarrow{d}N([0],\sigma_\gamma^2\Omega)$, Slutsky's Theorem [34] implies that:

$$\sqrt{N}(\hat{\beta}-\beta)\xrightarrow{d}N([0],\dfrac{\sigma_\gamma^2}{\xi^2}\Omega^{-1})$$

Note that holding the support of ε constant, one can reduce the interval (c_1, c_J). As $\delta\to 0$, the asymptotic variance of $\sqrt{N}(\hat{\beta}-\beta)$ may tend to zero, but cannot grow without bound. For example, if at $\delta=0, \exists\,\varepsilon>0$ such that $P(\varepsilon\leq k)\geq\varepsilon(k-c_1)$, all $k\in(c_1,c_J)$ $(\Rightarrow P(\varepsilon\geq k)\geq\varepsilon(c_J-k)$ all $k\in(c_1,c_J))$, then $\lim_{\delta\to 0}\dfrac{\sigma_\gamma^2}{\xi^2}=0$.

Also note that, for large samples, the parameters reliance on the supports vanishes. In contrast, the supports on the errors influence the computed covariance matrix. Finally, for non-homogenous errors, the covariance matrix estimator could be adjusted following a standard White's covariance correction.

3.3. Cross-Entropy Extensions

To extend the previous asymptotic results to the case of cross-entropy maximization [10], first suppose that $z_{k\ell} = z_{k\ell+1}$ and/or $v_{\ell} = v_{\ell+1}$ for some ℓ. Let $z_{k\ell}^*, \ell = 1,\ldots,J_k^*$ and $v_{\ell}^*, \ell = 1,\ldots,J^*$ denote the distinct values among the $z_{k\ell}$'s and v_{ℓ}^*'s, respectively, and let $a_{k\ell}$ and α_{ℓ} denote the respective multiplicities of the values $z_{k\ell}^*$ and v_{ℓ}^*. From Equations (3) and (5), $w_{i\ell}(\gamma(e_i(\tau))) \equiv w_{im}(\gamma(e_i(\tau)))$ if $v_{\ell} = v_m$ and $p_{k\ell}(\tau_k) \equiv p_{km}(\tau_k)$ if $z_{k\ell} = z_{km}$. Thus, the maximization problem given by Equation (2) and Conditions C1-C6 is equivalent to:

$$\max_{b,p,w}\left(-\sum_{k=1}^{K}\sum_{\ell=1}^{J_k^*} p_{k\ell}^* \ln\left(\frac{p_{k\ell}^*}{a_{k\ell}}\right) - \sum_{i=1}^{N}\sum_{\ell=1}^{J^*} w_{i\ell}^* \ln\left(\frac{w_{i\ell}^*}{\alpha_{\ell}}\right)\right) \qquad (11)$$

with obvious changes being made to C1-C6. The only alterations needed to the preceding proof are:

$$w_{i\ell}^*(\gamma(e_i(\tau))) = w_{\ell}(\gamma(e_i(\tau))) = \frac{\alpha_{\ell} e^{\gamma(e_i(\tau))v_{\ell}}}{\sum_{m=1}^{J^*} \alpha_m e^{\gamma(e_i(\tau))v_m}}, \ell = 1,\ldots,J^*, \text{ and} \qquad (12)$$

$$p_{k\ell}^*(\tau_k) = \frac{a_{k\ell} e^{\eta_k(\tau_k)z_{k\ell}}}{\sum_{m=1}^{J_k^*} a_{km} e^{\eta_k(\tau_k)z_{km}}}, k = 1,\ldots,K. \text{ (13)} \qquad (13)$$

More generally, the same representation (11)-(13) applies for any $a_{k\ell} > 0, \alpha_{\ell} > 0$. Furthermore, Equations (12) and (13) are homogeneous of degree zero in $(\alpha_1,\cdots,\alpha_{J^*})$ and $(a_{k1,}\cdots,a_{kJ_k^*})$, respectively. Thus, without loss of generality, the normalization conditions:

$$\sum_{\ell=1}^{J^*} \alpha_{\ell} \equiv 1 \text{ and } \sum_{\ell=1}^{J_k^*} a_{k\ell} \equiv 1$$

can be imposed.

Using Equations (11), (12), and (13), we have characterized the maximum cross entropy solution. Upon substitution of Equations (11)–(13) in the appropriate arguments, all results, including the results in the next section on statistical testing, apply to the maximum cross-entropy paradigm.

