

Article

## Generalized Maximum Entropy Analysis of the Linear Simultaneous Equations Model

Thomas L. Marsh<sup>1,\*</sup>, Ron Mittelhammer<sup>2</sup> and Nicholas Scott Cardell<sup>3</sup>

<sup>1</sup> School of Economic Sciences and IMPACT, Washington State University, Pullman, WA 99164, USA

<sup>2</sup> School of Economic Sciences and Statistics, Washington State University, Pullman, WA 99164, USA; E-Mail: mittelha@wsu.edu

<sup>3</sup> Salford Systems, San Diego, CA 92126, USA; E-Mail: scardell@gocougs.wsu.edu

\* Author to whom correspondence should be addressed; E-Mail: tl\_marshall@wsu; Tel.: +1-509-335-8597; Fax: +1-509-335-1173.

Received: 20 November 2013; in revised form: 17 January 2014 / Accepted: 28 January 2014 / Published: 12 February 2014

---

**Abstract:** A generalized maximum entropy estimator is developed for the linear simultaneous equations model. Monte Carlo sampling experiments are used to evaluate the estimator's performance in small and medium sized samples, suggesting contexts in which the current generalized maximum entropy estimator is superior in mean square error to two and three stage least squares. Analytical results are provided relating to asymptotic properties of the estimator and associated hypothesis testing statistics. Monte Carlo experiments are also used to provide evidence on the power and size of test statistics. An empirical application is included to demonstrate the practical implementation of the estimator.

**Keywords:** robustness; asymptotic normality; consistency; hypothesis testing; Monte Carlo experiments; Wald test

**MSC Codes:** 62

---

### 1. Introduction

The simultaneous equations model (SEM) is applied extensively in econometric-statistical studies. Examples of traditional estimators for the SEM include two stage least squares [1], three stage

least squares [2], limited information maximum likelihood [3], and full information maximum likelihood [4,5]. These estimators yield consistent estimates of structural parameters by correcting for simultaneity between the endogenous variables and the disturbance terms of the statistical model. However, in the presence of small samples or ill-posed problems, traditional approaches may provide parameter estimates with high variance and/or bias, or provide no solution at all. As an alternative to traditional estimators, we present a generalized maximum entropy estimator for the linear SEM and rigorously analyze its sampling properties in small and large sample situations including the case of contaminated error models.

Finite sampling properties of the SEM have been discussed in [6–10], where alternative estimation techniques that have potentially superior sampling properties are suggested. Specifically, they discussed limitations of asymptotically justified estimators in finite sample situations and the lack of research on estimators that have small sample justification. In a special issue of *The Journal of Business and Economic Statistics*, the authors of [11,12] examined small sample properties of generalized methods of moments estimators for model parameters and covariance matrices. References [13–15] pointed out that even small deviations from model assumptions in parametric econometric-statistical models that are only asymptotically justified can lead to undesirable outcomes. Moreover, Reference [16] singled out the extreme sensitivity of least squares estimators to modest departures from strictly Gaussian conditions as a justification for examining robust methods of estimation. These studies motivate the importance of investigating alternatives to parameter estimation methods for the SEM that are robust in finite samples and lead to improved prediction, forecasting, and policy analysis.

The principle of maximum entropy has been applied in a variety of modeling contexts. Reference [10] proposed estimation of the SEM based on generalized maximum entropy (GME) to deal with small samples or ill-posed problems, and defined a criteria that balances the entropy in both the parameter and residual spaces. The estimator was justified on information theoretic grounds, but the repeated sampling properties of the estimator and its asymptotic properties were not analyzed extensively. Reference [17] suggested an information theoretic estimator based on minimization of the Kullback-Leibler Information Criterion as an alternative to optimally-weighted generalized method of moments estimation that can accommodate weakly dependent data generating mechanisms. Subsequently, [18] investigated an information theoretic estimator based on minimization of the Cressie-Read discrepancy statistic as an alternative approach to inference in models whose data information was cast in terms of moment conditions. Reference [18] identified both exponential empirical likelihood (negative entropy) and empirical likelihood as special cases of the Cressie-Read power divergence statistic. More recently, [19,20] applied the Kullback-Leibler Information Criterion to define empirical moment equations leading to estimators with improved predictive accuracy and mean square error in some small sample estimation contexts. Reference [21] provided an overview of information theoretic estimators for the SEM. Reference [22] demonstrated that maximum entropy estimation of the SEM has relevant application to spatial autoregressive models wherein autocorrelation parameters are inherently bounded and in circumstances when traditional spatial estimators become unstable. Reference [23] examined the effect of management factors on enterprise performance using a GME SEM estimator. Finally, [24] estimated spatial structural equation models also extended to a panel data framework.

In this paper we investigate a GME estimator for the linear SEM that is fundamentally different from traditional approaches and identify classes of problems (e.g., contaminated error models) in

which the proposed estimator outperforms traditional estimators. The estimator: (1) is completely consistent with data and other model information constraints on parameters, even in finite samples; (2) has large sample justification in that, under regularity conditions, it retains properties of consistency and asymptotic normality to provide practitioners with means to apply standard hypothesis testing procedures; and (3) has the potential for improved finite sample properties relative to alternative traditional methods of estimation. The proposed estimator is a one-step instrumental variable-type estimator based on a nonlinear-in-parameters SEM model discussed in [1,7,25]. The method does not deal with data information by projecting it in the form of moment constraints but rather, in GME parlance, is based on data constraints that deal with the data in individual sample observation form. Additional information utilized in the GME estimator includes finite support spaces that are imposed on model parameters and disturbances, which allows users to incorporate *a priori* interval restrictions on the parameters of the model.

Monte Carlo (MC) sampling experiments are used to investigate the finite sample performance of the proposed GME estimator. In the small sample situations analyzed, the GME estimator is superior to two and three stage least squares based on mean square error considerations. Further, we demonstrate the improved robustness of GME relative to 3SLS in the case of contaminated error models. For larger sample sizes, the consistency of the GME estimator results in sampling behavior that emulates that of 2SLS and 3SLS estimators. Observations on power and size of asymptotic test statistics suggest that the GME does not dominate, nor is it dominated by, traditional testing methods. An empirical application is provided to demonstrate practical implementation of the GME estimator and to delineate inherent differences between GME and traditional estimators in finite samples. The empirical analysis also highlights the sensitivity of GME coefficient estimates and predictive fit to specification of error truncation points, underscoring the need for care in specifying the empirical error support.

## 2. The GME-Parameterized Simultaneous Equations Model

Consider the SEM with  $G$  equations, which can be written in matrix form as:

$$\mathbf{Y} \mathbf{\Gamma} + \mathbf{X} \mathbf{B} + \mathbf{E} = \mathbf{0} \quad (1)$$

where  $\mathbf{Y} = (\mathbf{y}_1 \dots \mathbf{y}_G)$  is a  $(N \times G)$  matrix of jointly determined endogenous variables,  $\mathbf{\Gamma} = (\mathbf{\Gamma}_1 \dots \mathbf{\Gamma}_G)$  is an invertible  $(G \times G)$  matrix of structural coefficients of the endogenous variables,  $\mathbf{X} = (\mathbf{x}_1 \dots \mathbf{x}_K)$  is a  $(N \times K)$  matrix of exogenous variables that has full column rank,  $\mathbf{B} = (\mathbf{B}_1 \dots \mathbf{B}_G)$  is a  $(K \times G)$  matrix of coefficients of exogenous variables, and  $\mathbf{E} = (\mathbf{\varepsilon}_1 \dots \mathbf{\varepsilon}_G)$  is a  $(N \times G)$  matrix of unobserved random disturbances. The standard stochastic assumptions of the disturbance vectors are that  $E[\mathbf{\varepsilon}_i] = \mathbf{0}$  for  $i = 1, \dots, G$  and  $E[\mathbf{\varepsilon}_i \mathbf{\varepsilon}_j'] = \sigma_{ij} \mathbf{I}_N$  for  $ij = 1, \dots, G$ . Letting  $\boldsymbol{\varepsilon} = \text{vec}(\mathbf{\varepsilon}_1 \dots \mathbf{\varepsilon}_G)$  denote the vertical concatenation of the vectors  $\mathbf{\varepsilon}_1, \dots, \mathbf{\varepsilon}_G$ , the covariance matrix is given by  $E[\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}'] = \boldsymbol{\Sigma} \otimes \mathbf{I}_N$  where the  $(G \times G)$  matrix  $\boldsymbol{\Sigma}$  contains the unknown  $\sigma_{ij}$ 's for  $ij = 1, \dots, G$ .

The reduced form model is obtained by post-multiplying Equation (1) by  $\mathbf{\Gamma}^{-1}$  and solving for  $\mathbf{Y}$  as:

$$\mathbf{Y} = \mathbf{X}(-\mathbf{B}\mathbf{\Gamma}^{-1}) + (-\mathbf{E}\mathbf{\Gamma}^{-1}) = \mathbf{X}\mathbf{\Pi} + \mathbf{V} \quad (2)$$

where  $\mathbf{\Pi} = (\boldsymbol{\pi}_1 \dots \boldsymbol{\pi}_G)$  is a  $(K \times G)$  matrix of reduced form coefficients and  $\mathbf{V} = (\mathbf{v}_1 \dots \mathbf{v}_G)$  is a  $(N \times G)$  matrix of reduced form disturbances. The reduced form for the  $i$ th endogenous variable is:

$$\mathbf{y}_i = \mathbf{X}\boldsymbol{\pi}_i + \mathbf{v}_i \tag{3}$$

The  $i$ th equation in Equation (1) can be rewritten in terms of a nonlinear structural parameter representation of the reduced form model as [1]:

$$\mathbf{y}_i = \mathbf{X}\mathbf{\Pi}_{(-i)}\boldsymbol{\gamma}_i + \mathbf{X}_i\boldsymbol{\beta}_i + \boldsymbol{\mu}_i = \mathbf{Z}\boldsymbol{\delta}_i + \boldsymbol{\mu}_i \tag{4}$$

where  $E[\mathbf{Y}_{(-i)}] = \mathbf{X}\mathbf{\Pi}_{(-i)}$ ,  $\boldsymbol{\mu}_i = \boldsymbol{\varepsilon}_i + (\mathbf{Y}_{(-i)} - E[\mathbf{Y}_{(-i)}])\boldsymbol{\gamma}_i$ ,  $\mathbf{Z}_i = (\mathbf{X}\mathbf{\Pi}_{(-i)} \mathbf{X}_i)$ , and  $\boldsymbol{\delta}_i = \text{vec}(\boldsymbol{\gamma}_i, \boldsymbol{\beta}_i)$ .

In general the notation  $(-i)$  in the subscript of a variable represents the explicit exclusion of the  $i$ th column vector, such as  $\mathbf{y}_i$  being excluded from  $\mathbf{Y}$  to form  $\mathbf{Y}_{(-i)}$ , in addition to the exclusion of any other column vectors implied by the structural restrictions. Then  $\mathbf{Y}_{(-i)}$  represents a  $(N \times G_i)$  matrix of  $G_i$  jointly dependent explanatory variables having nonzero coefficients in the  $i$ th equation,  $\boldsymbol{\gamma}_i$  is the corresponding  $(G_i \times 1)$  subvector of the structural parameter vector  $\boldsymbol{\Gamma}_i$ ,  $\mathbf{X}_i$  is a  $(N \times K_i)$  matrix that represents the  $K_i$  exogenous variables with nonzero coefficients in the  $i$ th equation, and  $\boldsymbol{\beta}_i$  is the corresponding  $(K_i \times 1)$  subvector of the parameter vector  $\mathbf{B}_i$ . It is assumed that the linear exclusion restrictions on the structural parameters are sufficient to identify each equation. The  $(K \times G_i)$  matrix of reduced form coefficients  $\mathbf{\Pi}_{(-i)}$  coincides with the endogenous variables in  $\mathbf{Y}_{(-i)}$ .

Historically, Equation (4) has provided motivation for two stage least squares (2SLS) and three stage least squares (3SLS) estimators. The presence of right hand side endogenous variables yields biased and inconsistent estimates for  $\mathbf{Y}_{(-i)}$  [1]. In 2SLS and 3SLS, the first stage is to approximate  $E[\mathbf{Y}_{(-i)}]$  by applying ordinary least squares (OLS) to the unrestricted reduced form model in Equation (2) and thereby obtain predicted values of  $\mathbf{Y}_{(-i)}$ . Then, using the predicted values to replace  $E[\mathbf{Y}_{(-i)}]$ , the second stage is to estimate the model in Equation (4) with OLS. In the event that the error terms are normally distributed, homoskedastic, and serially independent, the 3SLS estimator is asymptotically equivalent to the asymptotically efficient full-information maximum likelihood (FIML) estimator [21]. Under the same conditions, it is equivalent to apply FIML to either Equation (1) or to Equation (4) under the restriction  $\mathbf{\Pi} = -\mathbf{B}\boldsymbol{\Gamma}^{-1}$ .

### 2.1. GME Estimation of the SEM

Following the maximum entropy principle, the entropy of a distribution of probabilities  $\mathbf{q} = (q_1, \dots, q_N)'$ ,  $\sum_{n=1}^N q_n = 1$ , is defined by:

$$H(\mathbf{q}) = -\sum_{n=1}^N q_n \ln q_n$$

in [26]. The value of  $H(\mathbf{q})$  reaches a maximum when  $q_n = N^{-1}$  for  $n = 1, \dots, N$ , which characterizes the uniform distribution. Generalizations of the entropy function that have been examined elsewhere in the econometrics and statistics literature include the Cressie-Read power divergence statistic [18], Kullback-Leibler Information Criterion [27], and the  $\alpha$ -entropy measure [28]. We restrict our analysis

to the entropy objective function due to its efficiency and robustness properties [18], and its current universal use within the context of GME applications [9].

GME estimators previously proposed for the SEM include (a) the data constrained estimator for the general linear model, hereafter GME-D, which amounts to applying the GME principle to a vectorized version of the structural model in Equation (1); and (b) a two stage estimator analogous to 2SLS whereby GME-D is applied to the reduced form model in the first stage and to the structural model in the second stage, hereafter GME-2S. Alternatively, [10] applied the GME principle to the reduced form model in Equation (3) with the restriction  $\Pi = -B\Gamma^{-1}$  imposed, hereafter GME-GJM.

Our approach follows 2SLS and 3SLS in the sense that the restriction  $\Pi = -B\Gamma^{-1}$  is not explicitly enforced and that  $E[Y_{(-i)}]$  is algebraically replaced by  $X\Pi_{(-i)}$ . However, unlike 2SLS and 3SLS, our approach is formulated under the GME principle completely consistent with Equation (4) retained as a nonlinear constraint and concurrently solved with the unrestricted reduced form model in Equation (3) to identify structural and reduced form coefficient estimates. Reference [7] refers to Equations (3) and (4) as a nonlinear-in-parameters (NLP) form of the SEM model.

To formulate a GME estimator for the NLP model of the SEM, henceforth referred to as GME-NLP, parameters and disturbance terms of Equations (3) and (4) are reparameterized as convex combinations of reference support points and unknown convexity weights. Support matrices  $S^i$  for  $i = \pi, \gamma, \beta, z, w$  that identify finite bounded feasible spaces for individual parameters and weight vectors  $p^\beta, p^\gamma, p^\pi, z, w$  that consist of unknown parameters to be estimated are explicitly defined below. The parameters are redefined as  $\beta = \text{vec}(\beta_1, \dots, \beta_G) = S^\beta p^\beta$ ,  $\gamma = \text{vec}(\gamma_1, \dots, \gamma_G) = S^\gamma p^\gamma$ , and  $\pi = \text{vec}(\pi_1, \dots, \pi_G) = S^\pi p^\pi$ , while the disturbance vectors are defined as  $v = \text{vec}(v_1, \dots, v_G) = S^z z$ , and  $\mu = \text{vec}(\mu_1, \dots, \mu_G) = S^w w$ . Using these identities and letting  $p = \text{vec}(p^\beta, p^\gamma, p^\pi, z, w)$  the estimates of  $\pi, \gamma, \beta$  are obtained by solving the constrained GME problem:

$$\max_p \{-p' \ln p\} \tag{5}$$

subject to:

$$y = (I_G \otimes X)(S_{(-)}^\pi p^\pi)(S^\gamma p^\gamma) + X^\beta (S^\beta p^\beta) + S^w w \tag{6}$$

$$y = (I_G \otimes X)(S^\pi p^\pi) + S^z z \tag{7}$$

$$(I_{Q+2NG} \otimes \mathbf{1}'_M) p = \mathbf{1}_{Q+2NG} \tag{8}$$

The  $S^i$  support matrices (for  $i = \pi, \gamma, \beta, z, w$ ) present in Equations (6) and (7) consist of user supplied reference support points defining feasible spaces for parameters and disturbances. For example,  $S^w$  is given by:

$$S^w = \begin{pmatrix} S_1^w & 0 & \dots & 0 \\ 0 & S_2^w & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & S_G^w \end{pmatrix}_{(GN \times GNM)} \quad S_i^w = \begin{pmatrix} s_{1i}^{w'} & 0 & \dots & 0 \\ 0 & s_{2i}^{w'} & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & s_{Ni}^{w'} \end{pmatrix}_{(N \times NM)} \quad s_{ni}^w = \begin{pmatrix} s_{ni1}^w \\ s_{ni2}^w \\ \cdot \\ s_{niM}^w \end{pmatrix}_{(M \times 1)} \tag{9}$$

where the  $n$ th disturbance term of the  $g$ th equation with  $M$  support points is defined, in summation notation, as  $\mu_{ng} = \sum_{m=1}^M s_{ngm}^w w_{ngm}$ . Similarly, the  $k$ th  $\beta$  parameter of the  $g$ th equation is defined by  $\beta_{kg} = \sum_{m=1}^M s_{kgm}^\beta p_{kgm}^\beta$ . For notational convenience the number of support points have been defined as  $M \geq 2$  for both errors and parameters.

In Equation (6), the matrix  $S_{(-)}^\pi$  defines the reference supports for the block diagonal matrix  $f$ , while  $X^\beta = \text{diag}(X_1, \dots, X_G)$  is a  $(GN \times \bar{K})$  block diagonal matrix and  $y = \text{vec}(y_1, \dots, y_G)$  is a  $(GN \times 1)$  vector of endogenous variables. In Equations (6) and (7) the  $(NGM \times 1)$  vectors  $w = \text{vec}(w_{11}, \dots, w_{NG})$  and  $z = \text{vec}(z_{11}, \dots, z_{NG})$  represent vertical concatenations of sets of  $(M \times 1)$  subvectors for  $n = 1, \dots, N$  and  $g = 1, \dots, G$ , where each subvector  $w_{ng} = (w_{ng1}, \dots, w_{ngM})'$  and  $z_{ng} = (z_{ng1}, \dots, z_{ngM})'$  contains a set of  $M$  convex weights. Also  $p^\pi = \text{vec}(p_{11}^\pi, \dots, p_{KG}^\pi)$  is a  $(KGM \times 1)$  vector that consists of convex weights  $p_{kg}^\pi = (p_{kg1}^\pi, \dots, p_{kgM}^\pi)'$  for  $k = 1, \dots, K$  and  $g = 1, \dots, G$ . The  $(M\bar{G} \times 1)$  vector  $p^\gamma = \text{vec}(p_{11}^\gamma, \dots, p_{GG}^\gamma)$  and the  $(\bar{K}M \times 1)$  vector  $p^\beta = \text{vec}(p_{11}^\beta, \dots, p_{KG}^\beta)$  are similarly defined. Equation (8) contains the required adding up conditions for each of the sets of convexity weights used in forming the GME-NLP estimator. Nonnegativity of the weights is an inherent characteristic of the maximum entropy objective and does not need to be explicitly enforced with inequality constraints. Regarding notation in (8),  $I_G$  represents a  $(G \times G)$  identity matrix and  $\mathbf{1}_N$  is a  $(N \times 1)$  unit vector. Letting  $\bar{K} = \sum_{i=1}^G K_i$  denote the number of unknown  $\beta_{kg}$ 's and  $\bar{G} = \sum_{i=1}^G G_i$  denote the number of unknown  $\gamma_{ig}$ 's, then together with the  $KG$  reduced form parameters, the  $\pi_{kg}$ 's, the total number of unknown parameters in the structural and reduced form equations is  $Q = \bar{K} + \bar{G} + KG$ .

Optimizing the objective function defined in Equation (5) optimizes the entropy in the parameter and disturbance spaces for both the structural model in Equation (6) and the reduced form model in Equation (7). The optimized objective function can mitigate the detrimental effects of ill-conditioned explanatory and/or instrumental variables and extreme outliers due to heavy tailed sampling distributions. In these circumstances traditional estimators are unstable and often represent an unsatisfactory basis for estimation and inference [20,25,29].

We emphasize that the proposed GME-NLP is a data-constrained estimator. Equations (5)–(8) constitute a data-constrained model in which the regression models themselves, as opposed to moment conditions based on them, represent constraining functions to the entropy objective function. [16] pointed out that outside the Gaussian error model, estimation based on sample moments can be inefficient relative to other procedures. Reference [9] provided MC evidence that data-constrained GME models, making use of the full set of observations, outperformed moment-constrained GME models in mean square error. In the GME-NLP model, constraints Equations (6) and (7) remain completely consistent with sample data information in Equations (3) and (4).

We also emphasize that the proposed GME-NLP estimator is a one-step approach, simultaneously solving for reduced form and structural parameters. As a result, the nonlinear specification of Equation (6) leads to first order optimization conditions (Equation A15) derived in the Appendix) that are different from other multiple-step or asymptotically justified estimators. The most obvious difference is that the first order conditions do not require orthogonality between right hand side variables and error terms,

*i.e.*, GME-NLP relaxes the orthogonality condition between instruments and the structural error term. Perhaps more importantly, multiple-step estimators (e.g., 2SLS or GME-2S) only approximate the NLP model and ignore nonlinear interactions between reduced and structural form coefficients. Thus, constraints Equations (6) and (7) are not completely satisfied by multiple-step procedures, yielding an estimator that is not fully consistent with the entire information set underlying the specification of the model. Although this is not a critical issue in large sample estimation, as demonstrated below, estimation inefficiency can be substantial in small samples if multiple-step estimators do not adequately approximate the NLP model.

The proposed GME-NLP estimator has some econometric limitations similar to, and other limitations which set it apart from, 2SLS that are evident when inspecting Equations (5)–(8). Firstly, like 2SLS, the residuals in Equations (4) and (6) are not identical to those of the original structural model, nor are they the same as the reduced form error term, except when evaluated at the true parameter values. Secondly, the GME-NLP estimator does not attempt to correct for contemporaneous correlation among the errors of the structural equations. Although a relevant efficiency issue, contemporaneous correlation is left for future research. Thirdly, and perhaps most importantly, the use of bounded disturbance support spaces in GME estimation introduces a specification issue in empirical analysis that typically does not arise with traditional estimators. These issues are discussed in more detail ahead.

## 2.2. Parameter Restrictions

In practice, parameter restrictions for coefficients of the SEM have been imposed using constrained maximum likelihood or Bayesian regression [7,30]. Neither approach is necessarily simple enough to specify analytically nor estimate empirically, and each has its empirical advantages and disadvantages. For example, Bayesian estimation is well-suited for representing uncertainty with respect to model parameters, but can also require extensive MC sampling when numerical estimation techniques are required, as is often the case in non-normal, non-conjugate prior model contexts. In comparison to constrained maximum likelihood or Bayesian analysis, the GME-NLP estimator also enforces restrictions on parameter values, is arguably no more difficult to specify or estimate, and does not require the use of MC sampling in the estimation phase of the analysis. Moreover, and in contrast to constrained maximum likelihood or the typical parametric Bayesian analysis, GME-NLP does not require explicit specification of the distributions of the disturbance terms or of the parameter values. However, both the coefficient and the disturbance support spaces are compact in the GME-NLP estimation method, which may not apply in some idealized empirical modeling contexts.

Imposing bounded support spaces on coefficients and error terms has several implications for GME estimation. Consider support spaces for coefficients. Selecting bounds and intermediate reference support points provides an effective way to restrict parameters of the model to intervals. If prior knowledge about coefficients is limited, wider truncation points can be used to increase the confidence that the support space contains the true  $\beta$ . If knowledge exists about, say, the sign of a specific coefficient from economic theory, this can be straightforwardly imposed together with a reasonable bound on the coefficient.

Importantly, there is a bias-efficiency tradeoff that arises when parameter support spaces are specified in terms of bounded intervals. A disadvantage of bounded intervals is that they will generally introduce bias into the GME estimator unless the intervals happen to be centered on the true values of the parameters. An advantage of restricting parameters to finite intervals is that they can lead to increases in efficiency by lowering parameter estimation variability. In the MC analysis ahead, it is demonstrated that the bias introduced by bounded parameter intervals in the GME-NLP estimator can be much more-than compensated for by substantial decreases in variability, leading to notable increases in overall estimation efficiency.

In practice, support spaces for disturbances can always be chosen in a manner that provides a reasonable approximation to the true disturbance distribution because upper and lower truncation points can always be selected sufficiently wide to contain the true disturbances of regression models [31]. The number,  $M$ , of support points for each disturbance can be chosen to account for additional information relating to higher moments (e.g., skewness and kurtosis) of each disturbance term. MC experiments by [9] demonstrated that support points ranging from 2 to 10 are acceptable for empirical applications.

For the GME-NLP estimator, identifying bounds for the disturbance support spaces is complicated by the interaction among truncation points of the parameters and disturbance support points of both the reduced and structural form models. Yet, several informative generalizations can be drawn. First, [32] demonstrated that ordinary least squares-like behavior can be obtained by appropriately selecting truncation points of the GME-D estimator of the general linear model. This has direct implications to SEM estimation in that appropriately selected truncation points of the GME-2S estimator leads to 2SLS-like behavior. However, as demonstrated ahead, given the nonlinear interactions between the structural and reduced form models, adjusting truncation points of the GME-NLP does not necessarily lead to two stage like behavior in finite samples. Second, the reduced form model in Equation (3) and the nonlinear structural parameter representation of the reduced form model in Equation (4) have identical error structure at the true parameter values. Hence, in the empirical applications below, we specify identical support matrices for error terms of both the structural and reduced form models. Third, in the limiting case where the disturbance boundary points of the GME-NLP structural model expand in absolute value to infinity, the parameter estimates converge to the mean of their support points.

Given ignorance regarding the disturbance distribution, [9,10] suggest using a sample scale parameter and the multiple-sigma truncation rule to determine error bounds. For example, the three sigma rule for random variables states that the probability of a unimodal continuous random variable assuming outcomes distant from its mean by more than three standard deviations is at most 5% [33]. Intuitively, this multiple-sigma truncation rule provides a means of encompassing an arbitrarily large proportion of the disturbance support space. From the empirical evidence presented below, it appears that combining the three sigma rule with a sample scale parameter to estimate the GME-NLP model is a useful approach.

### 3. GME-NLP Asymptotic Properties and Inference

To derive consistency and asymptotic normality results for the GME-NLP estimator, we assume the following regularity conditions.

R1. The  $N$  rows of the  $(N \times G)$  disturbance matrix  $\mathbf{E}$  are independent random drawings from an  $G$ -dimensional population with zero mean vector and unknown finite covariance matrix  $\Sigma$ .



R2. The  $(N \times K)$  matrix  $\mathbf{X}$  of exogenous variables has rank  $K$  and consists of nonstochastic elements, with  $\lim_{N \rightarrow \infty} \left(\frac{1}{N} \mathbf{X}'\mathbf{X}\right) = \mathbf{\Omega}$  where  $\mathbf{\Omega}$  is a positive definite matrix.

R3. The elements  $\mu_{ng}$  of the vector  $\mathbf{v}_g = \boldsymbol{\mu}_g$  ( $n = 1, \dots, N, g = 1, \dots, G$ ) are independent and bounded such that  $c_{g1} + \omega_g \leq \mu_{ng} \leq c_{gM} - \omega_g$  for some  $\omega_g > 0$  and large enough positive  $c_{gM} = \square c_{g1}$ . The probability density function of  $\boldsymbol{\mu}$  is assumed to be symmetric about the origin with a finite covariance matrix.

R4.  $\pi_{kg} \in (\pi_{kgL}, \pi_{kgH})$ , for finite  $\pi_{kgL}$  and  $\pi_{kgH}$ ,  $\forall k = 1, \dots, K$  and  $g = 1, \dots, G$ .

$\gamma_{jg} \in (\gamma_{jgL}, \gamma_{jgH})$ , for finite  $\gamma_{jgL}$  and  $\gamma_{jgH}$ ,  $\forall (j \neq g) j, g = 1, \dots, G$ ; and  $\gamma_{gg} = -1$ .

$\beta_{kg} \in (\beta_{kgL}, \beta_{kgH})$ , for finite  $\beta_{kgL}$  and  $\beta_{kgH}$ ,  $\forall k = 1, \dots, K$  and  $g = 1, \dots, G$ .

R5. For the true  $\mathbf{B}$  and nonsingular  $\mathbf{\Gamma}$ , there exists positive definite matrices  $\boldsymbol{\Psi}_g$  ( $g = 1, \dots, G$ ) such that  $\lim_{N \rightarrow \infty} \left(\frac{1}{N} \mathbf{Z}_g' \mathbf{Z}_g\right) \rightarrow \boldsymbol{\Psi}_g$  where  $\mathbf{\Pi} = -\mathbf{B}\mathbf{\Gamma}^{-1}$ .

Condition R1 asserts that the disturbances are contemporaneously correlated. It also requires independence of the  $N$  rows of the  $(N \times G)$  disturbance matrix  $\mathbf{E}$ , which is stronger than the uncorrelated error assumptions introduced immediately following Equation (1). Conditions R1, R2, and R5 are typical assumptions made when deriving asymptotic properties for the 2SLS and 3SLS estimators of the SEM [1]. The condition R3 states that the supports of  $\mu_{ng}$  and  $v_{ng}$  are symmetric about the origin and can be contained in the interior of closed and bounded intervals  $[c_1, c_M]$ . Extending the lower and upper bounds of the interval by (possibly arbitrarily small)  $\omega_g > 0$  is a technical and computational convenience ensuring feasibility of the entropic solutions [32]. Condition R4 implies that the true value of the parameters  $\pi_{kg}, \gamma_{jg}, \beta_{kg}$  can be enclosed within a bounded interval.

### 3.1. Estimator Properties

The regularity conditions (R1)-(R5) provide a basic set of assumptions sufficient to establish asymptotic properties for the GME-NLP estimator of the SEM. For notational convenience let  $\boldsymbol{\theta} = \text{vec}(\boldsymbol{\pi}, \boldsymbol{\delta})$ , where we follow the standard convention that  $\boldsymbol{\delta} = \text{vec}(\boldsymbol{\delta}_1, \dots, \boldsymbol{\delta}_G)$ . The theorems for consistency and asymptotic normality are stated below with proofs in the Appendix.

**Theorem 1.** Under the regularity conditions R1-R5, the GME-NLP estimator,  $\hat{\boldsymbol{\theta}} = \text{vec}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\delta}})$ , is a consistent estimator of the true coefficient values  $\boldsymbol{\theta} = \text{vec}(\boldsymbol{\pi}, \boldsymbol{\delta})$ .

The intuition behind the proof is that without the reduced form component in Equation (7) the parameters of the structural component in Equation (6) are not identified. As shown in the Appendix, the reduced form component yields estimates that are consistent and contribute to identifying the structural parameters, and the structural component in Equation (7) ties the structural coefficients to the data and draws the GME-NLP estimates toward the true parameter values as the sample size increases.

**Theorem 2.** Under the conditions of Theorem 1, the GME-NLP estimator,  $\hat{\delta} = \text{vec}(\hat{\delta}_1, \dots, \hat{\delta}_G)$ , is asymptotically normally distributed as  $\hat{\delta} \overset{a}{\sim} N(\delta, \frac{1}{N} \Omega_\xi^{-1} \Omega_\Sigma \Omega_\xi^{-1})$ .

The asymptotic covariance matrix consists of  $\Omega_\xi = \text{diag}(\xi_1 \Psi_1, \dots, \xi_G \Psi_G)$ , which follows from R5 and  $\xi_g = E[\xi_{ng}^w]$  with  $\xi_{ng}^w = \frac{\partial \lambda^w(u_{ng})}{\partial u_{ng}} = \left( \sum_{m=1}^M (s_{ngm}^w)^2 w_{ngm} (\lambda^w(u_{ng})) - (u_{ng})^2 \right)^{-1}$ . The elements of  $\Omega_\Sigma$  are defined by  $\frac{1}{N} \mathbf{Z}'(\Sigma_\lambda \otimes \mathbf{I})\mathbf{Z} \rightarrow \Omega_\Sigma$ , where  $\mathbf{Z} = \text{diag}(\mathbf{Z}_1, \dots, \mathbf{Z}_G)$  and  $\Sigma_\lambda$  is a  $(G \times G)$  covariance matrix for the  $\lambda_{ng}^w$ 's.

Estimators of the SEM are generally categorized as “full information” (e.g., 3SLS or FIML) or “limited information” (e.g., 2SLS or LIML) estimators. GME-NLP is not a full information estimator because the estimator neither enforces the restriction  $\Pi = -\mathbf{B}\Gamma^{-1}$  nor explicitly characterizes the contemporaneous correlation of the disturbance terms. An advantage of GME-NLP is that it is completely consistent with data constraints in both small and large samples, because we concurrently estimate the parameters of the reduced form and structural models. As a limited information estimator, GME-NLP has several additional attractive characteristics. First, similar to other limited information estimators, it is likely to be more robust to misspecification than a full information alternative because in the latter case misspecification of any one equation can lead to inconsistent estimation of all the equations in the system [34]. Second, GME-NLP is easily applied in the case of a single equation,  $G = 1$ , and it retains the asymptotic properties identified above. Finally, the single equation case is a natural generalization of the data-constrained GME estimator for the general linear model.

### 3.2. Hypothesis Tests

Because the GME-NLP estimator  $\hat{\delta}$  is consistent and asymptotically normally distributed, asymptotically valid normal and chi-square test statistics can be used to test hypothesis about  $\delta$ . To implement such tests a consistent estimate of the asymptotic covariance of  $\hat{\delta}$ , or  $\Omega_\xi^{-1} \Omega_\Sigma \Omega_\xi^{-1}$ , is required. The matrix  $\Omega_\xi$  can be estimated using  $\xi_{ng}^w(\hat{\delta})$  above or alternatively by:

$$\hat{\xi}_g(\hat{\delta}) = \frac{1}{N} \sum_{n=1}^N \left( \sum_{m=1}^M (s_{ngm}^w)^2 w_{ngm} (\lambda^w(u_{ng}(\hat{\delta}))) - (u_{ng}(\hat{\delta}))^2 \right)^{-1}$$

In the former case based on  $\xi_{ng}^w(\hat{\delta})$ , which are the elements of  $\Xi_g^w$  as defined in the Appendix, then  $\hat{\Omega}_\xi = \text{diag}\left(\frac{1}{N} \hat{\mathbf{Z}}_1' (\hat{\Xi}_1^w \odot \hat{\mathbf{Z}}_1), \dots, \frac{1}{N} \hat{\mathbf{Z}}_G' (\hat{\Xi}_G^w \odot \hat{\mathbf{Z}}_G)\right)$ . In the latter case based on  $\hat{\xi}_g$  and  $\hat{\Psi}_g = \left(\frac{1}{N} \hat{\mathbf{Z}}_g' \hat{\mathbf{Z}}_g\right)$ , then  $\hat{\Omega}_\xi = \text{diag}(\hat{\xi}_1 \hat{\Psi}_1, \dots, \hat{\xi}_G \hat{\Psi}_G)$ . A straightforward estimate of  $\Omega_\Sigma$  can be constructed as  $\hat{\Omega}_\Sigma = \frac{1}{N} \hat{\mathbf{Z}}' (\hat{\Sigma}_\lambda \otimes \mathbf{I}) \hat{\mathbf{Z}}$ . The  $(G \times G)$  matrix  $\Sigma_\lambda$  can be estimated by  $\hat{\sigma}_{ij}^\lambda = \frac{1}{N} \lambda^w(u_{.i}(\hat{\delta}))' \lambda^w(u_{.j}(\hat{\delta}))$  for  $i, j = 1, \dots, G$ . Combining these elements, the estimated asymptotic covariance matrix of  $\hat{\delta}$  is defined as  $\hat{\text{Var}}(\hat{\delta}) = \frac{1}{N} \hat{\Omega}_\xi^{-1} \hat{\Omega}_\Sigma \hat{\Omega}_\xi^{-1}$ .

### 3.2.1. Asymptotically Normal Tests

Since  $Z = \frac{\hat{\delta}_{ij} - \delta_{ij}^0}{\sqrt{\hat{\text{Var}}(\hat{\delta}_{ij})}}$  is asymptotically  $N(0,1)$  under the null hypothesis  $H_0: \delta_{ij} = \delta_{ij}^0$ , the statistic  $Z$  can be used to test hypothesis about the values of the  $\delta_{ij}$ 's.

### 3.2.2. Wald Tests

To define Wald tests on the elements of  $\delta$ , let  $H_0: \mathbf{R}(\delta) = \mathbf{0}$  be the null hypothesis to be tested. Here  $\mathbf{R}(\delta)$  is a continuously differentiable  $L$ -dimensional vector function with rank  $\frac{\partial \mathbf{R}(\delta)}{\partial \delta} = L \leq K$ . In the special case of a linear null hypothesis  $H_0: \mathbf{R}\delta = \mathbf{r}$ , then  $\frac{\partial(\mathbf{R}\delta)}{\partial \delta} = \mathbf{R}$ . It follows from Theorem 5.37 in [35] that:

$$\sqrt{N} \left( \mathbf{R}(\hat{\delta}) - \mathbf{r} \right) \xrightarrow{d} N \left( [\mathbf{0}], \frac{\partial \mathbf{R}(\delta)}{\partial \delta}' \boldsymbol{\Omega}_\xi^{-1} \boldsymbol{\Omega}_\zeta \boldsymbol{\Omega}_\xi^{-1} \frac{\partial \mathbf{R}(\delta)}{\partial \delta} \right)$$

The Wald test statistic has a  $\chi^2$  limiting distribution with  $L$  degrees of freedom given as

$$W = \left( \mathbf{R}(\hat{\delta}) - \mathbf{r} \right)' \left( \frac{\partial \mathbf{R}(\hat{\delta})}{\partial \delta}' \hat{\text{Var}}(\hat{\delta}) \frac{\partial \mathbf{R}(\hat{\delta})}{\partial \delta} \right)^{-1} \left( \mathbf{R}(\hat{\delta}) - \mathbf{r} \right) \xrightarrow{d} \chi_L^2$$

under the null hypothesis.

## 4. Monte Carlo Experiments

For the sampling experiments we set up an overdetermined simultaneous system with contemporaneously correlated errors that is similar, but not identical, to empirical models discussed in [10,36,37]. Reference [10] provide empirical evidence of the performance of the GME-GJM estimator for both ill-posed (multicollinearity) and well-posed problems using a sample size of 20 observations. In this study we attempt to focus on both smaller and larger sample size performance of the GME-NLP estimator, the size and power of single and joint hypothesis tests, and the relative performance of GME-NLP to 2SLS and 3SLS. In addition, the performance of GME-NLP is compared to Golan, Judge, and Miller's GME-GJM estimator. The estimation performance measure is the mean square error (MSE) between the empirical coefficient estimates and the true coefficient values.

### 4.1. Parameters and Support Spaces

The parameters  $\Gamma$  and  $\mathbf{B}$  and the covariance structure  $\Sigma$  of the structural system in Equation (1) are specified as:

$$\Gamma = \begin{pmatrix} -1 & .267 & .087 \\ .222 & -1 & 0 \\ 0 & .046 & -1 \end{pmatrix} \quad \mathbf{B} = \begin{pmatrix} 6.2 & 4.4 & 4.0 \\ 0 & .74 & 0 \\ .7 & 0 & .53 \\ 0 & 0 & .11 \\ .96 & .13 & 0 \\ 0 & 0 & .56 \\ .06 & 0 & 0 \end{pmatrix} \quad \Sigma = \begin{pmatrix} 1 & -1 & -.125 \\ -1 & 4 & .0625 \\ -.125 & .0625 & 8 \end{pmatrix}$$

The exogenous variables are drawn from an *iid*  $N(0,1)$  distribution, while the errors for the structural equations are drawn from a multivariate normal distribution with mean zero and covariance  $\Sigma \otimes \mathbf{I}$  that is truncated at  $\pm 3$  standard deviations.

To specify the GME models, additional information beyond that traditionally used in 2SLS and 3SLS is required. Upper and lower bounds, as well as intermediate support points for the individual coefficients and disturbance terms, are supplied for the GME-NLP and GME-GJM models along with starting values for the parameter coefficients. The difference in specification of GME-GJM relative to GME-NLP is that in the former,  $\mathbf{\Pi} = -\mathbf{B}\mathbf{\Gamma}^{-1}$  replaces the structural model in Equation (6) and the GME-GJM objective function excludes any parameters associated with the structural form disturbance term. The upper and lower bounds of the support spaces specified for the structural and reduced form models are identical to [10] except that we use three rather than five support points. The supports are defined as  $s_{ik}^{\beta} = s_{ik}^{\pi} = (-5, 0, 5)'$  for  $k = 2, \dots, 7$ ,  $s_{i1}^{\beta} = s_{i1}^{\pi} = (-20, 0, 20)'$ , and  $s_{ij}^{\gamma} = (-2, 0, 2)'$  for  $ij = 1, 2, 3$ . The error supports for the reduced form and structural model were specified as  $s_{in}^z = s_{in}^w = (-\omega_i - 3\sigma_i, 0, \omega_i + 3\sigma_i)'$ , where  $\sigma_i$  is the standard deviation of the errors from the  $i$ th equation and from R3 we let  $\omega_i = 2.5$  to ensure feasibility. See appendix material for a more complete discussion of computational issues.

#### 4.2. Estimation Performance

Table 1 contains the mean values of the estimated  $\mathbf{\Gamma}$  parameters based on 1,000 MC repetitions for sample sizes of 5, 25, 100, 400, and 1,600 observations per equation. From this information, we can infer several implications about the performance of the GME estimators. For a sample size of five observations per equation, 2SLS and 3SLS estimators provide no solution due to insufficient degrees of freedom. For five and 25 observations the GME-NLP and GME-GJM estimators have mean values that are similar, although GME-NLP exhibits more bias. When the sample size is 100, the GME-NLP estimator generally exhibits less bias. Like 2SLS and 3SLS, the GME-NLP estimator is converging to the true coefficient values as  $N$  increases to 1,600 observations per equation (3SLS estimates are not reported for 1,600 observations).

In Table 2 the standard error (SE) and MSE are reported for 3SLS and GME-NLP. The GME-NLP estimator has uniformly lower standard error and MSE than does 3SLS. For small samples of 25 observations the MSE performance of the GME-NLP estimator is vastly improved relative to the 3SLS estimator, which is consistent with MC results from other studies relating to other GME-type estimators [9,32]. As the sample size increases from 25 to 400 observations, both the standard error and mean squared error of the 3SLS and GME-NLP converge towards each other. Interestingly, even at a sample size of 100 observations the GME-NLP mean squared error remains notably superior to 3SLS.

**Table 1.** Mean value of parameter estimates from 1000 Monte Carlo simulations using 2SLS, 3SLS, GME-GJM, and GME-NLP.

Obs	2SLS	3SLS	GME-GJM	GME-NLP
$\gamma_{21} = 0.222$				
5	-	-	0.331	0.353
25	0.165	0.186	0.304	0.311
100	0.207	0.220	0.357	0.259
400	0.219	0.222	0.373	0.234
1,600	0.223	-	0.393	0.227
$\gamma_{12} = 0.267$				
5	-	-	0.267	0.301
25	0.274	0.241	0.292	0.304
100	0.264	0.278	0.278	0.283
400	0.272	0.276	0.293	0.274
1,600	0.268	-	0.319	0.269
$\gamma_{32} = 0.046$				
5	-	-	0.144	0.158
25	0.067	0.103	0.107	0.144
100	0.044	0.048	0.101	0.083
400	0.039	0.040	0.095	0.053
1600	0.046	-	0.075	0.048
$\gamma_{13} = 0.087$				
5	-	-	0.197	0.223
25	0.115	0.114	0.182	0.208
100	0.084	0.085	0.165	0.139
400	0.083	0.083	0.155	0.100
1,600	0.088	-	0.153	0.093

### 4.3. Inference Performance

To investigate the size of the asymptotically normal test, the single hypothesis  $H_0: \gamma_{ij} = k$  was tested with  $k$  set equal to the true values of the structural parameters. Critical values of the tests were based on a normal distribution with a 0.05 level of significance. An observation on the power of the respective tests was obtained by performing a test of significance whereby  $k = 0$  in the preceding hypothesis. To complement this analysis, we investigated the size and power of a joint hypothesis  $H_0: \gamma_{21} = k_1, \gamma_{32} = k_2$  using the Wald test. The scenarios were analyzed using 1000 MC repetitions for sample sizes of 25, 100, and 400 per equation.

Table 3 contains the rejection probabilities for the true and false hypotheses of both the GME-NLP and 3SLS estimators. The single hypothesis test for the parameter  $\gamma_{21} = 0.222$  based on the asymptotically normal test responded well for GME-NLP (3SLS), yielding an estimated test size of 0.066 (0.043) and power of 0.980 (0.964) at 400 observations per equation. In contrast, for the remaining parameters, the size and power of the hypotheses tests were considerably less satisfactory. This is due in part to the second and third equations having substantially larger disturbance variability. For the joint hypothesis test based on the Wald test the size and power perform well for GME-NLP

(3SLS) with an estimated test size of 0.047 (0.047) and power of 0.961 (0.934) at 400 observations. Overall, the results indicate that based on asymptotic test statistics GME-NLP does not dominate, nor is it dominated by, 3SLS.

**Table 2.** Standard error (SE) and mean square error (MSE) of parameter estimates from 1000 Monte Carlo simulations using 3SLS and GME-NLP.

Obs	SE		MSE	
	3SLS	GME-NLP	3SLS	GME-NLP
$\gamma_{21} = 0.222$				
5	-	0.101	-	0.027
25	0.442	0.155	0.197	0.032
100	0.143	0.116	0.021	0.015
400	0.065	0.064	0.004	0.004
$\gamma_{12} = 0.267$				
5	-	0.103	-	0.012
25	1.281	0.166	1.641	0.029
100	0.459	0.183	0.211	0.034
400	0.198	0.149	0.039	0.022
$\gamma_{32} = 0.046$				
5	-	0.168	-	0.041
25	0.842	0.256	0.711	0.075
100	0.449	0.226	0.201	0.052
400	0.183	0.158	0.033	0.025
$\gamma_{13} = 0.087$				
5	-	0.120	-	0.033
25	0.669	0.202	0.448	0.055
100	0.269	0.188	0.073	0.038
400	0.133	0.121	0.018	0.015

#### 4.4. Further Results: 3-Sigma Rule and Contaminated Errors

Further MC results are presented to demonstrate the sensitivity of the GME-NLP to the sigma truncation rule (Table 4) and to illustrate robustness of the GME-NLP relative to 3SLS in the presence of contaminated error models (Table 5). Each of these issues play a critical role in empirical analysis of the SEM, while the latter can compound estimation problems especially in small sample estimation.

To obtain the results in Table 4, the error supports for the reduced form and structural model were specified as before with  $s_{in}^z = s_{in}^w = (-\omega_i - j\sigma_i, 0, \omega_i + j\sigma_i)'$  where  $\sigma_i$  is the standard deviation of the errors from the  $i$ th equation,  $j = 3, 4, 5$  and from R3  $\omega_i = 2.5$ , again for solution feasibility. The results exhibit a tradeoff between bias and MSE specific to the individual coefficient estimates. For  $\gamma_{21}$  the bias and the MSE decreases as the truncation points are shrunk from five to three sigma. In contrast, for the remaining coefficients in Table 4, the MSE increases as the truncation points are decreased. The bias decreases for  $\gamma_{32}$  and  $\gamma_{13}$  as the truncation points are shrunk, while the direction of bias is ambiguous for  $\gamma_{12}$ . Predominately, the empirical standard error of the coefficients decreased with

wider truncation points. Overall, these results underscore that the mean and standard error of GME-NLP coefficient values are sensitive to the choice of truncation points.

**Table 3.** Rejection Probabilities for True and False Hypotheses.

Single Hypotheses: Asymptotic Normal Test GME-NLP								
Obs	$\gamma_{21} = 0.222$	$\gamma_{21} = 0$	$\gamma_{12} = 0.267$	$\gamma_{12} = 0$	$\gamma_{32} = 0.046$	$\gamma_{32} = 0$	$\gamma_{13} = 0.087$	$\gamma_{13} = 0$
25	0.021	0.23	0.001	0.008	0.021	0.022	0.002	0.005
100	0.046	0.600	0.005	0.051	0.013	0.019	0.009	0.025
400	0.066	0.980	0.012	0.276	0.033	0.042	0.032	0.092
3SLS								
Obs	$\gamma_{21} = 0.222$	$\gamma_{21} = 0$	$\gamma_{12} = 0.267$	$\gamma_{12} = 0$	$\gamma_{32} = 0.046$	$\gamma_{32} = 0$	$\gamma_{13} = 0.087$	$\gamma_{13} = 0$
25	0.149	0.197	0.101	0.124	0.100	0.108	0.102	0.104
100	0.064	0.424	0.036	0.135	0.050	0.052	0.051	0.068
400	0.043	0.964	0.031	0.338	0.041	0.045	0.045	0.094
Joint Hypotheses: Asymptotic Chi-Square Wald Test								
	GME-NLP				3SLS			
	$\gamma_{21} = 0.222$	$\gamma_{21} = 0$	$\gamma_{32} = 0.046$	$\gamma_{32} = 0$	$\gamma_{21} = 0.222$	$\gamma_{21} = 0$	$\gamma_{32} = 0.046$	$\gamma_{32} = 0$
25		0.014		0.169		0.189		0.256
100		0.029		0.433		0.082		0.357
400		0.047		0.961		0.047		0.934

**Table 4.** Mean, standard error (SE), and mean square error (MSE) of parameter estimates from 1000 Monte Carlo simulations for GME-NLP with 3, 4, and 5-sigma truncation rules.

Obs	3-Sigma			4-Sigma			5-Sigma		
	Mean	SE	MSE	Mean	SE	MSE	Mean	SE	MSE
$\gamma_{21} = 0.222$									
25	0.311	0.155	0.030	0.336	0.133	0.031	0.345	0.111	0.033
100	0.259	0.116	0.015	0.277	0.111	0.015	0.292	0.108	0.017
400	0.234	0.064	0.004	0.244	0.066	0.005	0.247	0.063	0.005
$\gamma_{12} = 0.267$									
25	0.304	0.166	0.029	0.303	0.120	0.016	0.301	0.095	0.010
100	0.283	0.183	0.034	0.283	0.146	0.021	0.285	0.118	0.014
400	0.274	0.149	0.022	0.271	0.130	0.017	0.272	0.115	0.013
$\gamma_{32} = 0.046$									
25	0.144	0.256	0.075	0.144	0.203	0.051	0.164	0.152	0.037
100	0.083	0.226	0.052	0.101	0.199	0.042	0.113	0.158	0.029
400	0.053	0.158	0.025	0.063	0.137	0.019	0.068	0.128	0.017
$\gamma_{13} = 0.087$									
25	0.208	0.202	0.055	0.210	0.145	0.036	0.217	0.109	0.029
100	0.139	0.188	0.038	0.157	0.157	0.030	0.176	0.139	0.027
400	0.100	0.121	0.015	0.111	0.112	0.013	0.127	0.106	0.013

Results from Table 5 provide the mean and MSE of the distribution of coefficient estimates for 3SLS and GME-NLP when the error term is contaminated by outcomes from an asymmetric

distribution [14,15]. For a given percentage level  $\varphi$ , the errors for the structural equations are drawn from  $(1-\varphi)N(\mathbf{0}, \Sigma \otimes \mathbf{I}) + \varphi F(2,3)$  and then truncated at  $\pm 3$  standard deviations. We define  $F(2,3) = \text{Beta}(2,3) - 6$  and examine the robustness of 3SLS and GME-NLP with values of  $\varphi = 0.1, 0.5, \text{ and } 0.9$ . The error supports for the reduced form and structural model were specified with the three sigma rule. As evident in Table 5, when the percent of contamination induced in the error component of the SEM increases, performance of both estimators is detrimentally impacted. For 25 observations, the 3SLS coefficient estimates are much less robust to the contamination process than are the GME-NLP estimates as measured by the MSE values. At 100 observations the performance of 3SLS improves, but still remain less robust than GME-NLP.

**Table 5.** Mean and mean square error (in parentheses) of parameter estimates from 1000 Monte Carlo simulations for 3SLS and GME-NLP with contaminated normal distribution.

Obs	$0.90N(0, \Sigma) + 0.10F(2,3)$		$0.50N(0, \Sigma) + 0.50F(2,3)$		$0.10N(0, \Sigma) + 0.90F(2,3)$	
	3SLS	GME-NLP	3SLS	GME-NLP	3SLS	GME-NLP
$\gamma_{21} = 0.222$						
25	0.184 (0.159)	0.320 (0.032)	0.278 (0.406)	0.414 (0.064)	0.350 (1.404)	0.451 (0.082)
100	0.226 (0.023)	0.262 (0.016)	0.243 (0.082)	0.329 (0.037)	0.268 (0.204)	0.368 (0.050)
$\gamma_{12} = 0.267$						
25	0.262 (1.058)	0.309 (0.029)	0.427 (1.195)	0.385 (0.041)	0.608 (4.578)	0.422 (0.056)
100	0.267 (0.353)	0.282 (0.036)	0.356 (0.551)	0.339 (0.038)	0.374 (0.726)	0.364 (0.44)
$\gamma_{32} = .046$						
25	0.084 (0.794)	0.111 (0.067)	-0.009 (0.779)	0.105 (0.058)	-0.070 (2.489)	0.097 (0.062)
100	0.061 (0.326)	0.082 (0.049)	0.010 (0.395)	0.067 (0.048)	-0.003 (0.601)	0.075 (0.057)
$\gamma_{13} = 0.087$						
25	0.081 (0.330)	0.198 (0.048)	0.094 (0.401)	0.198 (0.056)	0.083 (1.366)	0.219 (0.067)
100	0.093 (0.061)	0.142 (0.036)	0.093 (0.059)	0.144 (0.038)	0.077 (0.124)	0.150 (0.055)

#### 4.5. Discussion

The performance of the GME-NLP estimator was based on a variety of MC experiments. In small and medium sample situations ( $\leq 100$  observations) the GME-NLP is MSE superior to 3SLS for the defined experiments. Increasing the sample size clearly demonstrated consistency of the GME-NLP estimator for the SEM. Regarding performance in single or joint hypothesis testing contexts, the empirical results indicate that the GME-NLP did not dominate, nor was it dominated by 3SLS.

The MC evidence provided above indicates that applying the multiple-sigma truncation rule with a sample scale parameter to estimate the GME-NLP model is a useful empirical approach. Across the 3,



4, and 5-sigma rule sampling experiments, GME-NLP continued to dominate 3SLS in MSE for 25, 100, and 400 observations per equation. For wider truncation points the empirical SE of the coefficients decreased. However, these results also demonstrate that the GME-NLP coefficients are sensitive to the choice of truncation points with no consensus in choosing narrower (3-sigma) over wider (5-sigma) truncation supports under a Gaussian error structure. We suggest that additional research is needed to *optimally* identify error truncation points.

Finally, the GME-NLP estimator exhibited more robustness in the presence of contaminated errors relative to 3SLS. The MC analysis illustrates that deviations from normality assumptions in asymptotically justified econometric-statistical models lead to dramatically less robust outcomes in small samples. Reference [9,16] emphasized that under traditional econometric assumptions, when samples are Gaussian in nature and sample moments are taken as minimal sufficient statistics, then no information may be lost. However, they point out that outside the Gaussian setting, reducing data constraints to moment constraints can be wasteful use of sample information and results in estimators that are less than fully efficient. The above MC analysis suggests that GME-NLP, which relies on full sample information but does not rely on a full parametric specification such as maximum likelihood, can be more robust to alternative error distributions.

## 5. Empirical Illustration

In this section, an empirical application is examined to demonstrate implementation of the GME-NLP estimator. It is the well known three-equation system that comprises the Klein Model I, which further benchmarks the GME-NLP estimator relative to least squares.

### 5.1. Klein Model

Klein's Model I was selected as an empirical application because it has been extensively applied in many studies. Klein's macroeconomic model is highly aggregated with relatively low parameter dimensionality, making it useful for pedagogical purposes. It is a three-equation SEM based on annual data for the United States from 1920 to 1941. All variables are in billions of dollars, which are constant dollars with base year 1934 (for a complete description of the model and data see [1,38]).

The model is comprised of three stochastic equations and five identities. The stochastic equations include demand for consumption, investment, and labor. Klein's consumption function is given as:

$$CN_t = \beta_{11} + \gamma_{11}(W_{1t} + W_{2t}) + \gamma_{21}P_t + \beta_{21}P_{t-1} + \varepsilon_{1t}$$

where  $CN_t$  is consumption,  $W_{1t}$  is wages earned by workers in the private sector,  $W_{2t}$  is wages earned by government workers,  $P_t$  is nonwage income (profit), and  $\varepsilon_{1t}$  is a stochastic error term. This equation describes aggregate consumption as a function of the total wage bill and current and lagged profit. The investment equation is given by:

$$I_t = \beta_{12} + \gamma_{12}P_t + \beta_{22}P_{t-1} + \beta_{32}K_{t-1} + \varepsilon_{12}$$

where  $I_t$  is net investment,  $K_t$  is the stock of capital goods at the end of the year, and  $\varepsilon_{2t}$  is a stochastic error term. This equation implies that net investment reacts to current and lagged profits, as well as beginning of the year capital stocks. The demand for labor is given by:

$$W_{1t} = \beta_{13} + \gamma_{13} E_t + \beta_{23} E_{t-1} + \beta_{33}(\text{Year} - 1931) + \varepsilon_{t3}$$

where  $E_t$  is a measure of private product and  $\varepsilon_{3t}$  is a stochastic error term. It implies that the wage bill paid by private industry varies with the current and lagged total private product and a time trend. A time trend is included to capture institutional changes over the period, primarily the bargaining strength of labor. The identities that complete the structural model include:

Total Product	$Y_t + TX_t = CN_t + I_t + G_t + W_{2t}$
Income	$Y_t = P_t + W_t$
Capital	$K_t = I_t + K_{t-1}$
Wage Bill	$W_t = W_{1t} + W_{2t}$
Private Product	$E_t = Y_t + TX_t - W_{2t}$

The first identity states that national income,  $Y_t$ , plus business taxes,  $TX_t$ , are equal to the sum of goods and services demanded by consumers,  $CN_t$ , plus investors,  $I_t$ , plus net government demands,  $G_t + W_{2t}$ . The second identity holds total income,  $Y_t$ , as the sum of profit,  $P_t$ , and wages,  $W_t$ , while the third implies that end-of-year capital stocks,  $K_t$ , are equal to investment,  $I_t$ , plus last years end-of-year capital stock,  $K_{t-1}$ . In the fourth identity,  $W_t$  is the total wage bill that is the sum of wages earned from the private sector,  $W_{1t}$ , and wages earned by the government,  $W_{2t}$ . The fifth identity states that private product,  $E_t$ , is the equal to income,  $I_t$ , plus business taxes,  $TX_t$ , less government wages,  $W_{2t}$ .

### 5.2. Klein Model I Results

Table 6 contains the estimates of the three stochastic equations using ordinary least squares (OLS), two stage least squares (2SLS), three stage least squares (3SLS), and GME-NLP. Parameter restrictions for GME-NLP were specified using the fairly uninformative reference support points  $(-50,0,50)'$  for the intercept,  $(-5,0,5)'$  for the slope parameters of the reduced form models and  $(-2,0,2)'$  for the slope parameters of the structural form models. Truncation points for the error supports of the structural model are specified using both three- and five-sigma rules.

For the given truncation points, the GME-NLP estimates of asymptotic standard errors are greater than those of the other estimators. It is to be expected that if more informative parameter support ranges had been used when representing the feasible space of the parameters, standard errors would have been reduced. In most of the cases, the parameter, standard error, and  $R^2$  measures were not particularly sensitive to the choice of error truncation point, although there were a few notable exceptions dispersed throughout the three equation system.

**Table 6.** Structural parameter estimates and standard errors (in parentheses) of Klein's Model I using OLS, 2SLS, 3SLS, and GME-NLP.

Structural Parameter	OLS	2SLS	3SLS	GME-NLP 3-sigma	GME-NLP 5-sigma
<b>Consumption</b>					
$\beta_{11}$	16.237 (1.303)	16.555 (1.468)	16.441 (12.603)	14.405 (2.788)	14.374 (2.625)
$\gamma_{11}$	0.796 (0.040)	0.810 (0.045)	0.790 (0.038)	0.772 (0.073)	0.750 (0.071)
$\gamma_{21}$	0.193 (0.091)	0.017 (0.131)	0.125 (0.108)	0.325 (0.372)	0.280 (0.306)
$\beta_{21}$	0.090 (0.091)	0.216 (0.119)	0.163 (0.100)	0.120 (0.332)	0.206 (0.274)
$R^2$	0.981	0.929	0.928	0.916	0.922
<b>Investment</b>					
$\beta_{12}$	10.126 (5.466)	20.278 (8.383)	28.178 (6.79)	8.394 (10.012)	9.511 (10.940)
$\gamma_{12}$	0.480 (0.097)	0.150 (0.193)	-0.013 (0.162)	0.440 (0.386)	0.358 (0.362)
$\beta_{22}$	0.333 (0.101)	0.616 (0.181)	0.756 (0.153)	0.340 (0.342)	0.350 (0.325)
$\beta_{32}$	-0.112 (0.027)	-0.158 (0.040)	-0.195 (0.033)	-0.100 (0.046)	-0.100 (0.051)
$R^2$	0.931	0.837	0.831	0.819	0.811
<b>Labor</b>					
$\beta_{13}$	1.497 (1.270)	1.500 (1.276)	1.797 (1.12)	2.423 (3.112)	1.859 (3.157)
$\gamma_{13}$	0.439 (0.032)	0.439 (0.040)	0.400 (0.032)	0.481 (0.255)	0.381 (0.178)
$\beta_{23}$	0.146 (0.037)	0.147 (0.043)	0.181 (0.034)	0.087 (0.272)	0.200 (0.180)
$\beta_{33}$	0.130 (0.032)	0.130 (0.032)	0.150 (0.028)	0.112 (0.091)	0.114 (0.085)
$R^2$	0.987	0.942	0.941	0.905	0.907

The Klein Model I benchmarks the GME-NLP estimator relative to OLS, 2SLS, and 3SLS. Comparisons are based on the sum of the squared difference (SSD) measures between GME-NLP and the OLS, 2SLS and 3SLS parameter estimates. Turning to the consumption model, the SSD is smallest (largest) between GME-NLP and OLS (3SLS) parameter estimates for both the three- and five-sigma rules (but only marginally). For example, the SSD between OLS (3SLS) and GME-NLP under the 3-sigma is 3.35 (4.15). Alternatively, for the labor model, the SSD is smallest (largest) between GME-NLP and 3SLS (OLS) parameter estimates for both the three- and five-sigma rules. The most dramatic differences arise in the investment model. For example, the SSD between OLS (3SLS) and GME-NLP under the 3-sigma is 3.00 (391.79). This comparison underscores divergences that exist between GME-NLP and 2SLS and 3SLS estimators. In addition to the information introduced by the

parameter support spaces, another reason for this divergence may be due to the fact that GME-NLP is a single-step estimator that is completely consistent with data constraints Equations (6) and (7), while 2SLS and 3SLS are multiple step estimators that only approximate the NLP model and ignore nonlinear interactions between reduced and structural form coefficients. The nonlinear specification of GME-NLP leads to first order optimization conditions (Equation (16) derived in the Appendix) that are different from other multiple-step or asymptotically justified estimators such as 2SLS and 3SLS. Overall, the SSD comparisons characterize finite samples differences in the GME-NLP estimator relative to more traditional estimators.

## 6. Conclusions

In this paper a one-step, data-constrained generalized maximum entropy estimator is proposed for the nonlinear- in- parameters model of the SEM (GME-NLP). Under the assumed regularity conditions, it is shown that the estimator is consistent and asymptotically normal in the presence of contemporaneously correlated errors. We define an asymptotically normal test (single scalar hypothesis) and an asymptotically chisquare-distributed Wald test (joint vector hypothesis) that are capable of performing hypothesis tests typically used in empirical work. Moreover, the GME-NLP estimator provides a simple method of introducing prior information into the model by means of informative supports on the parameters that can decrease the mean square error of the coefficient estimates. The reformulated GME-NLP model, which is optimized over the structural and reduced form parameter set, provides a computationally efficient approach for large and small sample sizes.

We evaluated the performance for the GME-NLP estimator based on a variety of Monte Carlo experiments and in an illustrative empirical application. In small and medium sample situations ( $\leq 100$  observations) the GME-NLP is mean square error superior to 3SLS for the defined experiments. Relative to 3SLS the GME-NLP estimator exhibited dramatically more robustness in the presence of contaminated error problems. These result illustrate advantages of a one-step, data-constrained estimator over multiple-step, moment-constrained estimators. Increasing the sample size clearly demonstrated consistency of the GME-NLP estimator for the SEM. The empirical results indicate that the GME-NLP did not dominate, nor was it dominated by, 3SLS in single or joint asymptotic hypothesis testing.

The three-equation Klein Model I was estimated as an empirical application of the GME-NLP method. Results of the Klein Model I benchmarked parameter estimates of GME-NLP relative to OLS, 2SLS, and 3SLS using the summed squared difference between parameter values of the estimators. GME-NLP was most similar to 2SLS and 3SLS for the consumption and labor demand equations, while it was most similar to OLS for the investment demand equation. In all, the empirical example also demonstrated some disadvantages of GME estimation in that coefficient estimates and predictive fit were somewhat sensitive to specification of error truncation points. This suggests additional research is needed to optimally identify error truncation points.

The analytical results in this study contribute toward establishing a rigorous foundation for GME estimation of the SEM and analogous properties of test statistics. It also furnishes a starting point for empirical economists desiring to apply maximum entropy to linear simultaneous systems (e.g., normalized quadratic demand systems used extensively in applied research). While empirical results are intriguing, this approach does not definitively solve the problem of estimating the SEM in

small samples or ill-posed problems, and underscores the need for continued research on problems of a number of problems in small sample estimation based on asymptotically justified estimators.

**Acknowledgments**

We thank George Judge (Berkeley) for helpful comments and suggestions. All errors remaining are the sole property of the authors.

**Conflicts of Interest**

The authors declare no conflict of interest.

**Appendix**

*A. Theorems and Proofs*

To facilitate both the derivation of the asymptotic properties and computational efficiency of the GME-NLP estimator, we reformulate the maximum entropy model into scalar notation that is completely consistent with Equations (5)–(8) (under the prevailing assumptions and the constraints Equations A1–A8 defined below). The scalar notation exhibits the flexibility to use different numbers of support points for each parameter or error term. However, we simplify the notation by using  $M$  support points for each parameter and error term.

Let  $\Delta$  represent a bounded, convex, and dense parameter space containing the  $(Q \times 1)$  vector of the reduced form and structural parameters  $\theta = \text{vec}(\theta^\pi, \theta^\gamma, \theta^\beta)$ . The reformulated constrained maximum entropy model is defined as

$$\max_{\theta, p^\pi, p^\gamma, p^\beta, z, w} = \left\{ \begin{aligned} & - \sum_{kgm} p_{kgm}^\pi \ln p_{kgm}^\pi - \sum_{igm} p_{igm}^\gamma \ln p_{igm}^\gamma - \sum_{kgm} p_{kgm}^\beta \ln p_{kgm}^\beta \\ & - \sum_{ngm} w_{ngm} \ln w_{ngm} - \sum_{ngm} z_{ngm} \ln z_{ngm} \end{aligned} \right\} \tag{A1}$$

subject to:

$$\sum_{m=1}^M s_{kgm}^\pi p_{kgm}^\pi = \theta_{kg}^\pi; \pi_{kgL} = s_{kg1}^\pi \leq \dots \leq s_{kgM}^\pi = \pi_{kgH} \tag{A2}$$

$$\sum_{m=1}^M s_{igm}^\gamma p_{igm}^\gamma = \theta_{ig}^\gamma; \gamma_{igL} = s_{ig1}^\gamma \leq \dots \leq s_{igM}^\gamma = \gamma_{igH} \tag{A3}$$

$$\sum_{m=1}^M s_{kgm}^\beta p_{kgm}^\beta = \theta_{kg}^\beta; \beta_{kgL} = s_{kg1}^\beta \leq \dots \leq s_{kgM}^\beta = \beta_{kgH} \tag{A4}$$

$$\sum_{m=1}^M s_{ngm}^w w_{ngm} = u_{ng} = y_{ng} - \mathbf{X}_{n\cdot} \left( \mathbf{\Pi}(\theta^\pi) \right)_{(-g)} \theta_g^\pi - (\mathbf{X}_{gn\cdot}) \theta_g^\beta; c_{g1} = s_{ng1}^w \leq \dots \leq s_{ngM}^w = c_{gM} \tag{A5}$$

$$\sum_{m=1}^M s_{ngm}^z z_{ngm} = v_{ng} = y_{ng} - (\mathbf{X}_{n\cdot}) \theta_g^\pi; c_{g1} = s_{ng1}^z \leq \dots \leq s_{ngM}^z = c_{gM} \tag{A6}$$

$$-S_{jgm}^i = -S_{jg(M+1-m)}^i \text{ for } m = 1, \dots, M \text{ (where for } M \text{ odd } S_{jg\frac{M+1}{2}}^i \equiv 0 \text{ and } i = w, z) \tag{A7}$$

$$\sum_{m=1}^M p_{kgm}^\pi = 1, \sum_{m=1}^M p_{igm}^\gamma = 1, \sum_{m=1}^M p_{kgm}^\beta = 1, \sum_{m=1}^M w_{ngm} = 1, \sum_{m=1}^M z_{ngm} = 1 \tag{A8}$$

Constraints A2-A6 define the reparameterized coefficients and errors with supports. In A5 the term  $\mathbf{\Pi}(\boldsymbol{\theta}^\pi)_{(-g)}$  is a  $(K \times G_g)$  matrix of elements  $\theta_{kg}^\pi$  that coincide with the endogenous variables in  $\mathbf{Y}_{(-g)}$ . The constraint A7 implies symmetry of the error supports about the origin and A8 defines the normalization conditions. The nonnegativity restrictions on  $p_{kgm}^\pi$ ,  $p_{igm}^\gamma$ ,  $p_{kgm}^\beta$ ,  $w_{ngm}$ , and  $z_{ngm}$  are inherently satisfied by the optimization problem and are not explicitly incorporated into the constraint set.

Next, we define the conditional entropy function by conditioning on  $\boldsymbol{\theta}^\pi = \boldsymbol{\tau}^\pi$ ,  $\boldsymbol{\theta}^\gamma = \boldsymbol{\tau}^\gamma$ , and  $\boldsymbol{\theta}^\beta = \boldsymbol{\tau}^\beta$ , or simply  $\boldsymbol{\theta} = \boldsymbol{\tau}$  where  $\boldsymbol{\theta} = \text{vec}(\boldsymbol{\theta}^\pi, \boldsymbol{\theta}^\gamma, \boldsymbol{\theta}^\beta)$  and  $\boldsymbol{\tau} = \text{vec}(\boldsymbol{\tau}^\pi, \boldsymbol{\tau}^\gamma, \boldsymbol{\tau}^\beta)$ . This yields

$$F(\boldsymbol{\tau}) = \max_{p^\pi, p^\gamma, p^\beta, z, w: \boldsymbol{\theta}=\boldsymbol{\tau}} \left\{ -\sum_{kgm} p_{kgm}^\pi \ln p_{kgm}^\pi - \sum_{igm} p_{igm}^\gamma \ln p_{igm}^\gamma - \sum_{kgm} p_{kgm}^\beta \ln p_{kgm}^\beta - \sum_{ngm} w_{ngm} \ln w_{ngm} - \sum_{ngm} z_{ngm} \ln z_{ngm} \right\} \tag{A9}$$

The optimal value of  $z_{ngm}$  in the conditionally-maximized entropy function is the solution to the Lagrangian  $L(z_{ng}, \eta_{ng}^z, \lambda_{ng}^z) = -\sum_{m=1}^M z_{ngm} \ln(z_{ngm}) + \eta_{ng}^z \left( \sum_{m=1}^M z_{ngm} - 1 \right) + \lambda_{ng}^z \left( \sum_{m=1}^M s_{ngm}^z z_{ngm} - v_{ng}(\boldsymbol{\tau}^\pi) \right)$  and is given by

$$z_{ngm} \left( \lambda_{ng}^z \left( v_{ng}(\boldsymbol{\tau}^\pi) \right) \right) = \frac{e^{\lambda_{ng}^z \left( v_{ng}(\boldsymbol{\tau}^\pi) \right) s_{ngm}^z}}{\sum_{\ell=1}^M e^{\lambda_{ng}^z \left( v_{ng}(\boldsymbol{\tau}^\pi) \right) s_{ng\ell}^z}}, m = 1, \dots, M \tag{A9}$$

while the optimal value  $w_{ngm}$ :

$$w_{ngm} \left( \lambda_{ng}^w \left( u_{ng}(\boldsymbol{\tau}) \right) \right) = \frac{e^{\lambda_{ng}^w \left( u_{ng}(\boldsymbol{\tau}) \right) s_{ngm}^w}}{\sum_{\ell=1}^M e^{\lambda_{ng}^w \left( u_{ng}(\boldsymbol{\tau}) \right) s_{ng\ell}^w}}, m = 1, \dots, M \tag{A10}$$

solves  $L(w_{ng}, \eta_{ng}^w, \lambda_{ng}^w) = -\sum_{m=1}^M w_{ngm} \ln(w_{ngm}) + \eta_{ng}^w \left( \sum_{m=1}^M w_{ngm} - 1 \right) + \lambda_{ng}^w \left( \sum_{m=1}^M s_{ngm}^w w_{ngm} - u_{ng}(\boldsymbol{\tau}) \right)$

The identities:

$$\sum_{m=1}^M s_{ngm}^z z_{ngm} \left( -\lambda_{ng}^z \left( v_{ng}(\boldsymbol{\tau}^\pi) \right) \right) = -\sum_{m=1}^M s_{ngm}^z z_{ngm} \left( \lambda_{ng}^z \left( v_{ng}(\boldsymbol{\tau}^\pi) \right) \right) \tag{A11}$$

and:

$$\sum_{m=1}^M s_{ngm}^w w_{ngm} \left( -\lambda_{ng}^w \left( u_{ng}(\boldsymbol{\tau}) \right) \right) = -\sum_{m=1}^M s_{ngm}^w w_{ngm} \left( \lambda_{ng}^w \left( u_{ng}(\boldsymbol{\tau}) \right) \right) \tag{A12}$$

follow from the symmetry of the support points around zero. Likewise the optimal values of  $p_{kgm}^\ell$  (for  $\ell = \pi, \gamma, \beta$ ) are respectively:

$$p_{kgm}^\ell \left( \lambda_{kg}^\ell \left( \tau_{kg}^\ell \right) \right) = \frac{e^{\lambda_{kg}^\ell \left( \tau_{kg}^\ell \right) s_{ngm}^\ell}}{\sum_{j=1}^M e^{\lambda_{kg}^\ell \left( \tau_{kg}^\ell \right) s_{ngj}^\ell}}, m = 1, \dots, M \tag{A13}$$

which satisfy  $L(p_{kg}^\ell, \eta_{ng}^\ell, \lambda_{ng}^\ell) = -\sum_{m=1}^M p_{kgm}^\ell \ln(p_{kgm}^\ell) + \eta_{kg}^\ell \left( \sum_{m=1}^M p_{kgm}^\ell - 1 \right) + \lambda_{kg}^\ell \left( \sum_{m=1}^M s_{ngm}^\ell p_{kgm}^\ell - \tau_{kg}^\ell \right)$ . For notational convenience we let  $\lambda_{ng}^z = \lambda^z(v_{ng}(\boldsymbol{\tau}^\pi))$ ,  $\lambda_{ng}^w = \lambda^w(u_{ng}(\boldsymbol{\tau}))$ ,  $\lambda_{kg}^\pi = \lambda_{kg}^\pi(\tau_{kg}^\pi)$ ,  $\lambda_{ig}^\gamma = \lambda_{ig}^\gamma(\tau_{ig}^\gamma)$  and  $\lambda_{kg}^\beta = \lambda_{kg}^\beta(\tau_{kg}^\beta)$  represent the optimal values of the Lagrangian multipliers. Substituting the solutions defined from Equations (A10), (A11), and (A14) into the conditional objective function yields the conditional maximum value entropy function:

$$F(\boldsymbol{\tau}) = -\sum_{kg} \left[ \lambda_{kg}^\pi \tau_{kg}^\pi - \ln \left( \sum_m \exp(\lambda_{kg}^\pi s_{kgm}^\pi) \right) \right] - \sum_{jg} \left[ \lambda_{jg}^\gamma \tau_{jg}^\gamma - \ln \left( \sum_m \exp(\lambda_{jg}^\gamma s_{jgm}^\gamma) \right) \right] \\ - \sum_{kg} \left[ \lambda_{kg}^\beta \tau_{kg}^\beta - \ln \left( \sum_m \exp(\lambda_{kg}^\beta s_{kgm}^\beta) \right) \right] - \sum_{ng} \left[ \lambda_{ng}^w u_{ng}(\boldsymbol{\tau}) - \ln \left( \sum_m \exp(\lambda_{ng}^w s_{ngm}^w) \right) \right] \\ - \sum_{ng} \left[ \lambda_{ng}^z v_{ng}(\boldsymbol{\tau}^\pi) - \ln \left( \sum_m \exp(\lambda_{ng}^z s_{ngm}^z) \right) \right] \tag{A14}$$

The gradient of  $F(\boldsymbol{\tau})$  in Equation (A16) is a  $(Q \times 1)$  vector  $\nabla(\boldsymbol{\tau}) = \text{vec}(\nabla_\pi(\boldsymbol{\tau}), \nabla_\gamma(\boldsymbol{\tau}), \nabla_\beta(\boldsymbol{\tau}))$  defined by:

$$\nabla(\boldsymbol{\tau}) = - \begin{pmatrix} \lambda^\pi(\boldsymbol{\tau}^\pi) \\ \lambda^\gamma(\boldsymbol{\tau}^\gamma) \\ \lambda^\beta(\boldsymbol{\tau}^\beta) \end{pmatrix} + \begin{pmatrix} (\mathbf{I}_G \otimes \mathbf{X}') & \left[ (\mathbf{I}_G + (\boldsymbol{\Gamma}(\boldsymbol{\tau}^\gamma)) \otimes \mathbf{X}') \right] \\ [\mathbf{0}] & \text{diag} \left( (\boldsymbol{\Pi}(\boldsymbol{\tau}^\pi)_{(-1)})' \mathbf{X}', \dots, (\boldsymbol{\Pi}(\boldsymbol{\tau}^\pi)_{(-G)})' \mathbf{X}' \right) \\ [\mathbf{0}] & \text{diag}(\mathbf{X}'_1, \dots, \mathbf{X}'_G) \end{pmatrix} \begin{pmatrix} \lambda^z \\ \lambda^w \end{pmatrix} \\ = -\boldsymbol{\lambda}(\boldsymbol{\tau}) + \mathbf{Z}^*(\boldsymbol{\tau}) \begin{pmatrix} \lambda^z \\ \lambda^w \end{pmatrix} \tag{A15}$$

Above,  $\boldsymbol{\Gamma}(\boldsymbol{\tau}^\gamma)$  is a  $(G \times G)$  matrix of elements  $\tau_{ig}^\gamma$  and  $\boldsymbol{\Pi}(\boldsymbol{\tau}^\pi)_{(-g)}$  is a  $(K \times G_g)$  matrix of elements  $\tau_{kg}^\pi$ . The Lagrangian multipliers are vertically concatenated into  $\boldsymbol{\lambda}^\pi, \boldsymbol{\lambda}^\gamma, \boldsymbol{\lambda}^\beta, \boldsymbol{\lambda}^w, \boldsymbol{\lambda}^z$ , where, for example, the vector  $\boldsymbol{\lambda}^w = \text{vec}(\boldsymbol{\lambda}_1^w, \dots, \boldsymbol{\lambda}_G^w)$  is of dimension  $(NG \times 1)$  and is made up of  $\boldsymbol{\lambda}_g^w = (\lambda_{1g}^w, \dots, \lambda_{Ng}^w)'$  for  $g = 1, \dots, G$ .

The  $(Q \times Q)$  Hessian matrix of the conditional maximum value  $F(\boldsymbol{\tau})$  in Equation (A15) is given by:

$$\mathbf{H}(\boldsymbol{\tau}) = -\frac{\partial \boldsymbol{\lambda}(\boldsymbol{\tau})}{\partial \boldsymbol{\tau}'} + \left( \frac{\partial \mathbf{Z}^*(\boldsymbol{\tau})}{\partial \boldsymbol{\tau}'} \right) (\mathbf{I} \otimes \boldsymbol{\lambda}^w) - \mathbf{Z}^*(\boldsymbol{\tau}) (\boldsymbol{\Xi}(\boldsymbol{\tau}) \odot \mathbf{Z}^*(\boldsymbol{\tau}))' \tag{A16}$$

where  $\odot$  denotes the Hadamard product (element wise) between two matrices. The  $(Q \times Q)$  diagonal matrix  $\frac{\partial \boldsymbol{\lambda}(\boldsymbol{\tau})}{\partial \boldsymbol{\tau}'} = \left( \frac{\partial \lambda^\pi}{\partial \tau'}, \frac{\partial \lambda^\gamma}{\partial \tau'}, \frac{\partial \lambda^\beta}{\partial \tau'} \right)$  is defined by:

$$\frac{\partial \lambda^\ell(\tau_{kg}^\ell)}{\partial \tau_{rt}^\ell} = \begin{cases} \left( \sum_{m=1}^M (s_{kgm}^\ell)^2 p_{kgm}^\ell - (\tau_{kg}^\ell)^2 \right)^{-1} & \text{if } k=r, g=t \\ 0 & \text{otherwise} \end{cases} \text{ for } \ell = \pi, \gamma, \beta$$

The components of the  $(Q \times QGN)$  matrix  $\frac{\partial \mathbf{Z}^*(\boldsymbol{\tau})}{\partial \boldsymbol{\tau}^r}$  are given by  $\frac{\partial \mathbf{Z}^*(\boldsymbol{\tau})}{\partial \boldsymbol{\tau}^r} = \left( \frac{\partial \mathbf{Z}^*(\boldsymbol{\tau})}{\partial \tau^{\pi r}}, \frac{\partial \mathbf{Z}^*(\boldsymbol{\tau})}{\partial \tau^{\gamma r}}, \frac{\partial \mathbf{Z}^*(\boldsymbol{\tau})}{\partial \tau^{\beta r}} \right)$ , where  $\frac{\partial \mathbf{Z}^*(\boldsymbol{\tau})}{\partial \tau^{\pi r}} = \left( \frac{\partial \mathbf{Z}^*(\boldsymbol{\tau})}{\partial \tau_{11}^\pi} \dots \frac{\partial \mathbf{Z}^*(\boldsymbol{\tau})}{\partial \tau_{KG}^\pi} \right)$  is a  $(KG \times KGN)$  sparse matrix of  $x_{nk} = s$ ,  $\frac{\partial \mathbf{Z}^*(\boldsymbol{\tau})}{\partial \tau^{\gamma r}} = \left( \frac{\partial \mathbf{Z}^*(\boldsymbol{\tau})}{\partial \tau_{11}^\gamma} \dots \frac{\partial \mathbf{Z}^*(\boldsymbol{\tau})}{\partial \tau_{GG}^\gamma} \right)$  is a  $(\bar{G} \times \bar{G}GN)$  matrix, and the  $(\bar{K} \times \bar{K}GN)$  matrix  $\frac{\partial \mathbf{Z}^*(\boldsymbol{\tau})}{\partial \tau^{\beta r}} = [\mathbf{0}]$ . Finally the matrix  $\Xi(\boldsymbol{\tau})$  is made up of derivatives of the Lagrangian multipliers  $\lambda^w$  and  $\lambda^z$ . It is defined as:

$$\begin{aligned} \Xi(\boldsymbol{\tau}) &= \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \otimes \text{diag}(\Xi_1^z(\boldsymbol{\tau}), \dots, \Xi_G^z(\boldsymbol{\tau})) + \begin{pmatrix} 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \otimes \begin{pmatrix} \Xi_1^w(\boldsymbol{\tau}) & \dots & \Xi_1^w(\boldsymbol{\tau}) \\ \vdots & \ddots & \vdots \\ \Xi_G^w(\boldsymbol{\tau}) & \dots & \Xi_G^w(\boldsymbol{\tau}) \end{pmatrix} \\ &+ \begin{pmatrix} 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \otimes \text{diag}(\Xi_1^w(\boldsymbol{\tau}), \dots, \Xi_G^w(\boldsymbol{\tau})) + \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix} \otimes \text{diag}(\Xi_1^w(\boldsymbol{\tau}), \dots, \Xi_G^w(\boldsymbol{\tau})) \end{aligned}$$

with:

$$\Xi_g^i(\boldsymbol{\tau}) = \begin{pmatrix} \xi_{1g}^i(\boldsymbol{\tau}) & \dots & \xi_{1g}^i(\boldsymbol{\tau}) \\ \vdots & \ddots & \vdots \\ \xi_{Ng}^i(\boldsymbol{\tau}) & \dots & \xi_{Ng}^i(\boldsymbol{\tau}) \end{pmatrix}; g=1, \dots, G; i=z, w$$

where:

$$\xi_{ng}^w(\boldsymbol{\tau}) = \left( \sum_{m=1}^M (s_{ngm}^w)^2 w_{ngm}(\lambda^w(u_{ng}(\boldsymbol{\tau}))) - (u_{ng}(\boldsymbol{\tau}))^2 \right)^{-1}$$

and:

$$\xi_{ng}^z(\boldsymbol{\tau}^\pi) = \left( \sum_{m=1}^M (s_{ngm}^z)^2 z_{ngm}(\lambda^z(v_{ng}(\boldsymbol{\tau}^\pi))) - (v_{ng}(\boldsymbol{\tau}^\pi))^2 \right)^{-1}$$

By the Cauchy-Swcharz inequality, symmetry assumption on the supports, and the adding up conditions, then  $-\frac{\partial \lambda(\boldsymbol{\tau})}{\partial \boldsymbol{\tau}} - \mathbf{Z}^*(\boldsymbol{\tau})(\Xi(\boldsymbol{\tau}) \odot \mathbf{Z}^*(\boldsymbol{\tau}))'$  is a negative definite matrix. Next, we prove consistency and asymptotic normality of the GME-NLP estimator.

**Theorem 1.** Under the regularity conditions R1-R5, the GME-NLP estimator,  $\hat{\boldsymbol{\theta}} = \text{vec}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\delta}})$ , is a consistent estimator of the true coefficient values  $\boldsymbol{\theta} = \text{vec}(\boldsymbol{\pi}, \boldsymbol{\delta})$ .

*Proof.* Let  $\Delta$  represent a bounded, convex, and dense parameter space such that the true coefficient values  $\boldsymbol{\theta} \in \Delta$ . Consider the just identified case. From Equations (5)–(8):

$$\max_{\mathbf{p}^\pi, \mathbf{p}^\gamma, \mathbf{p}^\beta, \mathbf{w}, \mathbf{z}} \{-\mathbf{w}' \ln \mathbf{w}\}$$

is not a function of  $\mathbf{p}^\pi$  or  $\mathbf{z}$  almost everywhere. Furthermore, it is not a function of the reduced form coefficients satisfying the identification conditions that are discussed after Equation (4). In addition the



nonstochastic terms  $-\mathbf{p}^{\pi'} \ln \mathbf{p}^{\pi}$ ,  $-\mathbf{p}^{\gamma'} \ln \mathbf{p}^{\gamma}$ , and  $-\mathbf{p}^{\beta'} \ln \mathbf{p}^{\beta}$  are asymptotically irrelevant terms that vanish in the convergence of the scaled Hessian or  $\frac{1}{N} \mathbf{H}$ . Accordingly the GME-NLP estimates of the reduced form parameters,  $\hat{\boldsymbol{\pi}}$ , are asymptotically and uniquely determined by:

$$\hat{\boldsymbol{\pi}} = \arg \max_{\boldsymbol{\tau}^{\pi}} \left\{ -\mathbf{z}(\boldsymbol{\tau}^{\pi})' \ln \mathbf{z}(\boldsymbol{\tau}^{\pi}) \right\}$$

subject to Equation (7) and a normalization condition in Equation (8). The  $\hat{\boldsymbol{\pi}}$  are consistent, or  $\hat{\boldsymbol{\pi}} \xrightarrow{p} \boldsymbol{\pi}$ , which is proved in the **Proposition** below.

Next define the conditional estimator

$$\hat{\boldsymbol{\delta}}(\boldsymbol{\tau}^{\pi}) = \left( \hat{\boldsymbol{\gamma}}(\boldsymbol{\tau}^{\pi}), \hat{\boldsymbol{\beta}}(\boldsymbol{\tau}^{\pi}) \right) = \arg \max_{\boldsymbol{\tau}^{\gamma}, \boldsymbol{\tau}^{\beta}; \boldsymbol{\tau}^{\pi}} F(\boldsymbol{\tau})$$

for  $\boldsymbol{\tau}^{\pi}$  in the parameter set that satisfies the identification conditions. By [32]:

$$\hat{\boldsymbol{\gamma}}(\boldsymbol{\tau}^{\pi}) \xrightarrow{p} \boldsymbol{\gamma}(\boldsymbol{\tau}^{\pi}) \text{ and } \hat{\boldsymbol{\beta}}(\boldsymbol{\tau}^{\pi}) \xrightarrow{p} \boldsymbol{\beta}(\boldsymbol{\tau}^{\pi})$$

and:

$$\hat{\boldsymbol{\gamma}}(\boldsymbol{\pi}) \xrightarrow{p} \boldsymbol{\gamma}(\boldsymbol{\pi}) \text{ and } \hat{\boldsymbol{\beta}}(\boldsymbol{\pi}) \xrightarrow{p} \boldsymbol{\beta}(\boldsymbol{\pi})$$

then by [39]

$$\left( \hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\gamma}}(\hat{\boldsymbol{\pi}}), \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\pi}}) \right) \xrightarrow{p} (\boldsymbol{\pi}, \boldsymbol{\gamma}, \boldsymbol{\beta})$$

which establishes consistency for the just identified case. Further results pertaining to the overidentified case are available from the authors upon request.

**Theorem 2.** Under the conditions of Theorem 1, the GME-NLP estimator,  $\hat{\boldsymbol{\delta}} = \text{vec}(\hat{\boldsymbol{\delta}}_1, \dots, \hat{\boldsymbol{\delta}}_G)$ , is asymptotically normally distributed as  $\hat{\boldsymbol{\delta}} \overset{a}{\sim} N\left(\boldsymbol{\delta}, \frac{1}{N} \boldsymbol{\Omega}_{\xi}^{-1} \boldsymbol{\Omega}_{\Sigma} \boldsymbol{\Omega}_{\xi}^{-1}\right)$ .

*Proof.* Let  $\hat{\boldsymbol{\delta}}$  be the GME-NLP estimator of  $\boldsymbol{\delta} = \text{vec}(\boldsymbol{\delta}_1, \dots, \boldsymbol{\delta}_G)$ . Expand the gradient vector in a Taylor series around  $\boldsymbol{\delta}$  to obtain:

$$\nabla(\hat{\boldsymbol{\delta}}) = \nabla(\boldsymbol{\delta}) + \mathbf{H}(\boldsymbol{\delta}^*)(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}) \tag{A17}$$

where  $\boldsymbol{\delta}^*$  is between  $\hat{\boldsymbol{\delta}}$  and  $\boldsymbol{\delta}$ . Since  $\hat{\boldsymbol{\delta}}$  is a consistent estimator of  $\boldsymbol{\delta}$ , then  $\boldsymbol{\delta}^* \xrightarrow{p} \boldsymbol{\delta}$ . Using this information and the fact that  $\nabla(\hat{\boldsymbol{\delta}}) = [\mathbf{0}]$  at the optimum, then:

$$\sqrt{N}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}) \overset{d}{=} \left(\frac{1}{N} \mathbf{H}(\boldsymbol{\delta})\right)^{-1} \left(\frac{1}{\sqrt{N}} \nabla(\boldsymbol{\delta})\right)$$

where both the left hand and right hand side terms have equivalent limiting distributions. Note that  $\frac{1}{N} \mathbf{H}(\boldsymbol{\delta}) = \frac{1}{N} \mathbf{Z}'(\boldsymbol{\Xi}^w \odot \mathbf{Z}) + 0_p \left(\frac{1}{N}\right)$  where  $\mathbf{Z}$  is the block diagonal matrix  $\mathbf{Z} = \text{diag}(\mathbf{Z}_1, \dots, \mathbf{Z}_G)$ . From regulatory conditions  $\frac{1}{N} \mathbf{H}(\boldsymbol{\delta}) \xrightarrow{p} \boldsymbol{\Omega}_{\xi} = \text{diag}(\xi_1 \boldsymbol{\Psi}_1, \dots, \xi_G \boldsymbol{\Psi}_G)$ , where  $\boldsymbol{\Omega}_{\xi}$  is a positive definite matrix. Because  $\xi_{ng} = \xi_{ng}^w(\boldsymbol{\theta}) = \xi_{ng}^z(\boldsymbol{\theta})$  are iid for  $n = 1, \dots, N$ , then  $\xi_g = E[\xi_{ng}]$ .

The scaled gradient term is asymptotically normally distributed as:

$$\frac{1}{\sqrt{N}} \nabla(\boldsymbol{\delta}) \xrightarrow{d} N([\mathbf{0}], \boldsymbol{\Omega}_\Sigma)$$

with covariance matrix  $\frac{1}{N} \mathbf{Z}'(\boldsymbol{\Sigma}_\lambda \otimes \mathbf{I}) \mathbf{Z} \rightarrow \boldsymbol{\Omega}_\Sigma$ , where  $\boldsymbol{\Sigma}_\lambda$  a  $(G \times G)$  covariance matrix for the  $\lambda_{\cdot g}^w$ 's (see [40,41]). From the above results and by applying Slutsky's Theorem:

$$\sqrt{N}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}) \overset{a}{\sim} N([\mathbf{0}], \boldsymbol{\Omega}_\xi^{-1} \boldsymbol{\Omega}_\Sigma \boldsymbol{\Omega}_\xi^{-1})$$

which yields the asymptotic distribution:

$$\hat{\boldsymbol{\delta}} \overset{a}{\sim} N\left(\boldsymbol{\delta}, \frac{1}{N} \boldsymbol{\Omega}_\xi^{-1} \boldsymbol{\Omega}_\Sigma \boldsymbol{\Omega}_\xi^{-1}\right)$$

**Proposition 1.** Under the assumptions of Theorem 1, the reduced form estimates of (3) are consistent,

$$\hat{\boldsymbol{\pi}} = \arg \max_{\boldsymbol{\tau}^\pi} \left\{ -\mathbf{z}(\boldsymbol{\tau}^\pi)' \ln \mathbf{z}(\boldsymbol{\tau}^\pi) \right\} \xrightarrow{p} \boldsymbol{\pi}.$$

*Proof.* With the exception that we account for contemporaneous correlation in the errors, this is the proof for consistency of the data-constrained GME estimator of the general linear model [32]. Consider the conditional maximum function:

$$F_R(\boldsymbol{\tau}^\pi) = -\sum_{ng} \left[ \lambda_{ng}^z v_{ng}(\boldsymbol{\tau}^\pi) - \ln \left( \sum_m \exp(\lambda_{ng}^z s_{ng}^z) \right) \right]$$

where  $v_{ng} = y_{ng} - \mathbf{X}_n \boldsymbol{\pi}_g$ .

We expand  $F_R(\boldsymbol{\tau}^\pi)$  about  $\boldsymbol{\pi}$  with a Taylor series expansion that yields:

$$F_R(\boldsymbol{\tau}^\pi) = F_R(\boldsymbol{\pi}) + \nabla(\boldsymbol{\pi})' (\boldsymbol{\tau}^\pi - \boldsymbol{\pi}) + \frac{1}{2} (\boldsymbol{\tau}^\pi - \boldsymbol{\pi})' \mathbf{H}_R(\boldsymbol{\pi}^*) (\boldsymbol{\tau}^\pi - \boldsymbol{\pi})$$

where  $\boldsymbol{\pi}^*$  lies between  $\boldsymbol{\tau}^\pi$  and  $\boldsymbol{\pi}$ . The gradient vector is given by  $\nabla_\pi = (\mathbf{I} \otimes \mathbf{X}') \boldsymbol{\lambda}^z$  and the Hessian matrix is  $\mathbf{H}_R = (\mathbf{I} \otimes \mathbf{X}') (\boldsymbol{\Xi}^z \odot (\mathbf{I} \otimes \mathbf{X}'))$ . The scaled gradient term is asymptotically normally distributed as  $\frac{1}{\sqrt{N}} \nabla_\pi(\boldsymbol{\pi}) \xrightarrow{d} N([\mathbf{0}], \boldsymbol{\Omega}_R)$  by a multivariate version of Liapounov's central limit theorem (see [40,41]). The covariance matrix is  $\frac{1}{N} \mathbf{Z}'_R (\boldsymbol{\Sigma}_\lambda \otimes \mathbf{I}) \mathbf{Z}_R \rightarrow \boldsymbol{\Omega}_R$  where  $\mathbf{Z}_R = (\mathbf{I} \otimes \mathbf{X})$  and  $\boldsymbol{\Sigma}_\lambda$  is a  $(G \times G)$  covariance matrix of the  $\lambda_{\cdot g}^z$ 's. Hence the gradient is bounded in probability. The value of the quadratic term in the Taylor expansion can be bounded above by:

$$\frac{1}{2} (\boldsymbol{\tau}^\pi - \boldsymbol{\pi})' \mathbf{H}_R(\boldsymbol{\pi}^*) (\boldsymbol{\tau}^\pi - \boldsymbol{\pi}) \leq \phi_s \frac{N \|\boldsymbol{\tau}^\pi - \boldsymbol{\pi}\|^2}{2}$$

The parameter  $\phi_s$  denotes the smallest eigenvalue of  $-\frac{1}{N} \mathbf{H}_R(\boldsymbol{\pi}^*)$  for any  $\boldsymbol{\pi}^*$  that lies between  $\boldsymbol{\tau}^\pi$  and  $\boldsymbol{\pi}$ , where  $\|a\| = \left[ \sum_{k=1}^K a_k^2 \right]^{1/2}$  denotes the standard vector norm.

Combining the elements from above, for all  $\varepsilon > 0$  the  $P\left(\max_{\tau:|\tau^\pi-\pi|>\varepsilon}\{F(\tau)-F(\delta)<0\}\right)\rightarrow 1$  as  $N\rightarrow\infty$ .

Thus,  $\hat{\pi} = \arg \max_{\tau^\pi} \left\{ -\mathbf{z}(\tau^\pi)' \ln \mathbf{z}(\tau^\pi) \right\} \xrightarrow{P} \pi$ .

### B. Model Estimation: Computational Considerations

To estimate the GME-NLP model, the conditional entropy function (Equation (A15)) was maximized. Note that the constrained maximization problem Equations (5)–(8) requires estimation of  $(Q + 2GNM)$  unknown parameters. Solving Equations (5)–(8) for  $(Q + 2GNM)$  unknowns is not computationally practical as the sample size,  $N$ , grows larger. For example, consider an empirical application with  $Q = 36$  coefficients,  $G = 3$  equations, and  $M = 3$  support points. Even for a small number of observations, say  $N = 50$ , the number of unknown parameters would be 936. In contrast, maximizing Equation (A15) requires estimation of only  $Q$  unknown coefficients for any real value of  $N$ .

The GME-NLP estimator uses the reduced and structural form models as data constraints with a dual objective function as part of its information set. To completely specify the GME-NLP model, support (upper and lower truncation and intermediate) points for the individual parameters, support points for each error term, and  $Q$  starting values for the parameter coefficients are supplied by the user. In the Monte Carlo analysis and empirical application, the model was estimated using the unconstrained optimizer OPTIMUM in the econometric software GAUSS. We used 3 support points for each parameter and error term. To increase the efficiency of the estimation process the analytical gradient and Hessian were coded in GAUSS and called in the optimization routine. This also offered an opportunity to empirically validate the derivation of the gradient, Hessian, and covariance matrix. Given suitable starting values the optimization routine generally converged within seconds for the empirical examples discussed above. Moreover, solutions were quite robust to alternative starting values.

### References

1. Theil, H. *Principles of Econometrics*; John Wiley & Sons: New York, NY, USA, 1971.
2. Zellner, A.; Theil, H. Three-stage least squares: Simultaneous estimation of simultaneous equations. *Econometrica* **1962**, *30*, 54–78.
3. Fuller, W.A. Some properties of a modification of the limited information estimator. *Econometrica* **1977**, *45*, 939–953.
4. Koopmans, T.C. *Statistical Inference in Dynamic Economic Models*; Cowles Commission Monograph 10; Wiley: New York, NY, USA, 1950.
5. Hausman, J.A. Full information instrumental variable estimation of simultaneous equations systems. *Ann. Econ. Soc. Meas.* **1974**, *3*, 641–652.
6. Zellner, A. Statistical analysis of econometric models. *J. Am. Stat. Assoc.* **1976**, *74*, 628–643.
7. Zellner, A. The finite sample properties of simultaneous equations = estimates and estimators bayesian and non-bayesian approaches. *J. Econom.* **1998**, *83*, 185–212.
8. Phillips, P.C.B. Exact Small Sample Theory in the Simultaneous Equations Model. In *Handbook of Econometrics*; Griliches, Z., Intrilligator, M.D., Eds.; Elsevier: New York, NY, USA, 1983.

9. Golan, A.; Judge, G.; Miller, D. *Maximum Entropy Econometrics: Robust Estimation with Limited Data*; John Wiley & Sons: New York, NY, USA, 1996.
10. Golan, A.; Judge, G.; Miller, D. Information Recovery in Simultaneous Equations Statistical Models. In *Handbook of Applied Economic Statistics*; Ullah, A., Giles, D., Eds.; Marcel Dekker: New York, NY, USA, 1997.
11. West, K.D.; Wilcox, D.W. A comparison of alternative instrumental variables estimators of a dynamic linear model. *J. Bus. Econ. Stat.* **1996**, *14*, 281–293.
12. Hanson, L.P.; Heaton, J.; Yaron, A. Finite-sample properties of some alternative GMM estimators. *J. Bus. Econ. Stat.* **1996**, *14*, 262–280.
13. Tukey, J.W. A Survey Sampling from Contaminated Distributions. In *Contributions to Probability and Statistics*; