Hacettepe Journal of Mathematics and Statistics
Volume 40 (3) (2011), 469 – 481

AN INFORMATION-THEORETIC ALTERNATIVE TO MAXIMUM LIKELIHOOD ESTIMATION METHOD IN ULTRASTRUCTURAL MEASUREMENT ERROR MODEL

Amjad D. Al-Nasser*†

Received 03 : 04 : 2009 : Accepted 06 : 10 : 2010

Abstract

In this paper, a data constrained generalized maximum entropy (GME) estimator for the general linear measurement error model is proposed. GME estimation, as developed by (A. Golan, G. Judge and D. Miller *A Maximum Entropy Econometrics: Robust Estimation with limited data* (Wiley, New York, 1996)), was formulated as a convex mixed-integer nonlinear optimization problem. Shannon entropy measures and its generalization, namely 'entropy of order r' by Tsallis and Rényi are briefly discussed. A Monte Carlo comparison is made with the classical maximum likelihood estimation (MLE) method. The results show that, with moderate sample size; the GME outperforms the MLE estimators in terms of mean squared error.

Keywords: Measurement error model, Generalized maximum entropy, Maximum likelihood, Entropy of order r.

2000 AMS Classification: 94 A 17, 62 H 12, 62 J 12.

*Department of Statistics, Yarmouk University, 21163 Irbid, Jordan.
E-mail: amjadyu@yahoo.com
†Current Address: Division of Area Planning & Residents Relations, WRM, Emirate of Abu Dhabi, UAE.

1. Introduction

The traditional maximum entropy formulation is based on the entropy-information measure of Shannon [27] to reflect the uncertainty about the occurrence of a collection of events. It is developed and described in Jaynes [18, 19] and used to recover the unknown probability distribution of underdetermined problems. Suppose we have a set of events $\{x_1, x_2, \ldots, x_K\}$ whose probabilities of occurrence are $p_1, p_2, \ldots, p_K$. Then using an axiomatic method to define a unique function to measure the uncertainty of a collection of events, Shannon [27] defines the entropy of the distribution (discrete events), as the average of self-information $H(P) = -\sum_{i=1}^{K} p_i \ln(p_i)$, where $0 \ln(0) = 0$. Thereafter, many entropy measures have been proposed as a generalization of Shannon's entropy, the most well known generalizations are Rényi's entropy measure and the Tsallis entropy measure, which were later known as entropy measures of order r.

Since the 1990's many attempts have been made to apply the method of maximum entropy in the area of linear models. In 1996, Golan *et al.* [14] proposed an estimator based on the maximum entropy formalism of Jaynes that they called the generalized maximum entropy (GME) estimator, by using a dual objective function. The logic of using the GME estimation method is that GME tends to dominate traditional estimation methods with small samples, it does not rely on any distributional assumption, also it is robust in fitting nonlinear models (Golan, [13]; Peeter, [23]); and it is also robust in case of strong collinearity of the independent variables (Paris, [22]; Ciavolino and Al-Nasser, [8]). The idea underlying the GME approach is to view each unknown parameter, and error term as an expected value of some proper probability distribution; then by maximizing the joint entropies subject to the data, represented by each unobserved value, and the requirement for proper probability distributions; better estimates can be achieved with less assumptions (Csiszar [10], Donho *et al.* [12], Golan, *et al.* [15], Al-Nasser [4], Al-Nasser [3], Golan [13], Caputo and Paris [5]; Ciavolino and Al-Nasser [8], Ciavolino and Dahlgaard [9], Al-Nasser [1]).

The remainder of this paper is divided into five sections. Section 2 presents the Ultrastructural Model. Section 3 presents the entropy measures and their generalizations. In Section 4 a generalized maximum entropy estimation approach idea is introduced to the ultrastructural model, Section 5 presents Monte Carlo evidence on the numerical performance of GME and maximum likelihood estimation (MLE), and the last section presents some concluding comments.

2. The ultrastructural ME model

Consider the simple linear relationship between two mathematical variables ξ and η

$$\eta = \alpha + \beta\xi,$$

where α is the intercept and β is the slope. The classical theory of regression analysis assumes that these variables are measured without error; particularly in the social sciences and natural science this assumption is often violated. Hence, this linear relationship is reformulated such that both variables are contaminated with measurement errors. Then the observed values can be defined by:

$$\begin{aligned} x_i &= \xi_i + \delta_i, \\ y_i &= \eta_i + \varepsilon_i, \ i = 1, 2, \ldots, n \end{aligned} \tag{1}$$

where δ and ε are the measurement errors associated with ξ and η, respectively. It is assumed that $\delta_1, \delta_2, \ldots, \delta_n$ are identically and independently distributed (*i.i.d*) with mean 0 and variance σ_δ^2. Similarly, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are *i.i.d* with mean 0 and variance σ_ε^2.

Further, suppose that the true values of the variable (ξ) have possibly different means; say $\mu_1, \mu_2, \ldots, \mu_n$, so that we can write

$$\xi_i = \mu_i + \omega_i, \ i = 1, 2, \ldots, n, \tag{2}$$

where $\omega_1, \omega_2, \ldots, \omega_n$ are *i.i.d* with mean 0 and variance σ_ω^2. Finally, assume that the distribution of $(\delta_1, \delta_2, \ldots, \delta_n)$, $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ and $(\omega_1, \omega_2, \ldots, \omega_n)$ are mutually independent of each other. This provides the specification of the ultrastructural model.