4. Statistical Tests

The GME estimator $\hat{\beta} = Z\hat{p}$ is consistent and asymptotically normally distributed. Therefore, asymptotically valid normal and χ^2 test statistics can be used to test hypotheses about β. For empirical implementation of such tests a consistent estimate of the asymptotic covariance matrix of $\hat{\beta}$ will be required. An estimate of $\frac{1}{N\xi^2}\Omega^{-1}$ is straightforwardly obtained by calculating $M(\hat{\beta})^{-1}(X'X)M(\hat{\beta})^{-1}$, where:

$$M(\hat{\beta}) = \sum_{i=1}^{N} \left(\frac{-X_{i\bullet}'X_{i\bullet}}{\sum_{\ell=1}^{J} v_{\ell}^2 w_{\ell}(\gamma(e_i(\hat{\beta}))) - e_i(\hat{\beta})^2} \right).$$

An estimate of the variance, σ_{γ}^2, of the γ_i's can be constructed as $\hat{\sigma}_{\gamma}^2(\hat{\beta}) = \frac{1}{N} \sum_{i=1}^{N} \gamma(e_i(\hat{\beta}))^2$. Then the asymptotic covariance matrix of $\hat{\beta}$ can be estimated by:

$$\widehat{Var}(\hat{\beta}) = \hat{\sigma}_{\gamma}^2(\hat{\beta})M(\hat{\beta})^{-1}(X'X)M(\hat{\beta})^{-1}.$$

Alternatively, ξ can be estimated by:

$$\hat{\xi}(\hat{\beta}) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\sum_{\ell=1}^{J} v_{\ell}^2 w_{\ell}(\gamma(e_i(\hat{\beta}))) - e_i(\hat{\beta})^2}.$$

Then:

$$\widehat{Var}(\hat{\beta}) = \frac{\hat{\sigma}_{\gamma}^2(\hat{\beta})}{\hat{\xi}^2(\hat{\beta})}(X'X)^{-1}.$$

4.1. Asymptotically Normal Tests

Because $T_Z = \dfrac{\hat{\beta}_k - \beta_k^0}{\sqrt{\widehat{Var}(\hat{\beta})_{kk}}}$ is asymptotically $N(0,1)$ under the null hypothesis $H_0 : \beta_k = \beta_k^0$, the statistic T_z can be used to test hypotheses about the values of the β_k's.

4.2. Wald Tests

Wald tests of linear restrictions on the elements of β can be expressed in the usual form. Let $H_0 : R\beta = r$ be the null hypothesis to be tested, where R is a $L \times K$ matrix with rank $(R) = L \le K$. Then $\sqrt{N}(R\hat{\beta} - r) \xrightarrow{d} N(0, R(\frac{\sigma_{\gamma}^2}{\xi^2}\Omega^{-1})R')$. Thus, the Wald test statistic has a χ^2 limiting distribution as:

$$T_W = (R\hat{\beta} - r)'(R(\widehat{Var}(\hat{\beta}))R')^{-1}(R\hat{\beta} - r) \xrightarrow{d} \chi_L^2$$

under the null hypothesis H_0. Similarly, for nonlinear restrictions $g(\beta) = [0]$, where $g(\beta)$ is a continuously differentiable L-dimensional vector function with $q = \frac{\partial g(b)}{\partial b}$ and rank $(q(\beta)) = L \le K$, it follows that:

$$T_W = g(\hat{\beta})'(q(\hat{\beta})'\widehat{Var}(\hat{\beta})q(\hat{\beta}))^{-1}g(\hat{\beta}) \xrightarrow{d} \chi_L^2.$$

4.3. Likelihood Ratio Tests

To establish a pseudo-likelihood ratio test of functional restrictions on the β vector, first note that:

$$F(\hat{\beta}) - F(\beta) \xrightarrow{d} \frac{1}{2}\left(\frac{1}{\sqrt{N}}G(\beta)\right)'\frac{1}{\xi}\Omega^{-1}\left(\frac{1}{\sqrt{N}}G(\beta)\right),$$

which follows from Equations (7) and (10) and the fact that $\frac{1}{N}H(\beta) \xrightarrow{p} -\xi\Omega$. Thus:

$$\frac{2\hat{\xi}(\hat{\beta})}{\hat{\sigma}_\gamma^2(\hat{\beta})}\left(F(\hat{\beta}) - F(\beta)\right) \xrightarrow{d} \chi_K^2.$$

Now let $\hat{\beta}_R$ be a restricted GME estimator of β. Thus, $\hat{\beta}_R = \underset{b:Rb=r}{\arg\max}(F(b))$ for a linear null hypothesis $H_0 : R\beta = r$, or $\hat{\beta}_R = \underset{b:g(b)=0}{\arg\max}(F(b))$ for a general null hypothesis $H_0 : g(\beta) = [0]$. As before, let $L = \text{rank}\ (R) \le K$ for a linear hypothesis or $L = \text{rank}\ (q(\beta)) \le K$ for a general hypothesis.

Then:

$$\frac{2\hat{\xi}(\hat{\beta})}{\hat{\sigma}_\gamma^2(\hat{\beta})}\left(F(\hat{\beta}) - F(\hat{\beta}_R)\right) \xrightarrow{d} \chi_L^2$$

under the null hypothesis.

Lagrange Multiplier Tests

Define *R, r, g, J,* and $\hat{\beta}_R$ as above. Then a Lagrangian multiplier test of functional restrictions on β can be based on the fact that:

$$\frac{1}{\hat{\sigma}_\gamma^2(\hat{\beta}_R)}G(\hat{\beta}_R)'(X'X)^{-1}G(\hat{\beta}_R) \xrightarrow{d} \chi_L^2$$

under the null hypothesis.

5. Monte Carlo Simulations

A Monte Carlo experiment was conducted to explore the sampling behavior of test situations based on the Generalized Maximum Entropy Estimator. The data were generated based on a linear model containing an intercept term, a dichotomous explanatory variable, and two continuously measured explanatory variables. The results of the Monte Carlo experiment also add additional perspective to simulation results relating the bias and mean square error to the maximum entropy estimator generated previously by [10].

The linear model $Y = X\beta + \varepsilon$ is specified as $Y = 2 + 1X_{\bullet 1} - 1X_{\bullet 2} + 3X_{\bullet 3} + \varepsilon$, where $X_{\bullet 1}$ is a discrete random variable such that $X_{i1} \overset{iid}{\sim}$ Bernoulli(.5), observations on the pair of explanatory random variables (X_{i2}, X_{i3}) are generated from iid outcomes of $N\left(\begin{pmatrix}2\\5\end{pmatrix}, \begin{pmatrix}1 & .5\\.5 & 1\end{pmatrix}\right)$ that are censored at the mean ± 3 standard deviations, and outcomes of the disturbance term are defined as $\varepsilon = \left(\sum_{i=1}^{12} U_i\right) - 6$, where $U_i \overset{iid}{\sim}$ Uniform(0,1). The support points for the disturbance terms were specified as $V = (-10, 0, 10)'$ (recall C2 and C3) for all experiments. Three different sets of support points were specified for the β-vector, given by:

$$Z_I = \begin{pmatrix} -2 & 2 & 6 \\ -3 & 1 & 5 \\ -5 & -1 & 3 \\ -1 & 3 & 7 \end{pmatrix},$$

$$Z_{II} = \begin{pmatrix} -3 & 1 & 5 \\ -4 & 0 & 4 \\ -4 & 0 & 4 \\ 0 & 4 & 8 \end{pmatrix},$$

and:

$$Z_{III} = \begin{pmatrix} -10 & 0 & 10 \\ -10 & 0 & 10 \\ -10 & 0 & 10 \\ -10 & 0 & 10 \end{pmatrix}$$

(recall C1). The support points in Z_I were chosen to be most favorable to the GME estimator, where the elements of the true β-vector are located in the center of their respective supports and the widths of the supports are relatively narrow. The supports represented by Z_{II} are tilted to the left of β_1 and β_2 and to the right of β_3 and β_4 by 1 unit, with the widths of the supports being the same as their counterparts in Z_I. The last set of supports represented by Z_{III} are wider and effectively define an upper bound of 10 on the absolute values of each of the elements of β.

To explore the respective sizes of the various tests presented in Section IV, the hypothesis $H_0 : \beta_2 = c$ was tested using the T_Z test, and the hypothesis $H_0 : \beta_2 = c, \beta_3 = d$ was tested using the Wald, pseudo-likelihood, and Lagrange Multiplier tests, with c and d set equal to the true values of β_2 and β_3, i.e., $c = 1$ and $d = -1$. Critical values of the tests were based on their respective asymptotic distributions and a 0.05 level of significance. An observation on the power of the respective tests was obtained by performing a test of significance whereby $c = d = 0$ in the preceding hypotheses. All scenarios were analyzed using 10,000 Monte Carlo repetitions, and sample sizes of $n = 25, 100, 400$, and 1,600 were examined. In the course of calculating values of the test statistics, both unrestricted and restricted (by $\beta_2 = c$ and/or $\beta_3 = d$) GME estimators needed to be calculated. Therefore, bias and mean square error measures relating to these and the least squares estimators were calculated as well. Monte Carlo results for the test statistics and for the unrestricted GME and OLS estimators are presented in Tables 1 and 2, respectively, while results relating to the restricted GME and OLS estimators are presented in Table 3. Because the choice of which asymptotic covariance matrix to use in calculating the T_Z and Wald tests was inconsequential, only the results for the second suggested covariance matrix representation are presented here.

Regarding properties of the test statistics, their behavior under a true H_0 is consistent with the behavior expected from the respective asymptotic distributions when n is large (sample size of 1600), their sizes being approximately .05 regardless of the choice of support for β. The sizes of the tests remain within 0.01 of their asymptotic size when n decreases to 400, except for the Lagrange Multiplier test under support Z_{II}, which has a slightly larger size. Across all support choices and ranging over all sample sizes from small to large, the sizes of the T_Z and Wald tests remain in the 0−0.10 range; for Z_I supports and small sample sizes, the sizes of the tests are substantially less than 0.05. Results were similar for the pseudo-likelihood and Lagrange Multiplier tests, except for the cases of Z_{II} support and $n \leq 100$, where the size of the test increased as high as 0.36 for the pseudo-likelihood test and 0.73 for the Lagrange multiplier test when $n = 25$.

Table 1. Rejection Probabilities for True $(\beta_2 = 1, \beta_3 = -1)$ and False $(\beta_2 = \beta_3 = 0)$ Hypotheses.

Supports	T_z		WALD		Pseudo-Likelihood		Lagrange Multiplier	
	H_0		H_0		H_0		H_0	
Z_I			$\beta_2 = 1$ $\beta_3 = -1$	$\beta_2 = 0$ $\beta_3 = 0$	$\beta_2 = 1$ $\beta_3 = -1$	$\beta_2 = 0$ $\beta_3 = 0$	$\beta_2 = 1$ $\beta_3 = -1$	$\beta_2 = 0$ $\beta_3 = 0$
	$\beta_2 = 1$	$\beta_2 = 0$						
$n = 25$	0.000	0.825	0.004	0.998	0.021	1.000	0.059	1.000
$n = 100$	0.017	0.999	0.022	1.000	0.038	1.000	0.056	1.000
$n = 400$	0.041	1.000	0.042	1.000	0.048	1.000	0.053	1.000
$n = 1600$	0.047	1.000	0.046	1.000	0.049	1.000	0.050	1.000
Z_{II}								
$n = 25$	0.101	0.047	0.080	0.894	0.357	0.980	0.734	0.995
$n = 100$	0.085	0.996	0.067	1.000	0.114	1.000	0.172	1.000
$n = 400$	0.053	1.000	0.048	1.000	0.058	1.000	0.066	1.000
$n = 1600$	0.052	1.000	0.052	1.000	0.055	1.000	0.057	1.000
Z_{III}								
$n = 25$	0.038	0.670	0.070	0.967	0.097	0.980	0.088	0.972
$n = 100$	0.045	0.999	0.050	1.000	0.057	1.000	0.052	1.000
$n = 400$	0.045	1.000	0.050	1.000	0.051	1.000	0.050	1.000
$n = 1600$	0.051	1.000	0.051	1.000	0.052	1.000	0.051	1.000

The powers of the tests were all substantial in rejecting false null hypotheses except for the T_Z test in the case of Z_{II} support and the smallest sample size, the latter result being indicative of a notably biased test. Overall, the choice of support did impact the power of tests for rejecting the errant hypotheses, although the effect was small for all but the T_Z test.

In the case of unrestricted estimators and the most favorable support choice (Z_I), the GME estimator dominated the OLS estimator in terms of MSE, and GME superiority was substantial for sample sizes of $n \leq 100$ (Table 2). The GME-Z_I estimator and, of course, the OLS estimator, were unbiased, with the GME-Z_I estimator exhibiting substantially smaller variances for smaller n. The choice of support has a significant effect on the bias and MSE of the GME estimator for small sample sizes. Neither the GME-Z_{II} or GME-Z_{III} estimator dominates the OLS estimator, although the GME-Z_{III} estimator is generally the better estimator across the various sample sizes. When $n = 25$, the GME-Z_{II} estimator offers notable improvement over OLS for estimating three of the four elements of β, but is significantly worse for estimating β_2. For larger sample sizes, the GME-Z_{II} estimator is generally inferior to the OLS estimator. Although the centers of the Z_{III} support are on average further from the true β's than are the centers of the Z_{II} support, the wider widths of the former result in a superior GME estimator.

The results for the restricted GME estimators in Table 3 indicate that under the errant constraints $\beta_2 = \beta_3 = 0$, the GME dominates the OLS estimator for all sample sizes and for all support choices. The superiority of the GME estimator is substantial for smaller sample sizes, but dissipates as sample size increases. The results suggest a misspecification robustness of the GME estimator that deserves further investigation.

Table 2. $E(\hat{\beta}_i)$ and Mean Square Error Measures–Unrestricted Estimators.

Estimator	$\beta_1 = 2$		$\beta_2 = 1$		$\beta_3 = -1$		$\beta_4 = 3$	
	$E(\hat{\beta}_1)$	MSE	$E(\hat{\beta}_2)$	MSE	$E(\hat{\beta}_3)$	MSE	$E(\hat{\beta}_4)$	MSE
GME-Z_I								
$n = 25$	2.000	0.015	1.001	0.038	−1.001	0.028	3.000	0.006
$n = 100$	2.003	0.034	1.003	0.026	−1.000	0.011	2.999	0.004
$n = 400$	2.000	0.032	1.001	0.009	−1.000	0.003	3.000	0.002
$n = 1600$	2.000	0.014	1.000	0.002	−1.000	0.001	3.000	0.001
GME-Z_{II}								
$n = 25$	1.022	0.977	0.484	0.309	−0.840	0.058	3.182	0.040
$n = 100$	1.306	0.519	0.826	0.056	−0.966	0.013	3.139	0.023
$n = 400$	1.672	0.141	0.960	0.010	−0.996	0.003	3.066	0.006
$n = 1600$	1.892	0.026	0.991	0.002	−1.000	0.001	3.022	0.001
GME-Z_{III}								
$n = 25$	1.278	0.757	0.946	0.131	−0.881	0.069	3.092	0.028
$n = 100$	1.709	0.252	0.995	0.037	−0.978	0.014	3.046	0.011
$n = 400$	1.914	0.068	0.999	0.010	−0.996	0.003	3.015	0.003
$n = 1600$	1.978	0.017	0.999	0.002	−0.999	0.001	3.004	0.001
OLS								
$n = 25$	1.997	1.342	1.002	0.181	−1.002	0.066	3.001	0.065
$n = 100$	2.009	0.283	1.003	0.041	−1.000	0.014	2.998	0.014
$n = 400$	2.001	0.068	1.001	0.010	−1.000	0.003	3.000	0.003
$n = 1600$	2.000	0.017	1.000	0.003	−1.000	0.001	3.000	0.001

Table 3. $E(\hat{\beta}_i)$ and Mean Square Error Measures – Restricted Estimators Under the Errant Restriction $\beta_2 = \beta_3 = 0$.

Estimator	$\beta_1 = 2$		$\beta_4 = 3$	
	$E(\hat{\beta}_1)$	MSE	$E(\hat{\beta}_4)$	MSE
GME-Z_I				
$n = 25$	2.078	0.041	2.681	0.011
$n = 100$	2.340	0.191	2.630	0.142
$n = 400$	2.689	0.537	2.600	0.196
$n = 1600$	2.898	0.832	2.520	0.232
GME-Z_{II}				
$n = 25$	1.064	0.915	2.885	0.018
$n = 100$	1.603	0.234	2.772	0.056
$n = 400$	2.330	0.169	2.630	0.140
$n = 1600$	2.776	0.628	2.543	0.210
GME-Z_{III}				
$n = 25$	1.686	0.589	2.750	0.084
$n = 100$	2.468	0.542	2.601	0.172
$n = 400$	2.842	0.823	2.530	0.225
$n = 1600$	2.958	0.948	2.508	0.243
OLS				
$n = 25$	3.011	3.342	2.497	0.342
$n = 100$	3.013	1.575	2.497	0.274
$n = 400$	3.005	1.138	2.499	0.256
$n = 1600$	2.999	1.030	2.500	0.251

Asymmetric Error Supports

We present further Monte Carlo simulations to show that regularity condition R2, which assumes symmetry of the disturbance term, is not a necessary condition for identification of the GME slope parameters. It is demonstrated below that if the supports of the error distribution asymmetric, then only the intercept term of the GME regression estimator is asymptotically biased.

The Monte Carlo experiments that follow are identical to those above except for specification of the user supplied support points for the error terms and the underlying true error distribution. To illustrative the impact of asymmetric errors, experiments are based on one set of support points symmetric about zero, $V_I = (-10,0,10)'$, and two sets of support points not symmetric about zero, $V_{II} = (-5,5,15)'$ and $V_{III} = (-5,0,15)'$. The support V_{II} is a simple translation of V_I by five positive units in magnitude and retaining symmetry centered about 5. The asymmetric support V_{III} translates the truncation points by five positive units in magnitude, but retains the center support point 0. The true error distribution is generated in two ways: a symmetric distribution specified as a N(0,1) distribution truncated at $(-3,3)$ and an asymmetric distribution specified as a Beta(3,2) translated and scaled from support (0,1) to $(-3,3)$ with mean 0.6. Supports on the parameter coefficients terms are retained as Z_I, providing symmetric support points about the true coefficient values.

Monte Carlo experiments presented in Table 4 and 5 are generated for sample sizes 25, 100, and 400 with 1,000 replications for each sample size. Consider when the true distribution is symmetric about zero. Slope coefficients for error supports that are not symmetric about zero appear biased in smaller sample sizes. However, the bias and MSE of the slope coefficients decrease as the sample sizes increases. Next, suppose the true distribution is asymmetric. For symmetric and asymmetric supports only the intercept terms are persistently biased, diverging from the true parameter values as the sample size increases. These results demonstrate the robustness of GME slope coefficients to asymmetric error distributions and user supplied supports.

Table 4. Mean and MSE of 1,000 Monte Carlo Simulations with True Distribution Symmetric. Symmetric and Asymmetric Error Supports and Coefficient Support Z_I.

Estimator	$\beta_1 = 2$		$\beta_2 = 1$		$\beta_3 = -1$		$\beta_4 = 3$	
	$E(\beta_1)$	MSE	$E(\beta_2)$	MSE	$E(\beta_3)$	MSE	$E(\beta_4)$	MSE
GME-Z_I,V_I								
25	2.002	0.016	1.003	0.042	−1.000	0.030	2.997	0.007
100	2.000	0.033	1.001	0.026	−1.002	0.011	3.002	0.004
400	2.000	0.035	1.001	0.010	−0.998	0.003	2.999	0.002
GME-Z_I,V_{II}								
25	1.259	0.585	0.815	0.101	−1.009	0.048	2.209	0.636
100	0.208	3.258	0.804	0.071	−0.944	0.020	2.381	0.389
400	−1.144	9.903	0.868	0.028	−0.959	0.005	2.640	0.132
GME-Z_I,V_{III}								
25	1.506	0.271	0.875	0.069	−1.005	0.038	2.476	0.282
100	0.752	1.598	0.875	0.045	−0.961	0.015	2.602	0.163
400	−0.235	5.024	0.925	0.015	−0.977	0.004	2.794	0.044
OLS								
25	2.014	1.321	1.007	0.204	−0.998	0.069	2.993	0.065
100	1.999	0.280	1.001	0.042	−1.002	0.014	3.002	0.014
400	2.001	0.075	1.001	0.011	−0.997	0.003	2.999	0.003

Table 5. Mean and MSE of 1000 Monte Carlo Simulations with True Distribution Asymmetric. Symmetric and Asymmetric Error Supports and Coefficient Support Z_I.

Estimator	$\beta_1=2$		$\beta_2=1$		$\beta_3=-1$		$\beta_4=3$	
	$E(\beta_1)$	MSE	$E(\beta_2)$	MSE	$E(\beta_3)$	MSE	$E(\beta_4)$	MSE
GME-Z_I,V_I								
25	2.089	0.031	1.038	0.060	−1.005	0.041	3.094	0.018
100	2.233	0.108	1.023	0.033	−1.006	0.016	3.071	0.010
400	2.427	0.229	1.015	0.012	−1.004	0.005	3.033	0.004
GME-Z_I,V_{II}								
25	1.358	0.449	0.843	0.103	−1.021	0.057	2.305	0.496
100	0.410	2.583	0.826	0.073	−0.966	0.019	2.463	0.294
400	−0.860	8.209	0.890	0.025	−0.966	0.006	2.700	0.092
GME-Z_I,V_{III}								
25	1.597	0.190	0.905	0.075	−1.019	0.049	2.574	0.193
100	0.964	1.129	0.889	0.055	−0.967	0.020	2.674	0.112
400	0.126	3.553	0.946	0.016	−0.981	0.005	2.835	0.030
OLS								
25	2.600	2.324	1.041	0.261	−1.009	0.097	2.998	0.099
100	2.616	0.813	1.001	0.052	−0.999	0.020	2.997	0.021
400	2.610	0.471	1.003	0.013	−1.000	0.005	2.997	0.005

6. Further Results

Unbiased GME Estimation. It is apparent from the proof of the theorem in Section 3 that the $-\sum_{\ell=1}^{J_k} p_{k\ell}\ln(p_{k\ell})$ terms are asymptotically uninformative. It is instructive to note that if these terms are deleted from the GME objective function and the resulting objective function is then maximized through choosing b and w subject to constraints C2–C4 and C6, the resulting GME estimator is in fact *unbiased* for estimating β. This follows because the ε_i's are iid mean zero and symmetrically distributed around zero, and the new estimator, say $\tilde{\beta}$, is such that $\tilde{\beta}-\beta$ is a symmetric function of the ε_i's.

Bayesian Analogues. As pointed out by [35] maximum entropy methods can be motivated as an empirical Bayes rule. We expand on their analogy by noting a strong formal parallel to the traditional Bayesian framework of inference. In particular, one can view $-\sum_{k=1}^{K}\sum_{\ell=1}^{J_k} p_{k\ell}\ln(p_{k\ell})$ as the maximum entropy analogue to the log of a non-normalized Bayesian prior and $-\sum_{i=1}^{N}\sum_{\ell=1}^{J} w_{i\ell}\ln(w_{i\ell})$ as the maximum entropy analogue to the non-normalized log of the probability density kernel or log-likelihood function. For any given set of support points Z and V, we can define functions f_{β_k} and f_ε by:

$$f_{\beta_k}(b_k) = \frac{e^{-\sum_{\ell=1}^{J_K} p_{k\ell}(b_k)\ln(p_{k\ell}(b_k))}}{\int_{z_{k1}}^{z_{kJ_k}} e^{-\sum_{\ell=1}^{J_K} p_{k\ell}(\tau_k)\ln(p_{k\ell}(\tau_k))} d\tau_k} \text{ and } f_\varepsilon(x) = \frac{e^{-\sum_{\ell=1}^{J} w_{k\ell}(x)\ln(w_\ell(x))}}{\int_{v_1}^{v_J} e^{-\sum_{\ell=1}^{J} w_\ell(y)\ln(w_\ell(y))} dy}$$

Then for $\varepsilon \overset{iid}{\sim} f_\varepsilon$, the maximum likelihood estimator of β is $\tilde{\beta}$, and if one adds priors $\beta_k \overset{ind}{\sim} f_{\beta_k}$, then $\hat{\beta}$

is the Bayesian posterior mode estimator of β. We note the following consequences of these equivalences. First, if the support points $v_1, \ldots, v_J$ can be chosen so that f_ε is very close to the true distribution of ε, then the GME estimator should be nearly asymptotically efficient. Second, in finite samples the prior information influences $\hat{\beta}$ such that $\hat{\beta}$ is generally not unbiased. Third, the support points used in the GME estimator have no particular relationship to the points of support of the distribution of a discrete random variable. The distributions f_ε and f_{β_k} are absolutely continuous for any choice of Z and V.

The previous Monte Carlo results illustrate the Bayesian-like character of the maximum entropy results. The GME with reasonably narrow points of support centered on the true values of β dominated the OLS estimator and was sometimes far better. On the other hand, the GME performed poorly when the points of support were similarly narrow and mis-centered by only one-eighth the range of the points of support. In the latter case, mean squared errors were often much worse than OLS and biases were often substantial. Finally, wider points of support, even though they were the most mis-centered of the cases examined, were quite similar to OLS results for moderate to large sample sizes, and provided some degree of improvement over OLS for small samples.

Finally, the GME approach is a special case of generalized cross entropy, which incorporates a reference probability distribution over support points. This allows a direct method of including prior information, akin to a Bayesian framework. However, in a classical sense, the empirical estimation strategies are inherently different.

GME Calculation Method. The conditional maximum entropy formulation (2) utilized in the proof of asymptotic results represents the basis for a computationally efficient method of obtaining GME estimates. In particular, maximizing $F(\tau)$ through choice of τ involves a nonlinear search over a vector of relatively low dimension (K) as opposed to searching over the ($KM + NJ$) dimensional space of (p,w) values. In the process of concentrating the objective function, note that the needed Lagrange multiplier functions $\eta_k(\tau_K)$ and $\gamma(e_i(\tau))$ can be expressed as elementary functions for three support points or less, and still exist in closed form (using inverse hyperbolic functions) for support vectors having five elements. As a point of comparison, the calculation of GME estimates in the Monte Carlo experiment with N = 1,600 was completed in a matter of seconds on a 133 mhz personal computer. Such a calculation would be intractable, let alone efficient, in the space of (p,w) values. We note further that the dual algorithm of [10] would still involve a search over a space of dimension N = 1,600, which would be infeasible here and in other problems in which the number of data points is large.

7. Conclusions

We have shown that the data-constrained GME estimator of the GLM is consistent and asymptotically normal as long as the coefficients and errors obey the constraints of the constrained maximum entropy problem. Furthermore, we have demonstrated the possibility that the GME estimator can be asymptotically efficient. Thus, depending on the distribution of the errors, GME may be more or less efficient than alternatives such as least squares. We performed Monte Carlo tests showing that the quality of the GME estimates depends on the quality of the supports chosen. The Monte Carlo results suggest that GME with wide supports will often perform better than OLS while providing some robustness to misspecification.

We have shown how all the conventional types of asymptotic tests can be calculated for GME estimates. In the Monte Carlo study these asymptotic tests performed extremely well in samples of 400 or more. In smaller samples the tests performed less well, particularly when the supports were narrow, although some of the results were quite acceptable. We have also demonstrated that all our results can be applied to a maximum cross-entropy estimator. While our focus has been on asymptotic properties, we have also shown how the entropy terms involving the coefficients play a role analogous to a Bayesian prior. Furthermore, these terms are asymptotically uninformative and can be omitted if the researcher wishes to use an unbiased GME estimator.

References

1. Imbens, G.; Spady, R.; Johnson, P. Information Theoretic Approaches to Inference in Moment Condition Models. *Econometrica* **1998**, *66*, 333–357.
2. Imbens, G. A New Approach to Generalized Method of Moments Estimation, Harvard Institute of Economic Research Discussion Paper No. 1633; Harvard University: Cambridge, MA, USA, 1993.
3. Kitamura, Y.; Stutzer, M. An Information-Theoretic Alternative to Generalized Method of Moments Estimation. *Econometrica* **1997**, *65*, 861–874.
4. Cressie, N.; Read, T.R.C. Multinomial goodness of fit tests. *J. Roy. Stat. Soc B* **1984**, *46*, 440–464.
5. Pompe, B. On Some Entropy Measures in Data Analysis. *Chaos Solutions Fractals* **1994**, *4*, 83–96.
6. Seidenfeld, T. Entropy and Uncertainty. *Philo. Sci.* **1986**, *53*, 467–491.
7. Judge, G.G.; Mittelhammer, R.C. *An Information Theoretic Approach to Econometrics*; Cambridge University Press: Cambridge, UK, 2012.
8. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.
9. Jaynes, E.T. Information Theory and Statistical Mechanics. In *Statistical Physics*; Ford, K., Ed.; Benjamin: New York, NY, USA, 1963; p. 181.
10. Golan, A.; Judge, G.; Miller, D. *Maximum Entropy Econometrics: Robust Estimation with Limited Data*; Wiley & Sons: New York, NY, USA, 1996.
11. Zellner, A.; Highfield, R.A. Calculation of maximum entropy distribution and approximation of marginal posterior distributions. *J. Econometrics* **1998**, *37*, 195–209.
12. Soofi, E. Information Theoretic Regression Methods. In *Applying Maximum Entropy to Econometric Problems (Advances in Econometrics)*; Fomby, T., Hill, R.C., Eds.; Emerald Group Publishing Limited: London, UK, 1997.
13. Ryu, H.K. Maximum entropy estimation of density and regression functions. *J. Econometrics* **1993**, *56*, 397–440.
14. Golan, A.; Judge, G.; Perloff, J.M. A maximum entropy approach to recovering information from multinomial response data. *JASA* **1996**, *91*, 841–853.
15. Vinod, H.D. Maximum Entropy Ensembles for Time Series Inference in Economics. *Asian Econ.* **2006**, *17*, 955–978.
16. Holm, J. Maximum entropy Lorenz curves. *J. Econometrics* **1993**, *59*, 377–389.
17. Marsh, T.L.; Mittelhammer, R.C. Generalized Maximum Entropy Estimation of a First Order Spatial Autoregressive Model. In *Spatial and Spatiotemporal Econometrics (Advances in Econometrics)*; Pace, R.K., LeSage, J.P., Eds; Emerald Group Publishing Limited: London, UK, 2004.

18. Krebs, T. Statistical Equilibrium in One-Step Forward Looking Economic Models. *JET* **1997**, *73*, 365–394.

19. Golan, A.; Judge, G.; Karp, L. A maximum entropy approach to estimation and inference in dynamic models or counting fish in the sea using maximum entropy. *JEDC* **1996**, *20*, 559–582.

20. Kattuman, P.A. On the size Distribution of Establishments of Large Enterprises: An Analysis for UK Manufacturing; University of Cambridge: Cambridge, UK, 1995.

21. Callen, J.L.; Kwan, C.C.Y.; Yip, P.C.Y. Foreign-Exchange Rate Dynamics: An Empirical Study Using Maximum Entropy Spectral Analysis. *J. Bus. Econ. Stat.* **1985**, *3*, 149–155.

22. Bellacicco, A.; Russo, A. Dynamic Updating of Labor Force Estimates: JARES. *Labor* **1991**, *5*, 165–175.

23. Sengupta, J.K. The maximum entropy approach in production frontier estimation. *Math. Soc. Sci.* **1992**, *25*, 41–57.

24. Fraser, I. An application of maximum entropy estimation: The demand for meat in the United Stated Kingdom. *Appl. Econ.* **2000**, *32*, 45–59.

25. Lev, B.; Theil, H. A Maximum Entropy Approach to the Choice of Asset Depreciation. *J. Accounting Res.* **1978**, *16*, 286–293.

26. Stuzer, M. A simple nonparametric approach to derivative security valuation. *J. Financ.* **1996**, *51*, 1633–1652.

27. Buchen, P.W.; Kelly, M. The Maximum Entropy Distribution of an Asset Inferred from Option Prices. *J. Financ. and Quant. Anal.* **1996**, *31,* 143-159.

28. Preckel, P.V. Least squares and entropy: A penalty function perspective. *Am. J. Agr. Econ.* **2001**, *83*, 366–377.

29. Paris, Q.; Howitt, R. An Analysis of Ill-Posed Production Problems Using Maximum Entropy. *Am. J. Agr. Econ.* **1998**, *80*, 124–138.

30. Lence, S.H.; Miller, D.J. Recovering Output-Specific Inputs from Aggregate Input Data: A Generalized Cross-Entropy Approach. *Am. J. Agr. Econ.***1998**, *80*, 852–867.

31. Miller, D.J.; Plantinga, A.J. Modeling Land Use Decisions with Aggregate Data. *Am. J. Agr. Econ.* **1999**, *81*, 180–194.

32. Fernandez, L. Recovering Wastewater Treatment Objectives: An Application of Entropy Estimation for Inverse Control Problems. In *Advances in Econometrics, Applying Maximum Entropy to Econometric Problems*; Fomby, T., Hill, R.C., Eds.; Jai Press Inc.: London, UK, 1997.

33. White, H. *Asymptotic Theory for Econometricians*; Academic Press: New York, NY, USA, 1984.

34. Rao, C.R. *Linear Statistical Inference and Its Applications*, 2nd ed; Wiley & Sons: New York, NY, USA, 1973.

35. Miller, D.; Judge, G.; Golan, A. *Robust Estimation and Conditional Inference with Noisy Data*; University of California: Berkeley, CA, USA, 1996.