Generalized maximum entropy (GME) estimator: formulation and a monte carlo study

Eruygur, H. Ozan

26 May 2005

Paper presented at the
VII. National Symposium on Econometrics and Statistics
Istanbul, Turkey, May 26-27, 2005

GENERALIZED MAXIMUM ENTROPY (GME) ESTIMATOR: FORMULATION AND A MONTE CARLO STUDY

H. Ozan ERUYGUR

Middle East Technical University
Department of Economics
Ankara 06531 Turkey
eruygur@gmail.com

ÖZET

Entropinin kökeni 19. yüzyıla kadar uzanır. Belirsizlik ölçütü olarak geliştirilmesi ise Shannon (1948) tarafından olmuştur. Yaklaşık 10 yıl sonra 1957'de Jaynes, Shannon'un entropisini özellikle kötü tanımlanmış problemler için bir tahmin ve çıkarım methodu olarak Maksimum Entropi (ME) ilkesi adıyla formüle etmiştir. Yakın tarihte, Golan *et al.* (1996) Genelleştirilmiş Maksimum Entropi (GME) tahmincisini geliştirerek yeni bir tartışmayı başlatmıştır. Bu yazı iki kısımdan oluşmaktadır. İlk kısım, bu yeni tekniğin (GME) formülasyonu üzerinedir. İkinci kısımda ise bir Monte Carlo çalışmasıyla normal dağılmayan hata terimleri durumunda GME'nin tahmin sonuçları tartışılacaktır.

Anahtar Kelimeler: Entropi, Maksimum Entropi, ME, Genelleştirilmiş Maksimum Entropi, GME, Monte Carlo, Shannon Entropisi, Normal dağılmayan hata terimi.

ABSTRACT

The origin of entropy dates back to 19th century. In 1948, the entropy concept as a measure of uncertainty was developed by Shannon. A decade after in 1957, Jaynes formulated Shannon's entropy as a method for estimation and inference particularly for ill-posed problems by proposing the so called Maximum Entropy (ME) principle. More recently, Golan *et al.* (1996) developed the Generalized Maximum Entropy (GME) estimator and started a new discussion in econometrics. This paper is divided into two parts. The first part considers the formulation of this new technique (GME). Second, by Monte Carlo simulations the estimation results of GME will be discussed in the context of non-normal disturbances.

Keywords: Entropy, Maximum Entropy, ME, Generalized Maximum Entropy, GME, Monte Carlo Experiment, Shannon's Entropy, Non-normal disturbances.

I. INTRODUCTION

The origin of entropy dates back to 19th century. In 1948, the entropy concept as a measure of uncertainty (state of knowledge) was developed by Shannon in the context of communication theory. A decade after in 1957, *E. T. Jaynes* formulated Shannon's entropy as a method for estimation and inference particularly for *ill-posed* problems by proposing the so called *Maximum Entropy (ME) principle*. More recently, Golan *et al.* (1996) developed the Generalized Maximum Entropy (GME) estimator and started a new discussion in econometrics.

Our paper is divided into two parts. The first part considers the formulation of this new technique of Generalized Maximum Entropy (GME). In the second part, by performing Monte Carlo simulations, we will discuss the estimation results of GME in the case of non-normal distributed disturbances.

II. THE GENERALIZED MAXIMUM ENTROPY APPROACH

Suppose that we observe a *T*-dimensional vector **y** of noisy indirect observations on an unknown and unobservable *K*-dimensional parameter vector **β**, where **y** and **β** are related through the following linear model relationship

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{u} \qquad \text{(EQ.1)}$$

where **X** is the *TxK* known matrix of explanatory variables and **u** is a *Tx1* noise (disturbance) vector.

In order to be able to use Jaynes' Maximum Entropy principle for the estimation of regression parameters, the parameters **β** must be written in terms of probabilities because of the fact that the *arguments* of the Shannon's maximum entropy function[I] are *probabilities.* Following Golan *et al* (1996), if we define M ≥ 2 equally distanced discrete *support values*,

z_{km}, as the possible realizations of β_k with corresponding probabilities p_{km}, we can convert each parameter β_k as follows:

$$\beta_k = \sum_{m=1}^{M} z_{km} p_{km} \text{ , for } k=1, 2, 3\ldots,K, \quad \text{where } M \geq 2 \qquad \text{(EQ.2)}$$

Let us define the M dimensional vector of equally distanced discrete points (support space) as $\mathbf{z}'_{\mathbf{k}}$=[z_{k1}, z_{k2},…, z_{kM}] and associated M dimensional vector of probabilities as $\mathbf{p_k}$=[p_{k1}, p_{k2},…, p_{kM}]′. Now, we can rewrite β in (EQ.1) as

$$\beta = \mathbf{Zp} = \begin{bmatrix} \mathbf{z}'_1 & 0 & . & 0 \\ 0 & \mathbf{z}'_2 & . & . \\ . & . & . & 0 \\ 0 & . & 0 & \mathbf{z}'_{\mathbf{K}} \end{bmatrix} \begin{bmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ . \\ \mathbf{p}_{\mathbf{K}} \end{bmatrix} \qquad \text{(EQ.3)}$$

where **Z** is a *block diagonal KxKM* matrix of support points with

$$\mathbf{z}'_{\mathbf{k}}\mathbf{p}_{\mathbf{k}} = \sum_{m=1}^{M} z_{km} p_{km} = \beta_k \text{ for } k=1,2,\ldots, K, \qquad m=1,2,\ldots,M \qquad \text{(EQ.4)}$$

where $\mathbf{p_k}$ is a M dimensional *proper probability vector*^{II} corresponding to a M dimensional vector of weights $\mathbf{z_k}$. Recall that the last vector, $\mathbf{z_k}$, defines the *support space* of β_k**.** By this way, each parameter is converted from the real line into a well-behaved set of proper probabilities defined over the supports.

As can be seen, the implementation of the maximum entropy formalism allowing for *unconstrained parameters* starts by *choosing* a set of discrete points by researcher based on his *a priori* information about the value of parameters to be estimated, where these set of discrete points are called the *support space* for all parameters. In most cases, where researchers are uninformed as to the sign and magnitude of the unknown β_k, they should specify a support space that is uniformly symmetric around zero with end points of large magnitude, say $\mathbf{z}_\mathbf{k}'$=[*-C, -C/2, 0, C/2, C*] for *M*=5 and for some scalar *C* [Golan *et al.*, 1996:77].

Similarly, we can also transform the *noises* **u** as follows [Golan *et al.*, 1996:87]:

$$u_t = \sum_{j=1}^{J} v_{tj} w_{tj} \text{, for } t\text{=1, 2, …, } T\text{,} \qquad \text{where } \; J \geq 2 \qquad \text{(EQ.5)}$$

Notice that by this conversation, Golan *et al.* (1996:121) propose a transformation of the possible outcomes for $\mathbf{u_t}$ to the interval [0,1] by defining a set of discrete support points $\mathbf{v}_\mathbf{t}'$=[v_{t1}, v_{t2},…, v_{tJ}] which is distributed uniformly and evenly around zero (such that v_{t1}=-v_{tJ} for each *t* if we assume that the error distribution is symmetric and centered about **0**)[III] and a vector of corresponding unknown probabilities $\mathbf{w_t}$=[w_{t1}, w_{t2},…, w_{tJ}]′ where *J*≥2. Now, we can rewrite **u** in (EQ.1) as

$$\mathbf{u} = \mathbf{Vw} = \begin{bmatrix} \mathbf{v}_1' & 0 & . & 0 \\ 0 & \mathbf{v}_2' & . & . \\ . & . & . & 0 \\ 0 & . & 0 & \mathbf{v}_\mathbf{T}' \end{bmatrix} \begin{bmatrix} \mathbf{w_1} \\ \mathbf{w_2} \\ . \\ \mathbf{w_T} \end{bmatrix} \qquad \text{(EQ.6)}$$

with

$$\mathbf{v}_\mathbf{t}'\mathbf{w_t} = \sum_{j=1}^{J} v_{tj} w_{tj} = u_t \qquad \text{for } t\text{=1,2,…, } T \text{ and } j\text{=1,2,…,}J \qquad \text{(EQ.7)}$$

In (EQ.4) and (EQ.7) the support spaces $\mathbf{z_k}$ and $\mathbf{v_t}$ are chosen to span the relevant parameter spaces for each {β_k} and {u_t}, respectively. As for the determination of support bounds for disturbances, Golan *et al* (1996) recommend using the "three-sigma rule" of Pukelsheim (1994) to establish bounds on the error components: the lower bound is v_L = -3σ_y and the upper bound is v_U = 3σ_y, where σ_y is the (empirical) standard deviation of the sample **y**. For example if J= 5, then $\mathbf{v}_\mathbf{t}'$=(-3σ_y, -1.5σ_y, 0, 1.5σ_y, 3σ_y) can be used.

Under this reparameterization, the inverse problem with noise given in (EQ.1) may be rewritten as

$$\mathbf{y}{=}\mathbf{X}\boldsymbol{\beta}{+}\mathbf{u} = \mathbf{XZp} + \mathbf{Vw} \qquad \text{(EQ.8)}$$

Jaynes (1957) demonstrates that *entropy is additive for independent sources of uncertainty*. The details of this property can be found in Kapur and Kesavan (1992:31-32). Therefore, assuming the unknown weights on the parameter and the noise supports for the linear regression model are independent, we can jointly recover the unknown parameters and disturbances (noises or errors) by solving the constrained optimization problem of *max* $H(\mathbf{p,w}){=}{-}\mathbf{p}'\ln\mathbf{p}{-}\mathbf{w}'\ln\mathbf{w}$ subject to $\mathbf{y}{=}\mathbf{XZp}{+}\mathbf{Vw}$.

Hence, given the reparameterization in (EQ.8) where $\{\beta_k\}$ and $\{u_t\}$ are transformed to have the properties of probabilities, in scalar notation the GME formulation for a noisy inverse problem may be stated as

$$\max_{\mathbf{p,w}} H(\mathbf{p},\mathbf{w}) = -\sum_{k=1}^{K}\sum_{m=1}^{M} p_{km}.\ln p_{km} - \sum_{t=1}^{T}\sum_{j=1}^{J} w_{tj}.\ln w_{tj} \qquad \text{(EQ.9)}$$

subject to the constraints

$$\sum_{k=1}^{K}\sum_{m=1}^{M} x_{tk} z_{km} p_{km} + \sum_{j=1}^{J} w_{tj} v_{tj} = y_t\,, \qquad \text{for } t{=}1, 2,\ldots,T. \qquad \text{(EQ.10)}^{\text{IV}}$$

$$\sum_{m=1}^{M} p_{km} = 1\,, \qquad \text{for } k{=}1, 2,\ldots,K. \qquad \text{(EQ.11)}$$

$$\sum_{j=1}^{J} w_{tj} = 1\,, \qquad \text{for } t{=}1, 2,\ldots,T. \qquad \text{(EQ.12)}$$

where (EQ.10) is the data (or, consistency) constraint whereas (EQ.11) and (EQ.12) provide the required adding-up constraints for probability distributions of $\{p_{km}\}$ and $\{w_{tj}\}$, respectively.