This model can be reduced to the functional measurement error model if we assume that the ξ_i are fixed; i.e. $\omega_i = 0, \ i = 1, 2, ...n$. Also it can be reduced to the structural measurement model if we assume $\mu_i = \mu_j, \ \forall i, j = 1, 2, \ldots, n$. On the other hand, if $\delta_i = 0, \ i = 1, 2, ...n$, then the ultrastructural model reduces to the classical regression model with no measurement error; for more details see, Dolby (1976) and Cheng and Van Ness (1999).

Assuming Normal measurement errors, then the MLE for the parameters of the ultrastructural model are unidentifiable, Srivastava and Shalabh [26]. Hence, to solve the ultrastructural model (1-2), the MLE method required additional assumptions; Dolby [11] derived the MLE estimates when the ratios $\lambda = \frac{\sigma_\varepsilon^2}{\sigma_\delta^2}$, and $\nu = \frac{\sigma_\omega^2}{\sigma_\delta^2}$ are known. Then, under the customary assumptions the MLE for the ultrastructural model results in a quintic equation for $\hat{\beta}$;

$$h(\hat{\beta}) = \nu S_{xx}\hat{\beta}^5 + (3\nu\lambda S_{xx} - \nu S_{yy})\,\hat{\beta}^3 - 2\lambda\,(\nu - 1)\,S_{xy}\hat{\beta}^2 + \left\{2\lambda^2\,(\nu + 1)\,S_{xx} - \lambda\,(\nu + 2)\,S_{xx}\right\}\hat{\beta} - 2\lambda^2\,(\nu + 1)\,S_{xy} = 0$$

This generally should be solved by iterative numerical methods. However, Gleser [16] showed that the likelihood is maximized on the $\sigma_\omega^2 = 0$ boundary of the parameter set. Thus, the ML estimates for the ultrastructural model are just the ML estimates of the functional model; which leads to

$$\hat{\beta} = \frac{(S_{yy} - \lambda S_{xx}) + ((S_{yy} - \lambda S_{xx})^2 + 4\lambda S_{xy})^{1/2}}{2S_{xy}}.$$

The ML solution of the other parameters will be:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \quad \hat{\mu}_i = \frac{\lambda x_i + \hat{\beta}(y_i - \hat{\alpha})}{\lambda + \hat{\beta}^2},$$

where $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2/n$, $S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2/n$, $S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})/n$, $\bar{x} = \sum_{i=1}^n x_i/n$ and $\bar{y} = \sum_{i=1}^n y_i/n$.

For more details see Cheng and Van Ness [7], Carroll *et al.* [6].

3. Entropy measures and higher order entropies

For a random vector $X = \{x_1, \ldots, x_K\}$ with probability distribution $p = \{p_1, \ldots, p_K\}$, such that $\sum_{i=1}^K p_i = 1$; then any entropy measure function H should satisfy the following requirements (Kapur, [21]):

(1) It should be a function of $p_1, p_2, \ldots, p_K$; i.e., $H = H_K(p) = H_K\,(p_1, p_2, \ldots, p_K)$.
(2) The entropy measure function should be a continuous function of $p_1, p_2, \ldots, p_K$.
(3) H should be a symmetric function of its arguments.
(4) It should not change if an impossible outcome is added to the probability scheme i.e.

$$H_{K+1}\,(p_1, p_2, \ldots, p_K, 0) = H_K\,(p_1, p_2, \ldots, p_K)\,.$$

(5) $H_K\,(p_1, p_2, \ldots, p_K) = 0$ when $p_i = 1, \ p_j = 0, \ j \neq i, \ i = 1, 2, ..., K$.

(6) H_K has a maximum value when all probabilities are uniform; $p_1 = p_2 = \cdots = p_K = \frac{1}{K}$.
(7) The maximum value of H_K should increase as K increases.

3.1. Shannon Entropy. Shannon [27] defines the entropy of the distribution (discrete events) $X = \{x_1, \ldots, x_K\}$ with corresponding probabilities $p = \{p_1, \ldots, p_K\}$ as

$$H(p) = -\sum_{i=1}^{K} p_i \ln(p_i), \tag{3}$$

where $0 \ln 0 = 0$. The quantity $(-\ln(p_i))$ is called *the amount of self information of the event* x_i. More than thirty measures of entropy have been introduced in the literature on Information Theory generalizing Shannon's entropy (Taneeja, [28]). Thus the idea of generalized entropies started with Rényi [25] who characterized a scalar parametric entropy as *entropy of order* r, which includes Shannon's entropy as a limiting case. Later on Tsallis [29] proposed another generalization to distinguish a non extensive system from an extensive system.

3.2. Rényi Entropy. Rényi [25] defines a higher order entropy, or the so called *entropy of order* r, as

$$H_r^R(p) = \frac{1}{1-r} \ln \sum_k p_k^r \tag{4}$$

It is an important measure function in ecology and statistics as indices of diversity. This entropy function satisfies additional properties that are used for an information theoretical proof of the Central Limit Theorem:

(1) $H_r^R(p)$ is concave in P for $r \in [0, 1]$, but concavity breaks down for $r > 1$.
(2) $H_r^R(p)$ is decreasing in r and is additive/extensive for all r.

Later, Kapur [21] proposed a generalized Rényi's measure further to give a measure of entropy of order r and type s.

$$H_{r,s}(p) = \frac{1}{1-r} \ln \frac{\sum_{i=1}^{n} p_i^{r+s-1}}{\sum_{i=1}^{n} p_i^s}, \ r \neq 1, \ s > 0, \ r + s - 1 > 0.$$

This reduces to Rényi's measure when $s = 1$, and to Shannon's measure when $s = 1$, $r \to 1$.

3.3. Tsallis Entropy. Tsallis [29] used the r-parameter to distinguish a non extensive system from an extensive system, and defines entropy of order r as

$$H_r^T(p) = \frac{\sum_k p_k^r - 1}{1-r}, \ r > 0, \ r \neq 1.$$

Note that $H_r^T(p)$ is concave in p for $r > 1$, and in this case the system is called a *non extensive system* or a *non-classical system*. A similar non additive entropy measure proposed by Havrada and Charvat [17] obtained the first non-additive measure of entropy:

$$H^r(p) = \frac{\sum_{i=1}^{n} p_i^r}{2^{1-r} - 1}, \ r \neq 1, \ r > 0.$$

It should be noted that Tsallis' entropy is simply related to Rényi's [25] entropy; in the two cases, Shannon's entropy is obtained in the limit case as $r \to 1$.

3.4. Comparisons of Entropy Measures. To demonstrate the benefit of using the GME estimation method with different entropy measures, consider the dice problem introduced by Jaynes [20].

Suppose we have a six-sided die that can take on the values $k = 1, 2, 3, \ldots, 6$, with unknown probabilities $p = (p_1, p_2, \ldots, p_6)'$ such that the six probabilities must sum to 1. Also, suppose we have an additional piece of information μ at the beginning. The problem is clearly ill-posed or underdetermined because there are six unknown probabilities, but we only have two pieces of information.

For instance, suppose we expect the die to be roughly 'fair', and the observed average matches the mean of the discrete uniform distribution, $\mu = 3.5$. Then, many would assert that the underlying distribution is discrete uniform because the sample information matches our prior beliefs. However, if $\mu \neq 3.5$, the sample information suggests that the underlying distribution is not likely to be uniform. Then the GME is used to solve this problem by maximizing $H(p)$ subject to; $\sum_{k=1}^{6} p_k = 1$, and $\sum_{k=1}^{6} k p_k = \mu$.

The lagrangian function for this problem is

$$L = H(p) + \lambda\left(\mu - \sum_{k=1}^{6} k p_k\right) + \gamma\left(1 - \sum_{k=1}^{6} p_k\right),$$

where λ and γ are Lagrangian multipliers. Then by solving the first-order conditions, the GME probability distribution places the weights

$$\hat{p}_k = \frac{\exp\left(-k\hat{\lambda}\right)}{\sum_{k=1}^{6} \exp\left(-k\hat{\lambda}\right)},$$

$$\hat{p}_k = \frac{\left[\left(k\hat{\lambda} + \hat{\gamma}\right)\sum_{k=1}^{6} (\hat{p}_k)^{\alpha}\right]^{\frac{1}{\alpha-1}}}{\sum_{k=1}^{6}\left[\left(k\hat{\lambda} + \hat{\gamma}\right)\sum_{k=1}^{6} (\hat{p}_k)^{\alpha}\right]^{\frac{1}{\alpha-1}}}, \text{ and}$$

$$\hat{p}_k = \frac{\left[\left(k\hat{\lambda} + \hat{\gamma}\right)\right]^{\frac{1}{\alpha-1}}}{\sum_{k=1}^{6}\left[\left(k\hat{\lambda} + \hat{\gamma}\right)\right]^{\frac{1}{\alpha-1}}}$$

based on Shannon, Réyni and Tsallis; respectively.

The optimal solution of Jayne's dice problem can be obtained by using a numerical optimization package, the estimated maximum entropy distributions, for various values $\mu = 2, 3.5, 4$, are given in (Table 1 – Table 3):

Table 1. Optimal distribution of Jayen's dice problem when $\mu = 2$

		P_1	P_2	P_3	P_4	P_5	P_6	$H(p)$
Shannon		0.477	0.255	0.137	0.073	0.038	0.020	1.37
Tsallis	$r = 0.20$	0.560	0.180	0.099	0.066	0.049	0.046	3.62
	$r = 0.50$	0.530	0.205	0.114	0.069	0.047	0.034	2.37
	$r = 0.80$	0.499	0.236	0.126	0.071	0.041	0.027	1.66
Rényi	$r = 0.20$	0.554	0.185	0.102	0.067	0.052	0.040	4.87
	$r = 0.50$	0.525	0.212	0.114	0.068	0.048	0.033	4.37
	$r = 0.80$	0.502	0.232	0.126	0.069	0.043	0.027	6.66

Table 2. Optimal distribution of Jayen's dice problem when $\mu = 3.5$

		P_1	P_2	P_3	P_4	P_5	P_6	$H(p)$
Shannon		0.167	0.167	0.166	0.166	0.166	0.168	1.79
Tsallis	$r = 0.20$	0.167	0.167	0.167	0.166	0.167	0.166	3.99
	$r = 0.50$	0.167	0.167	0.166	0.166	0.167	0.167	2.90
	$r = 0.80$	0.166	0.166	0.168	0.167	0.166	0.166	2.15
Rényi	$r = 0.20$	0.167	0.166	0.166	0.167	0.167	0.167	5.24
	$r = 0.50$	0.167	0.166	0.166	0.167	0.167	0.167	4.90
	$r = 0.80$	0.167	0.167	0.167	0.166	0.166	0.167	7.15