The solution for $\hat{p}_{km}$ is

$$\hat{p}_{km}^{GME} = \frac{e^{-\sum_{t=1}^{T}\hat{\lambda}_t z_{km} x_{tk}}}{\sum_{m=1}^{M} e^{-\sum_{t=1}^{T}\hat{\lambda}_t z_{km} x_{tk}}} \qquad \text{where } \Omega_k^p(\hat{\lambda}_t) = \sum_{m=1}^{M} e^{-\sum_{t=1}^{T}\hat{\lambda}_t z_{km} x_{tk}} \qquad \text{(EQ.13)}$$

The solution for $\hat{w}_{tj}$

$$\hat{w}_{tj}^{GME} = \frac{e^{-\hat{\lambda}_t v_{tj}}}{\sum_{j=1}^{J} e^{-\hat{\lambda}_t v_{tj}}} \qquad \text{where } \Omega_t^w(\hat{\lambda}_t) = \sum_{j=1}^{J} e^{-v_{tj}\hat{\lambda}_t} \qquad \text{(EQ.14)}$$

Notice that, in the expressions above, $\hat{\lambda}_t$ represent the dual value of data constraint. Substituting the solutions of $\hat{p}_{km}$ and $\hat{w}_{tj}$ into (EQ.2) and (EQ.5) produces the GME estimates of β_k and u_t, as

$$\hat{\beta}_k^{GME} = \sum_{m=1}^{M} \hat{p}_{km} z_{km} , \qquad \text{for k=1,2,…,K} \qquad \text{(EQ.15)}$$

and

$$\hat{u}_t^{GME} = \sum_{j=1}^{J} \hat{w}_{tj} v_{tj} , \qquad \text{for t=1,2,…,T} \qquad \text{(EQ.16)}$$

As can be seen, the GME estimates depend on the optimal Lagrange multipliers $\hat{\lambda}_t$ for the model constraints. There is no closed-from solution for $\hat{\lambda}_t$, and hence *no closed form solution* for **p**, **w**, **β** and **u**. Therefore numerical optimization techniques should be used to obtain the solutions and solutions must be found numerically.

III. A MONTE CARLO STUDY WITH NON NORMAL DISTURBANCES

In this section a Monte Carlo study is carried out to test the precision performance of GME estimators in the case of non normal distributed disturbances. For the sake of comparison, the simulation consists of three groups of data generation processes. First we generated a data set with normal disturbances. For this purpose, a normal distribution with zero mean and unit variance is used. Second, in order to represent a highly skewed error distribution, a chi-square distribution with unit degrees of freedom [$\chi^2(1)$] is used. Third, another chi square distribution but this time with 4 degrees of freedom [$\chi^2(4)$] is generated in order to represent a distribution which is less skewed.

For every generated data set, the model parameters are estimated with both GME and OLS and the whole procedure is repeated 1 000 times for each sample size. The sample sizes used are T=5, 10, 15, 20, 30, 40 and 50. The model estimated is a simple regression model with a constant term. The true value of the constant term is set to 0.5, while the true value of the slope parameter is set to 1.75. The econometric software Shazam 10.0 is used for the estimation of Monte Carlo trials.

The performances of the OLS and GME estimators are evaluated using the measures of absolute bias (ABIAS) and root mean square error (RMSE). Absolute bias is calculated as the absolute value of the difference between average estimate and the true value of the parameter. Root mean square error is, as known, the square root of mean square error (MSE) which is given by the sum of the squared bias and the empirical variance. The sum of all RMSE is denoted by SRMSE, and the sum of all ABIAS is denoted by SABIAS.

For each type of error distribution [N(0,1); $\chi^2(1)$ and $\chi^2(4)$] the simulation results are presented by four figures. First figure represents the SRMSE and SABIAS of OLS estimator. Second figure shows the SRMSE and SABIAS of GME estimator. Third figure plots the SRMSE of both OLS and GME in one graph, whereas the fourth one plots the SABIAS of both OLS and GME in one graph.

First, we will investigate the simulation results for the case of normally distributed disturbances. If we examine Figures 1A and 1B, we can see that sum of absolute bias (SABIAS) of GME decreases to near zero around sample size of fifteen (T=15), whereas that happens around sample size of ten (T=10) for the OLS estimator. Hence, we see that OLS estimator is much faster in the case of normal disturbances in terms of decrease in absolute bias. However, one should note that for a very small sample size such as five (T=5), the sum of absolute bias (SABIAS) of OLS is about three times and the root mean square error (SRMSE) is about one and a half times higher than that of GME. This is a remarkable result in favor of GME based on the criteria of absolute bias in the case of very low sample sizes such as five (T=5). Another important point that Figures 1A and 1B reveal is the fact that the behaviors of OLS and GME estimators in terms of both sum of root mean square error (SRMSE) and absolute bias (SABIAS) follow very similar patterns, starting from a sample size of fifteen (T=15). A further important point that we can note from Figures 1A and 1B is that the SRMSE of both OLS and GME estimators' decreases with increasing sample size, which is a behavior that indicates the *consistency* of estimators. Note that the consistency properties of data constrained GME estimator is first established by Mittelhammer and Cardell (1997) and later developed more by Mittelhammer, Cardell and Marsh (2002). Some more details can be found in Mittelhammer R., G. Judge, and D. Miller (2000).

Recalling that the MSE is the sum of the squared bias and the empirical variance, Figure 1B shows that, while the bias represents only a small part of the RMSE, the standard errors of GME estimates constitutes the important part. For small sample sizes this could lead to poor estimates, however, this situation can be handled by including out of sample information that can be introduced into the GME estimator. In other words, the employment of some prior information would decrease the variance of estimators at small sample sizes without introducing a strong additional bias. An easy way to include such out of sample information is to use the priors on parameters or disturbance support spaces. Hence, in addition to the good performance of GME in terms of absolute bias and root mean square in small sample sizes, the root mean square error (SRMSE) can also be much more reduced by incorporating out of sample information using support spaces of parameters and disturbances. However, in this study we will not go into further details of this issue and leave it as an important topic for another study. Notice that, in this study we do not include any prior information using support spaces: support spaces are defined as zero centered symmetrically large intervals.

In Figure 1C, the graph of sum of root mean square errors (SRMSE) of both OLS and GME is presented. It is clear that, until sample size of about thirty (T=30), the SRMSE of GME is always lower than that of OLS. Particularly for low sample sizes (lower than T=15), the SRMSE performance of GME is remarkable. With the beginning of relatively large sample sizes (T>30), the SRMSE of OLS appear to be lower than that of GME but at very small amounts.

The large differences in absolute bias (SABIAS) of OLS and GME in small sample sizes such as five (T=5) is notable, which is the case depicted in Figure 1D. If we compare Figure 1D with figures 2D and 3D we can see that in the case of normal disturbances (Figure 1D) the absolute bias (SABIAS) of OLS is lower than that of GME, starting from a sample size of about ten (T=10). However, the same performance of OLS cannot be seen in the case of non normal disturbances (Figure 2D and Figure 3D). In these cases, the absolute bias (SABIAS) of GME is always lower than that of OLS, although this difference becomes very low after a sample size thirty (T=30). Another remarkable result form our simulation is that, with higher degrees of freedom (df=4), the absolute bias of GME is much more lower when compared to unit degrees of freedom (df=1) as depicted in Figure 2D. In addition, this behavior of the absolute bias for the GME is long lasting until a sample size of fifteen (T=15) with higher degrees of freedom when compared to unit degrees of freedom (df=1).

Consulting to Figure 2A and 3A, one can observe that SRMSE and SABIAS of OLS becomes very close to each other particularly after the sample size of thirty (T=30), whereas the same pattern is seen for GME after a sample size of only twenty (T=20). Finally, in figures 2C and 3C, SRMSE performance of GME is compared with that of OLS. When these figures are examined, the first important point to note is the highly low SRMSE of GME estimator in small sample sizes (T<15) when compared with that of OLS. As stated before, contrary to normally distributed disturbances situation, in the case of non normal disturbances the SRMSE and SABIAS of GME estimator is always lower even in large sample sizes, although the gap is very low and closes faster for low sample sizes starting from fifteen (T=15).

IV. CONCLUSION

In first part of this article we have presented the general formulation of the newly developed GME estimator of Golan *et al*. (1996). In the second part of the paper, a Monte Carlo simulation study is performed to evaluate the precision performance of the GME compared with the performance of the OLS estimator. From the results that we have outlined briefly in the previous section, we can conclude that the performance of the GME estimator is remarkably good when compared to that of the OLS estimator, especially for small sample sizes. In addition, in case of non normal disturbances, this performance becomes prominently better. Of course, the findings in this paper are not analytical findings and the results are based on the model framework of the Monte Carlo study. However, the findings are notably promising at least within the structure of our model of simulation. The findings in this paper give some clues for the use of GME estimators. First, in the case of small samples because of data unavailability, the GME estimator can give better results when compared with OLS. Second, even in the case of non normal distributed disturbances, performance of the GME estimator is good and in addition, our Monte Carlo studies show that its performance is better than normal disturbances case.

Apart from the small sample sizes, there is another important feature of GME estimator that we have not dealt with in this study: we can use the GME approach even in the case of negative degrees of freedom, in other words, in the case of *ill-posed* problems. With all these advantages discussed, we think that this newly developed estimator is a good method particularly for the case of insufficient data and in the framework of non normal disturbances.

FIGURES

Figure 1A.

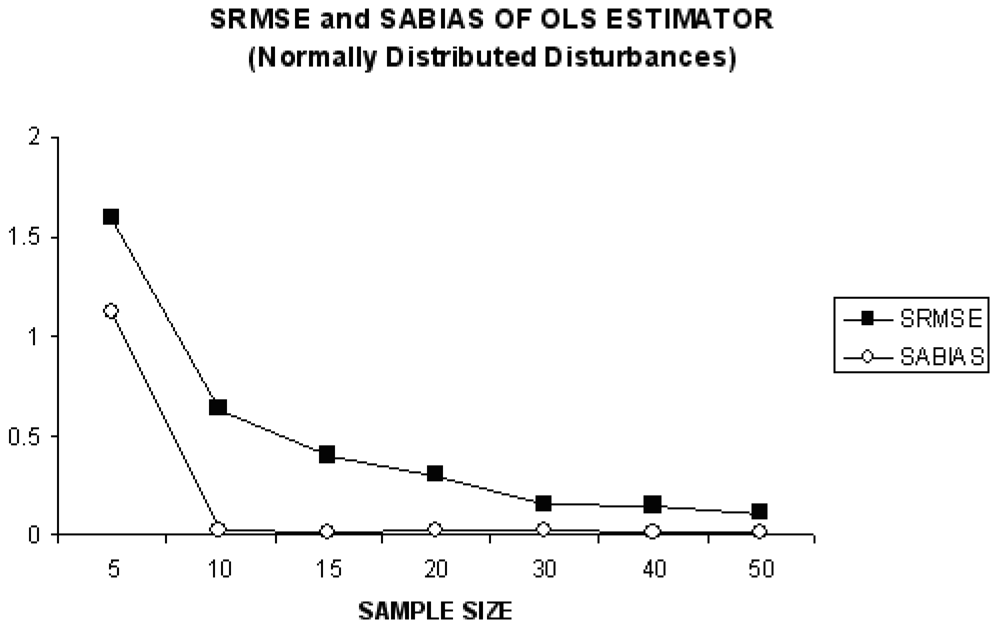


Figure 1B

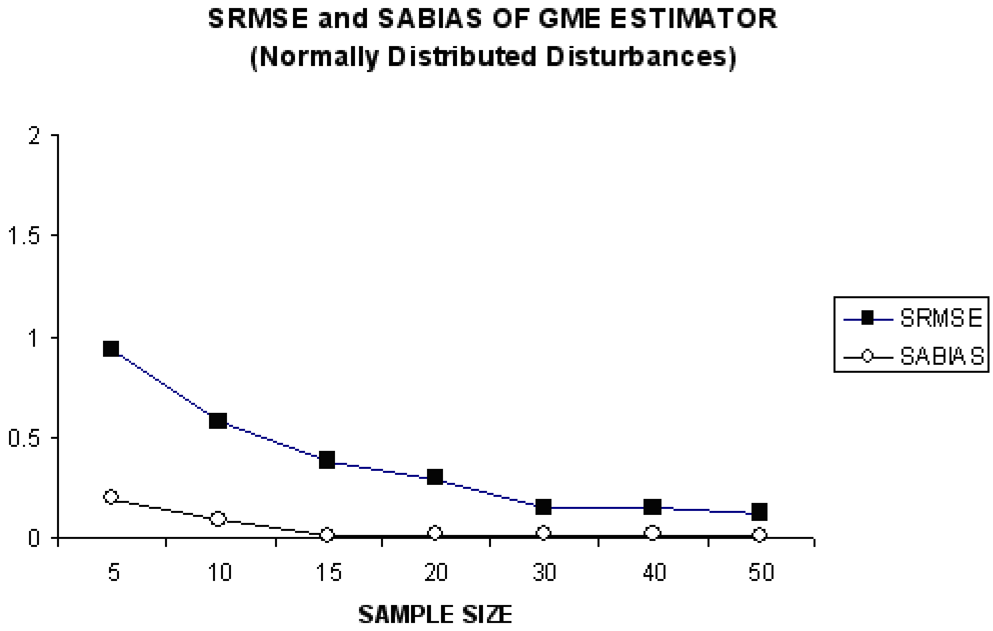


Figure 1C.

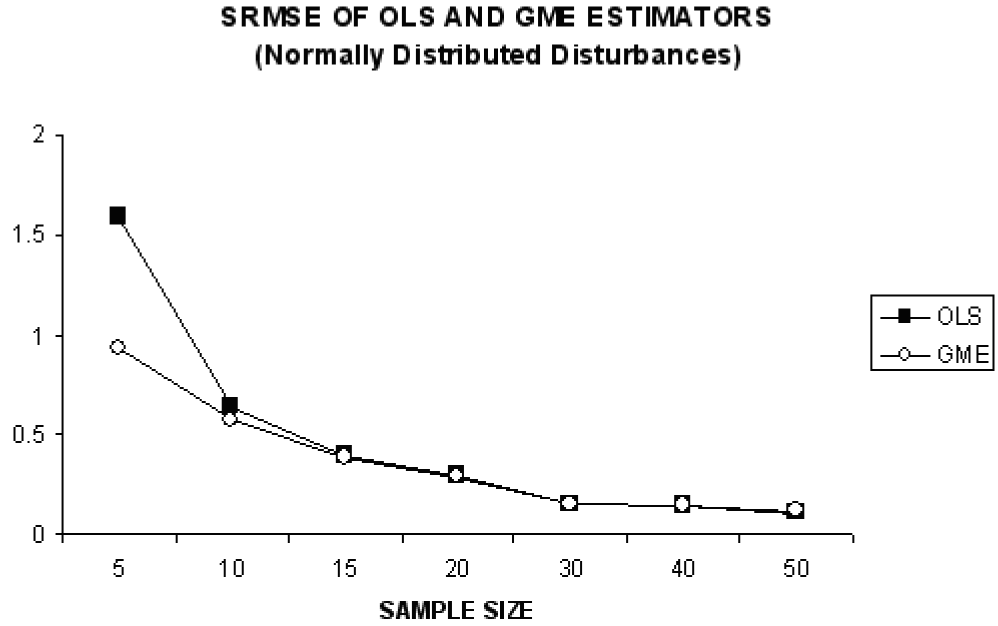


Figure 1D

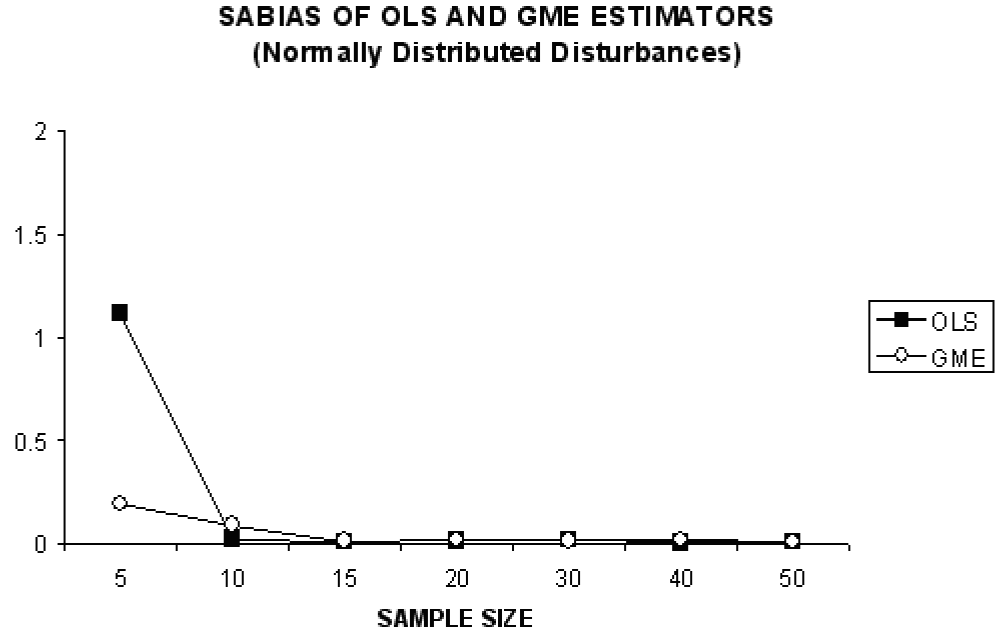


Figure 2A.

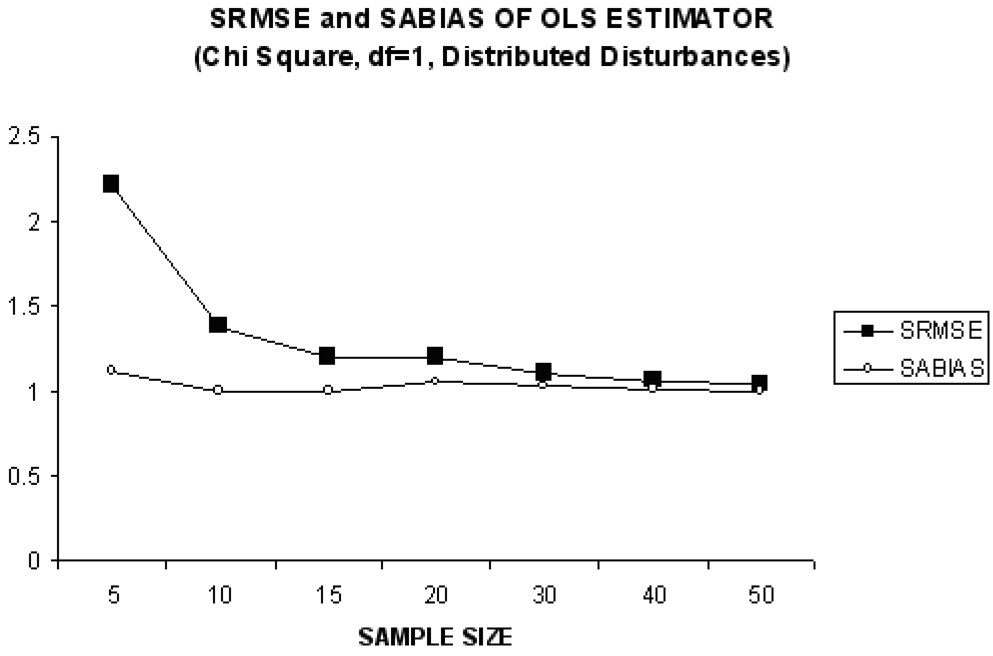


Figure 2B

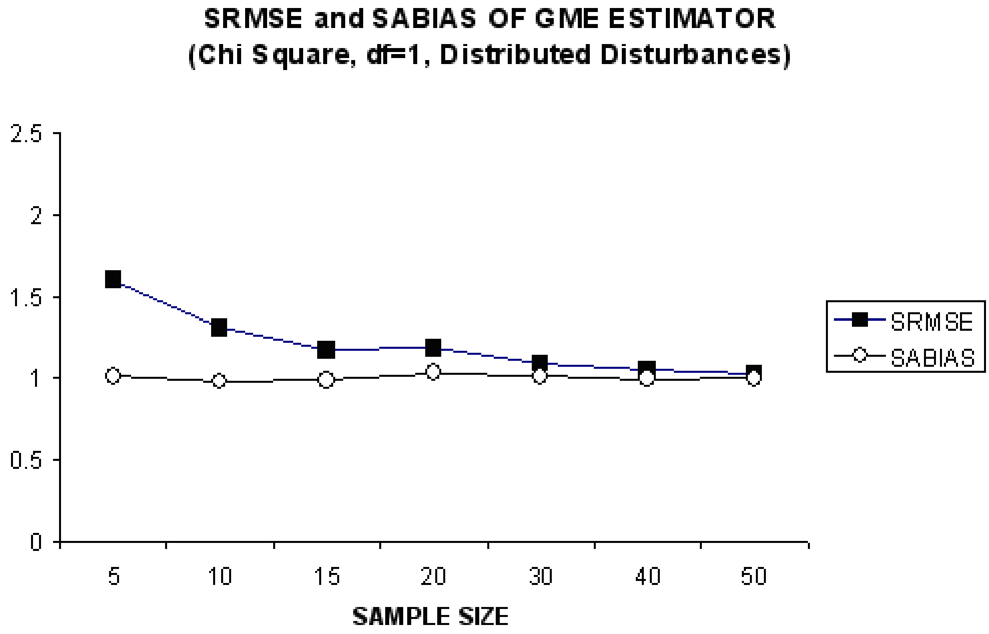


Figure 2C.

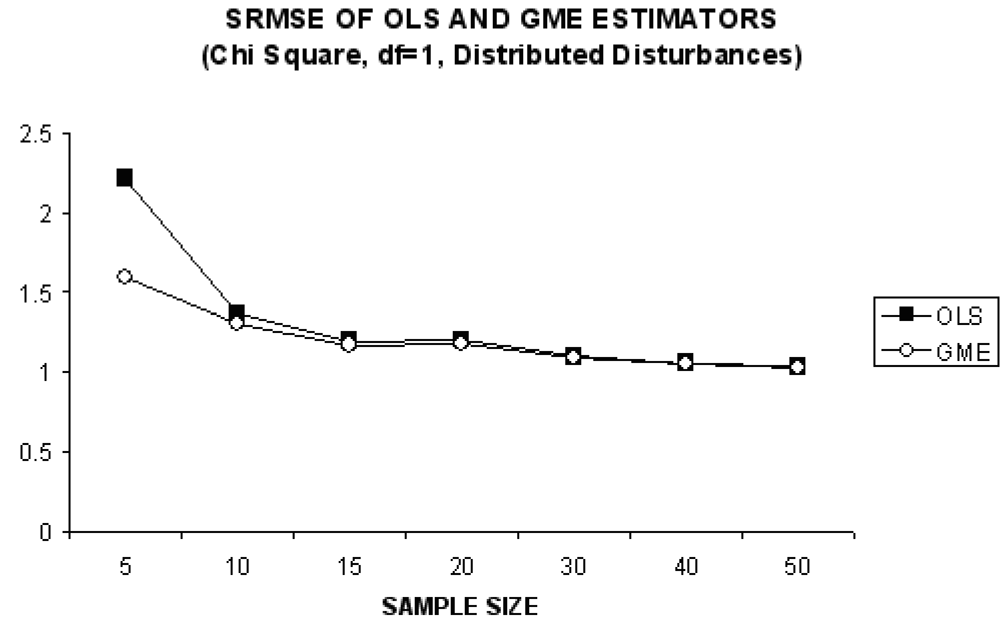


Figure 2D

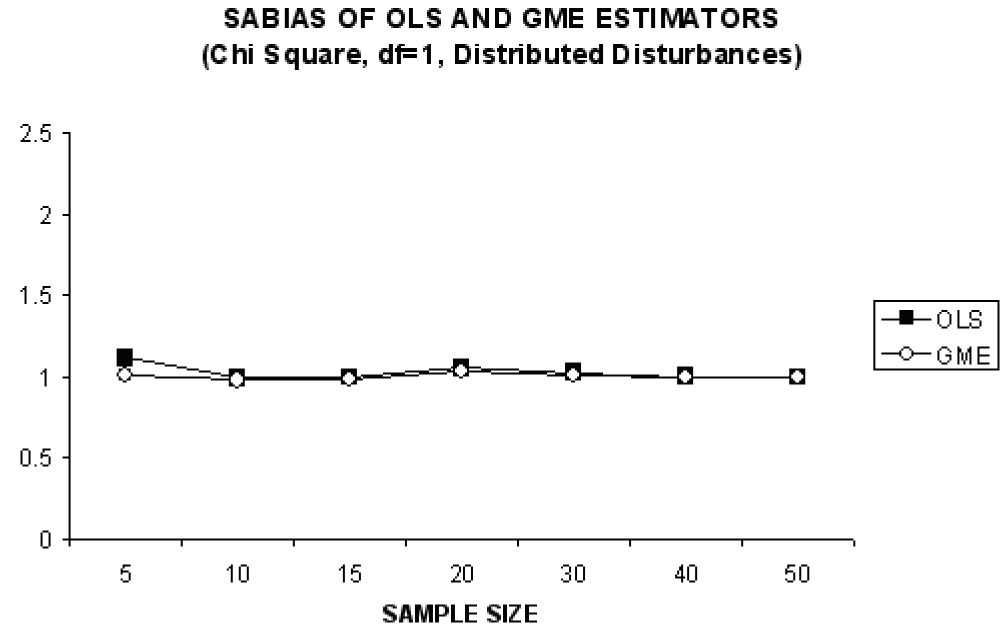


Figure 3A.

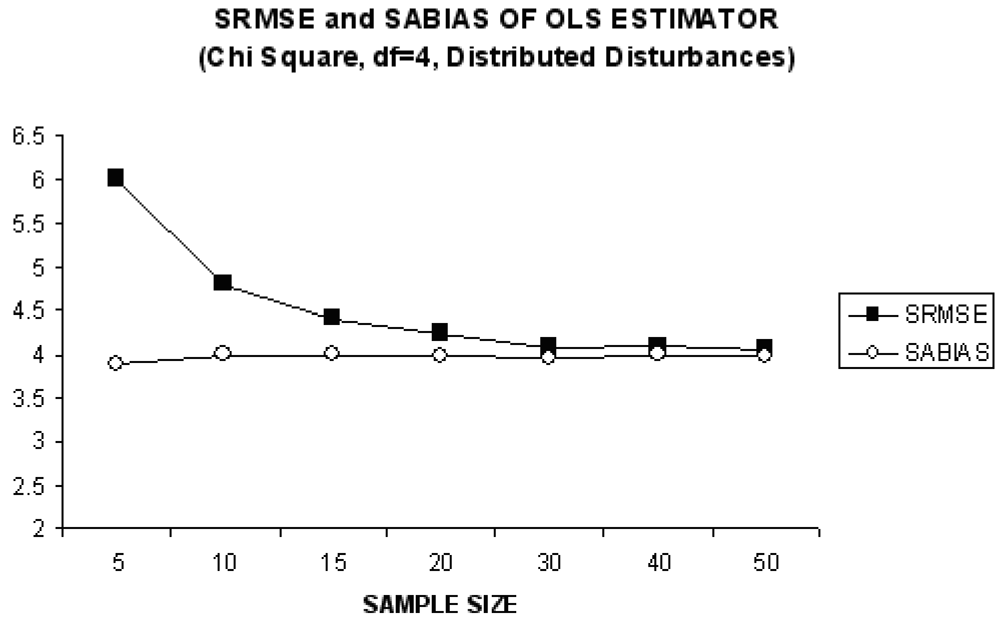


Figure 3B

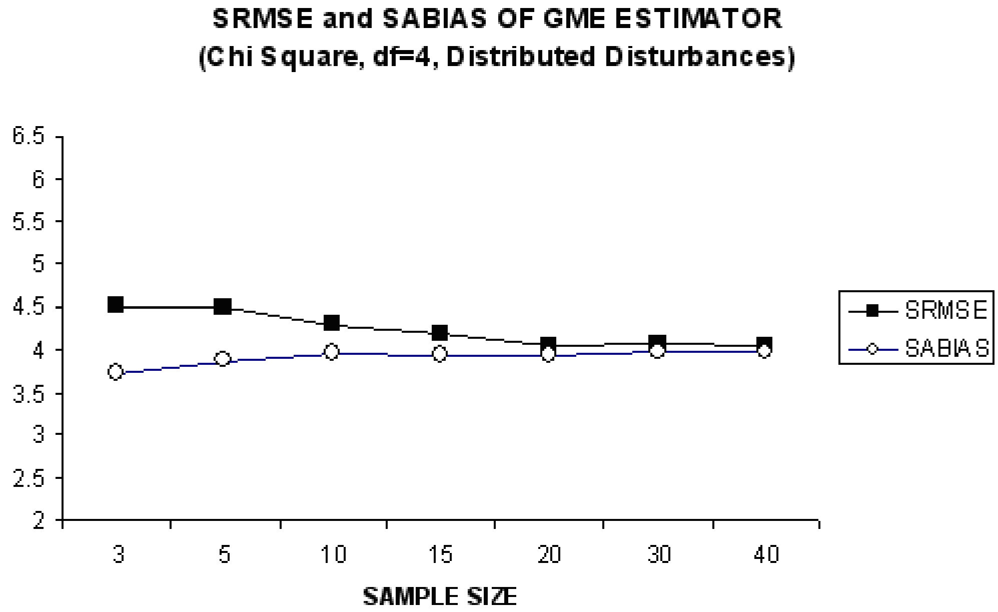


Figure 3C.

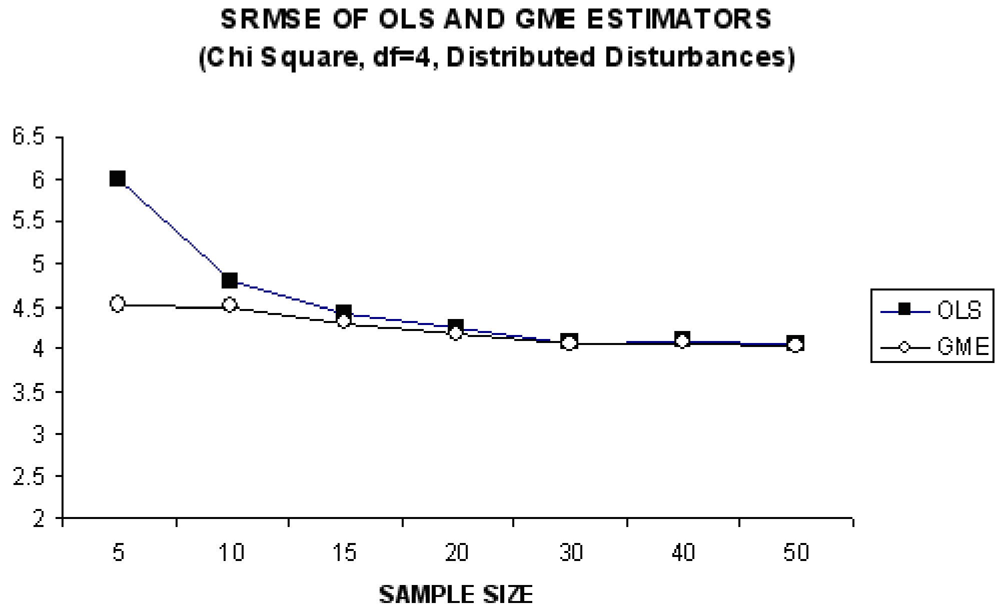
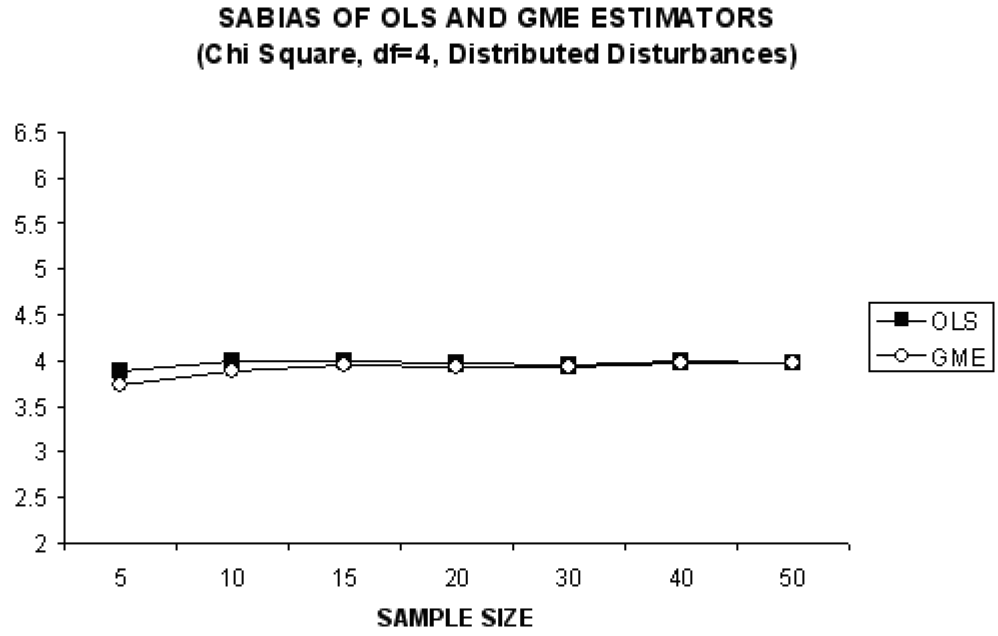


Figure 3D



ENDNOTES

[I] For properties of Shannon' Entropy measure, see Kapur and Kesavan (1992:24).

[II] A *proper probability vector* is characterized by *two* properties: $p_{km}\geq 0$, $\forall$ m=1,...,M and $\sum_{m=1}^{M} p_{km} = 1$

[III] Note that $J\geq 2$ points may be used to express or recover additional information about u_t (e.g. skewness or kurtosis). For example if we assume that the noise distribution is skewed such that $u_t \sim \chi^2(4)$, then v=[$-\sqrt{2}$, $2\sqrt{2}$] can be used as support space for noise representing the skewness.

[IV] One can easily show that $\sum_{k=1}^{K} x_{tk}\beta_k = \sum_{k=1}^{K} x_{tk} \sum_{m=1}^{M} z_{km} p_{km} = \sum_{k=1}^{K}\sum_{m=1}^{M} x_{tk} z_{km} p_{km}$

REFERENCES

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

Golan, A., Judge, G. and Miller D. (1996), *Maximum Entropy Econometrics: Robust Estimation With Limited Data*, John Wiley & Sons.

Kapur, J.N., & Kesavan, H.K. (1992), *Entropy Optimization Principles with Applications,* Academic Press, London.

Mittelhammer, R. and S. Cardell (1997), "On the Consistency and Asymptotic Normality of the Data Constrained GME Estimator of the GML", *Working Paper*, Washington State University, Pullman, WA.

Mittelhammer R., G. Judge, and D. Miller (2000), *Econometric Foundations,* Cambridge University Press.

Mittelhammer, R., S. Cardell and L. Marsh T. (2002), "The Data Constrained GME Estimator of the GML: Asymptotic Theory and Inference", *Working Paper*, Washington State University, Pullman, WA.

Pukelsheim. F. (1994). "The Three Sigma Rule", *American Statistician,* 48, 88-91.