Table 3. Optimal distribution of Jayen's dice problem when $\mu = 4.0$

		P_1	P_2	P_3	P_4	P_5	P_6	$H(p)$
Shannon		0.104	0.122	0.146	0.174	0.208	0.246	1.75
Tsallis	$r = 0.20$	0.108	0.123	0.142	0.169	0.202	0.256	3.95
	$r = 0.50$	0.106	0.123	0.143	0.172	0.203	0.252	2.85
	$r = 0.80$	0.104	0.123	0.145	0.173	0.206	0.249	2.11
Rényi	$r = 0.20$	0.108	0.125	0.140	0.166	0.205	0.255	5.20
	$r = 0.50$	0.107	0.124	0.142	0.169	0.206	0.252	4.85
	$r = 0.80$	0.105	0.121	0.148	0.173	0.204	0.250	7.11

Comparing the result for different entropy measures, it could be noted that Shannon's entropy measure produces the smallest values of the entropy function $H(p)$ for all cases. However, all entropy measures give almost the same results with fair dice when the first moment is known, and are equivalent to the theoretical case.

4. Generalized maximum entropy

To illustrate the GME estimation approach, consider the general linear model: $Y_i = f(X_i,\ \boldsymbol{\beta}) + \varepsilon_i,\ i = 1, 2, \ldots n$, where the Y_i 's are responses, f is a known function of the unknown parameter vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_K)'$ and the covariate vector $X_i = (x_{i1}, x_{i2}, \ldots, x_{iK})',\ i = 1, 2, \ldots, n$ whose values are assumed to be known, and $\varepsilon_i,\ i = 1, 2, \ldots, n$ are random errors.

In GME, the model is fitted after some reformulation of the unknown parameters $\boldsymbol{\beta}$, and the unknown error terms $\epsilon_i,\ i = 1, 2, 3, \ldots, n$, if they are not in probability format. This can be done by reparameterizing their possible outcome values probabilistically as a convex combination of random variables. This combination is presented as the expected value of some proper probability distribution. For each unknown, assume that there exists a discrete probability distribution that is defined over the parameter space $[0, 1]$; by a set of equally distanced discrete points; then the formulation of the model parameters will be of the form $\boldsymbol{\beta} = ZP$; where Z is a $(K \times KR)$ matrix and P is a KR-vector of weights such that $p_k > 0$ and $P_k' 1_R = 1$ for each k. Simply, each $\beta_k,\ k = 1, 2, \ldots, K$ can be defined by a set of equally distanced discrete points $Z_k' = [z_{k1}, z_{k2}, z_{k3}, \ldots, z_{kR}]$, where $R \geq 2$ with corresponding probabilities $P_k' = [p_{k1}, p_{k2}, p_{k3}, \ldots, p_{kR}] \in [0, 1]$. That is,

$$\beta_k = \sum_{r=1}^{R} z_{kr} p_{kr},\ \sum_{r=1}^{R} p_{kr} = 1,\ 0 \leq p_{kr} \leq 1,\ k = 1, 2, \ldots, K.$$

The support points Z for $\boldsymbol{\beta}$ are known. Golan *et al.* [14] suggested that these values can be specified uniformly symmetric around 0 with large values of the lower and the upper bounds. For example, one can select the support point $z = (-c, 0, c)$, given that c is a large value (i.e. $c = 100$ or 1000, ..., etc). Moreover, assuming one specifies Z to span the true values of $\boldsymbol{\beta}$, then the GME is a consistent estimator, which is an advantage of the GME estimation method.

The disturbance ϵ_i can be treated in a similar fashion. For a set of $T \geq 2$ support points ϵ_i assumed to be bounded between two finite values v_t, v_T, which are symmetric around zero and with corresponding unknown probability weights w_{1t}, w_{nT}. That is, each error term may be modeled as; $\boldsymbol{\varepsilon} = \mathbf{V}W$, where $\mathbf{V}$ is a $(n \times nT)$ matrix and $\mathbf{W}$ is a nT-dimensional vector of weights; that is to say:

$$\varepsilon_i = \sum_{j=1}^{T} v_{ij} w_{ij}, \ \sum_{j=1}^{T} w_{ij} = 1, \ 0 \leq w_{ij} \leq 1, \ i = 1, 2, \ldots, n.$$

Placing bounds for v_j is difficult in practice. Alternatively, Chebychev's inequality (see, Pukelsheim [24]) may be used as a conservative means of specifying sets of error bounds. The empirical GME literature indicates that, in general, the number of support values R for unknown parameters is 5, and the number of support values T of the error term is 3; see for example Paris [22], Al-Nasser [2], Al-Nasser [1], and Golan [13].

Now, using the reparameterized unknowns $\boldsymbol{\beta} = \mathbf{Z}P$ and $\boldsymbol{\varepsilon} = \mathbf{V}W$, we rewrite the general linear model as follows:

$$y = f(x, ZP) + VW$$

Then the maximum entropy principle may be stated in terms of scalar summations with two nonnegative probability components, and the GME estimators can be achieved by solving the following non-linear programming problem:

$$\begin{aligned} &\text{Maximize } H(P, W) \\ &\text{subject to the following constraints;} \\ &\text{(i)} \quad y = f(x, ZP) + VW, \\ &\text{(ii)} \quad 1_K = \left(I_K \otimes 1_R'\right) P, \\ &\text{(iii)} \quad 1_n = \left(I_n \otimes 1_J'\right) W, \end{aligned} \tag{5}$$

where $H(P, W)$ could be any of the entropy measure functions. Note that $\otimes$ is the Kronecker product, 1_K is a K-dimensional vector of ones. The GME system in (5) is a non-linear programming system and can be solved by applying the lagrangian method, in which after finding the lagrangian function, one solves the first order conditions. For more details see Golan *et al.* [14].

4.1. GME procedure applied to the ultrastructural model. Without loss of generality, assume that the ξ_i are fixed; i.e. $\omega_i = 0, \ i = 1, 2, \ldots, n$. The compound model of the Ultrastructural model given in (1-2) can be rewritten as follows:

$$y_i = \alpha + \beta (x_i - \delta_i) + \varepsilon_i, \ i = 1, 2, \ldots, n. \tag{6}$$

Then by using the GME we can solve the problem after some reformulation of the unknown parameters α and β, and the unknown error terms δ_i, and ϵ_i, $i = 1, 2, 3, \ldots, n$. In the compound model (6), there are two unknown parameters which are reparametrized as

$$\boldsymbol{\alpha} = \mathbf{A}Q; \text{ where } 1_R' Q = 1,$$

where $\mathbf{A}$ is a row vector of size R and $\mathbf{Q}$ a R-dimensional column vector of weights. The slope will be in the form

$$\boldsymbol{\beta} = ZP; \text{ where } 1'_K P = 1,$$

where $\mathbf{Z}$ is a row vector of size K and $\mathbf{P}$ is a K-dimensional column vector of weights. Also, there are two error terms that are reparametrized in the following terms

$$\boldsymbol{\delta} = \mathbf{V}^* W^* \text{ where } (\mathbf{I}_n \otimes 1'_T) W^* = 1_n,$$
$$\boldsymbol{\varepsilon} = \mathbf{V}\mathbf{W}; \text{ where } (I_n \otimes 1'_J) W = 1_n.$$

Noting that, $\boldsymbol{V}^*$ and $\boldsymbol{V}$ are $(n \times nT)$ and $(n \times nJ)$ matrices, respectively; and $\mathbf{W}^*$ and $\mathbf{W}$ are a nT-dimensional and nJ-dimensional vectors of weights; respectively. Based on these reparametrizations, the new ultrastructural model can be rewritten as

$$Y = AQ + X\left(ZP - V^* W^*\right) + VW$$

4.2. GME solution: Shannon's entropy based solution. Generalized Maximum Entropy (GME) for the model given in (6) may be formulated as a nonlinear programming (NP) system. The NP selects q, p, w^* and $w \geq 0$ to maximize the joint Shannon entropy:

$$H(Q, P, W, W^*) = -Q' \ln(Q) - P' \ln(P) - W' \ln(W) - W^{*\prime} \ln(W^*)$$
Subject to
$$Y = AQ + X\left(ZP - V^* W^*\right) + VW,$$
$$1'_R Q = 1,$$
$$1'_K P = 1,$$
$$\left(I_n \otimes 1'_T\right) W^* = 1_n,$$
$$\left(I_n \otimes 1'_J\right) W = 1_n.$$

Here, we have $3n + 2$ constraints and $R + K + n(T + J)$ unknowns. The solution of this system can be found by deriving the first order conditions of the Lagrangian function:

$$L = H(Q, P, W, W^*) + \gamma'[Y - AQ - X(ZP - V^* W^*) - VW] + \theta'_1[1 - 1'_R Q] + \theta'_2[1 - 1'_K P] + \psi'[(I_n \otimes 1'_T)W^* + \zeta'[(I_n \otimes 1'_J)W,$$

where, $\gamma' \in \mathbb{R}^n$, $\theta_1 \in \mathbb{R}^R$, $\theta_2 \in \mathbb{R}^K$, $\psi \in \mathbb{R}^n$, and $\zeta \in \mathbb{R}^n$ are the associated vectors of Lagrangian multipliers. Taking the gradient of L to derive the first order condition, and solving these conditions, the GME solution selects the most uniform distribution consistent with the information provided in the data and the add up constraints:

$$\hat{Q} = \exp(-A\, 1'_n \hat{\gamma}) \bigodot \left\{1'_R \exp(-A\, 1'_n \hat{\gamma})\right\}^{-1},$$
$$\hat{P} = \exp(-Z\, 1'_n \hat{\gamma}(X - V^* \hat{W}^*)) \bigodot 1'_K \left\{\exp(-Z\, 1'_n \hat{\gamma}(X - V^* \hat{W}^*))\right\}^{-1},$$
$$\hat{W} = \exp(-V' \hat{\gamma}) \bigodot \left\{(I_n \otimes 1_J 1'_J) \exp(-V' \hat{\gamma})\right\}^{-1},$$
$$\hat{W}^* = \exp(-V^{*\prime} \hat{\gamma} 1'_K Z \hat{P}) \bigodot \left\{(I_n \otimes 1_T 1'_T) \exp(-V^{*\prime} \hat{\gamma} 1'_K Z \hat{P})\right\}^{-1}.$$

Here, $\bigodot$ is the Hadamard (element wise) product. To simplify matters; the individual probabilities take the forms:

$$\hat{q}_r = \frac{\exp\left(-a_r \sum_{i=1}^{n} \hat{\gamma}_i\right)}{\sum_{r=1}^{R} \exp\left(-a_r \sum_{i=1}^{n} \hat{\gamma}_i\right)},\ r = 1, 2, ..., R;$$

$$\hat{p}_k = \frac{\exp\left(-z_k \sum_{i=1}^{n} \hat{\gamma}_i \left(x_i - \sum_{t=1}^{T} v_t^* \hat{w}_{it}^*\right)\right)}{\sum_{k=1}^{K} \exp\left(-z_k \sum_{i=1}^{n} \hat{\gamma}_i \left(x_i - \sum_{t=1}^{T} v_t^* \hat{w}_{it}^*\right)\right)},\ k = 1, 2, 3, \ldots, K;$$

$$\hat{w}_{ij} = \frac{\exp(-\hat{\gamma}_i v_j)}{\sum_{j=1}^{J} \exp(-\hat{\gamma}_i v_j)},\ i = 1, 2, 3, \ldots, n,\ j = 1, 2, 3, \ldots, J;$$

$$\hat{w}_{it}^* = \frac{\exp\left(-\hat{\gamma}_i v_t^* \sum_{k=1}^{K} z_k \hat{p}_k\right)}{\sum_{t=1}^{T} \exp\left(-\hat{\gamma}_i v_t^* \sum_{k=1}^{K} z_k \hat{p}_k\right)},\ i = 1, 2, 3, \ldots, n,\ t = 1, 2, 3, \ldots, T.$$

Then the intercept and slope of the model (6) can be estimated as

$$\begin{aligned} \widehat{\alpha} &= A\widehat{Q}, \\ \widehat{\beta} &= Z\widehat{P}. \end{aligned} \tag{7}$$

4.1. Lemma. *The GME estimators given in* (7) *are unbiased estimators.*

Proof. Simply by taking the expected value of $\hat{\alpha}$, we have:

$$E(\widehat{\alpha}) = E\Big(\sum_r a_r \hat{p}_r\Big) = E(a) \sum_r \hat{p}_r.$$

Since $\sum_r \widehat{p_r} = 1$, then

$$E(a) = \sum_r a_r p_r = \alpha.$$

Therefore, $\widehat{\alpha}$ is an unbiased estimator of the intercept. Consequently, the estimated variance of $\widehat{\alpha}$ is:

$$\begin{aligned} \mathrm{Var}\,(\widehat{\alpha}) &= \mathrm{Var}\left(\sum_r a_r \overset{\wedge}{p_r}\right) = \sum_{r=1}^{R} \widehat{p_r^2} \mathrm{var}(a_r) \\ &= \sum \widehat{p_r^2} \left[\left(\sum a_r^2 \widehat{p_r}\right) - \left(\sum a_r \widehat{p_r}\right)^2\right] \\ &= \sum \widehat{p_r^2} \left[\left(\sum a_r^2 \frac{\exp\left(-\widehat{\gamma_r} a_r\right)}{\sum_{r=1}^{R} \exp\left(-\widehat{\gamma_r} a_r\right)}\right) - \left(\sum a_r \frac{\exp\left(-\widehat{\gamma_r} a_r\right)}{\sum_{r=1}^{R} \exp\left(-\widehat{\gamma_r} a_r\right)}\right)^2\right]. \end{aligned}$$

In a similar way we can prove that $\widehat{\beta}$ is an unbiased estimator of the slope; and consequently we can derive the associated variance. Also, similarly we can derive the estimators based on the other entropy measures. □

5. Monte Carlo experiments for the sensitivity analysis

Monte Carlo experiments were performed to comment on the choice of the support points and number of support points of the unknown parameters and error terms in the GME formulations. For fixed sample size, $n = 20$, a simulation study was carried out by generating 1000 samples according to the ultrastructural relationship $y_i = 1 + x_i + \varepsilon_i$ and $x_i = \frac{2+i}{2} + \delta_i,\ i = 1, 2, \ldots, n$. The error terms were generated independently from the standard normal, i.e., $\varepsilon \sim N(0, 1)$ and $\delta \sim N(0, 1)$. Then a comparison between the

GME and MLE estimation methods was made in terms of bias and Mean Squared Error (MSE):

$$\text{MSE}(\hat{\beta}) = \frac{1}{1000}\sum_{i=1}^{1000}\left(\hat{\beta}_i - \beta\right)^2,$$

and

$$\text{Bias}(\hat{\beta}) = \frac{1}{1000}\sum_{i=1}^{1000}\left(\hat{\beta}_i - \beta\right).$$

Three experiments were conducted:

Experiment 1. The support parameter space of the error terms, $\mathbf{V}^*$ and $\mathbf{V}$ were fixed to be three in the intervals $[-3S_x, 0, 3S_x]$ and $[-3S_y, 0, 3S_y]$ for δ and ε; respectively. Then, this experiment was conducted for selecting the support values (a, z) and the number (R, K) of support values for the unknown parameters $\alpha = AQ$ and $\beta = ZP$; i.e.,

$$\{a_i,\ i = 1, 2, \ldots, R\};\ \{z_j,\ j = 1, 2, \ldots, K\}.$$

In the first part of this experiment, we fixed the number of these support values to be 3 in the intervals [-1, 0, 1] to [-500, 0, 500]. The results of this simulation study shown in Table 4 indicate that the best support values for both parameters should be in the interval [-100, 0, 100]. Hereafter, in the second part of this experiment, we start to increase the number of support values within this interval so as to have 4, 5, 6 or 7 support points, and allocate them in an equidistant fashion. The results in Table 4 indicate that the greatest improvement in precision comes when R and K are equal to 7. Moreover, it could be noted that for all choices of the parameter spaces the GME results were more accurate and more efficient than the MLE results.

Table 4. Selecting the parameters supports

Method		$\hat{\alpha}$		$\hat{\beta}$	
		Bias	MSE	Bias	MSE
MLE		-0.1465	0.6190	-0.0903	0.0549
Shannon	$[-1, 0, 1]$	-0.0906	0.0824	-0.0546	0.0323
	$[-10, 0, 10]$	-0.0765	0.0599	-0.0445	0.0296
	$[-100, 0, 100]$	-0.0579	0.0386	-0.0471	0.0311
	[-500,0,500]	-0.0647	0.0465	-0.0450	0.0294
	Increasing number of support points				
	$[-100, -50, 50, 100]$	-0.0608	0.0405	-0.0464	0.0308
	$[-100, -50, 0, 50, 100]$	-0.0744	0.0582	-0.0443	0.0291
	$[-100, -50, -25, 25, 50, 100]$	-0.0588	0.0393	-0.0466	0.0307
	$[-100, -50, -25, 0, 25, 50, 100]$	-0.0544	0.0354	-0.0436	0.0278

Experiment 2. To check the impact of the error terms support space, and based on the experimental design outlines above, the Monte Carlo trial were repeated under the results of Experiment 1. The number of support values for each of the parameters $\mathbf{A}$ and $\mathbf{Z}$; were fixed to be 7 support values within the interval [-100, 0, 100]. Also, this experiment started by fixing the number of support values (J, T) to be 3, then the experiment was repeated by shifting the support values $\mathbf{V}^*$ and $\mathbf{V}$ in the interval $[hS, 0, hS]$, where $h = 1, 2, \ldots, 7$. Under the simulation assumptions, the results in Table 5 indicate that

the greatest improvement in the precision comes from using $h = 3$. In a similar way to Experiment 1, the simulation was repeated by increasing the number of support points in the interval [-3S, 0, 3S] for each error term. Taking into consideration both parameters, the greatest improvement comes when the number of support points is equal to 3.

Table 5. Selecting the error terms support

Method		$\hat{\alpha}$		$\hat{\beta}$	
		Bias	MSE	Bias	MSE
MLE		-0.1465	0.6190	-0.0903	0.0549
Shannon	$[-S, 0, S]$	-0.0540	0.0361	-0.0445	0.0282
	$[-2S, 0, 2S]$	-0.0534	0.0337	-0.0437	0.0279
	$[-3S, 0, 3S]$	-0.0544	0.0354	-0.0436	0.0278
	$[-4S, 0, 4S]$	-0.0557	0.0384	-0.0470	0.0311
	$[-5S, 0, 5S]$	-0.0714	0.0541	-0.0449	0.0296
	Increasing the number of support points				
	$[-3S, -1.5S, 1.5S, 3S]$	-0.0819	0.0672	-0.0397	0.0254
	$[-3S, -1.5S, 0, 1.5S, 3S]$	-0.0771	0.0633	-0.0405	0.0260
	$[-3S, -1.5S, -0.75S, 0.75S, 1.5S, 3S]$	-0.0603	0.0432	-0.0426	0.0266
	$[-3S, -1.5S, -0.75S, 0, 0.75S, 1.5S, 3S]$	-0.0644	0.0467	-0.0366	0.0219

Experiment 3. Based on the results of the previous experiments, which were related to the choice of parameter supports, this experiment starts to increase the sample size; $n = 20, 30, 40, 50$ and 100. For the higher entropy the order value $r = 0.5$ was used. The simulation results in Table 6 show that the GME estimators have a lower MSE and lower bias for all sample sizes for all entropy measures.

Table 6. Comparisons between GME and MLE

n	Method	Measure function	$\hat{\alpha}$		$\hat{\beta}$	
			Bias	MSE	Bias	MSE
20	GME	Shannon	-0.0544	0.0354	-0.0436	0.0278
		Tsallis	0.0557	0.0487	-0.065	0.0431
		Rényi	0.0546	0.0492	-0.066	0.0431
	MLE		-0.1465	0.6190	-0.0903	0.0549
30	GME	Shannon	-0.0488	0.0313	-0.0449	0.0284
		Tsallis	-0.0496	0.0364	-0.0488	0.0361
		Rényi	-0.0539	0.0369	-0.0497	0.0395
	MLE		-0.1303	0.6721	-0.0896	0.0501
40	GME	Shannon	-0.0396	0.0218	-0.0398	0.0240
		Tsallis	-0.0475	0.0298	-0.0495	0.0267
		Rényi	-0.0487	0.0302	-0.0518	0.0325
	MLE		-0.0769	0.4213	-0.0798	0.0481

Table 6. (Continued)

n	Method	Measure function	$\hat{\alpha}$		$\hat{\beta}$	
			Bias	MSE	Bias	MSE
50	GME	Shannon	-0.0411	0.0211	-0.0382	0.0239
		Tsallis	-0.0426	0.0304	-0.0471	0.0282
		Rényi	-0.0439	0.0309	-0.0473	0.0292
	MLE		-0.0818	0.4186	-0.0764	0.0397
100	GME	Shannon	-0.0375	0.0191	-0.0394	0.0234
		Tsallis	0.0391	0.0208	0.0484	0.0241
		Rényi	0.0505	0.0219	0.0484	0.0242
	MLE		-0.0734	0.3381	-0.0789	0.0268

6. Concluding remarks

The maximum entropy estimator introduced in this study applies the definition of Shannon's entropy; or its generalization; of discrete random variables. This study gives the researcher a more precise method for estimating the parameters of the ultrastructural model by applying the GME estimation approach based on Shannon entropy or any of its generalizations. The main idea of using GME is to improve the parameter estimation in the generalized measurement error models and to reduce the additional assumptions that are needed in the traditional MLE. In fact, all what the GME needs to be applicable can be obtained from the sample or can be specified by the researchers experiences. The Monte Carlo simulations provides good evidence for the superiority of the GME based Shannon entropy, as well as the other entropies; Tsallis and Rényi, on the MLE in terms of MSE. Hence, the GME can be considered a good alternative in estimating the ultrastructural models.

Due to lack of space, Tables 4 and 5 give only the results for Shannon entropy. The author will be happy to send interested readers tables giving the results for Tsallis and Rényi entropy by e-mail.

Acknowledgements

The author wish to express appreciation to the Editorial team of "Hacettepe Journal of Mathematics and Statistics" and to the referees for improving the manuscript.

References

[1] Al-Nasser, A. *Measuring Customer Satisfaction: An Information - Theoretic Approach* (Lambert Academic Publishing AG & Co. KG., Germany, 2010).

[2] Al-Nasser, A. *Entropy type estimator to simple linear measurement error models*, Austrian Journal of Statistics **34** (3), 283–294, 2005.

[3] Al-Nasser, A. *Estimation of multiple linear functional relationships*, Journal of Modern Applied Statistical Methods **3** (1), 181–186, 2004.

[4] Al-Nasser, A. *Customer satisfaction measurement models: Generalized maximum entropy approach*, Pakistan Journal of Statistics **19** (2), 213–226, 2003.

[5] Caputo, M and Paris, Q. *Comparative statics of the generalized maximum entropy estimator of the general linear model*, European Journal of Operational Research **185** (1), 195–203, 2008.

[6] Carroll, R. J., Ruppert, D. and Stefanski, L. A. *Measurement Error in Nonlinear Models* (Chapman and Hall, London, 1995).

[7] Cheng, C.-L. and Van Ness, J. W. *Statistical Regression with Measurement Error* (Arlond, New York, 1999).

[8] Ciavolino, E and Al-Nasser, A. *Comparing generalized maximum entropy and partial least squares methods for structural equation models*, Journal of Nonparametric Statistics **21** (8), 1017–1036, 2009.

[9] Ciavolino, E and Dahlgaard, J. *Simultaneous equation model based on the generalized maximum entropy for studying the effect of management factors on enterprise performance*, Journal of Applied Statistics **36** (7), 801–815, 2009

[10] Csiszar, I. *Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems*, The Annals of Statistics **19**, 2032–2066, 1991.

[11] Dolby, G. R. *The ultra-structural model: A synthesis of the functional and structural relations*, Biometrika **63**, 39–50, 1976.

[12] Donho, D. L, Johnstone, I. M, Hoch, J. C, and Stern, A. S. *Maximum entropy and nearly black object*, J. Royal, Statistical Society, Ser B **54**, 41–81, 1992.

[13] Golan, A. *Information and entropy econometrics - A review and synthesis*, Foundations and Trends in Econometrics **2** (1-2), 1-145, 2008.

[14] Golan, A., Judge, G. and Miller, D. *A Maximum Entropy Econometrics: Robust Estimation with limited data* (Wiley, New York, 1996).

[15] Golan, A., Judge, G. and Perloff, J. *Estimation and inference with censored and ordered multinomial response data*, J. Econometrics **79**, 23–51, 1997.

[16] Gleser, L. J. *A note on G. R. Dolby's unreplicated ultrastructural model*, Biometrika **72**, 117–124, 1985.

[17] Havrada, J. H. and Charvat, F. *Quantification methods of classification process: Concept of structural α-entropy*, Kybernetika **3**, 30–35, 1967.

[18] Jaynes, E. T. *Information and Statistical Mechanics I*, Physics Review **106**, 620–630, 1957.

[19] Jaynes, E. T. *Information and Statistical Mechanics II*, Physics Review **108**, 171–190, 1957.

[20] Jaynes, E. T. *Information Theory and Statistical Mechanics*, in Statistical Physics, K. Ford (ed.), (Benjamin, New York, 181, 1963).

[21] Kapur J. N. *Maximum Entropy Models in Science and Engineering* (John Wiley & Sons, New York, 1989)

[22] Paris, Q. *Multicollinearity and maximum entropy estimators*, Economics Bulletin **3** (11), 1–9, 2001.

[23] Peeters, L. *Estimating a random-coefficients sample-selection model using generalized maximum entropy*, Economics Letters **84**, 87–92, 2004.

[24] Pukelsheim, F. *The three sigma rule*, The American Statistician **48** (2), 88–91, 1994.

[25] Rényi, A. *On measures of information and entropy* (Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability, 1960), 547–561, 1961.

[26] Srivastava, A. and Shalabh, K. *Consistent estimation for the non-normal ultrastructural model*, Statist. Probab. Lett. **34**, 67–73, 1997.

[27] Shannon, C,E. *A mathematical theory of communication*, Bell System Technical Journal **27**, 379–423, 1948.

[28] Taneja, I. J. *Generalized Information Measures and Their Applications*, On-line book: http://www.mtm.ufsc.br/˜taneja/book/book.html, 2001.

[29] Tsallis, C. *Possible generalization of Boltzmann-Gibbs statistics*, J. Statistical Physics **52**, 479–487, 1988.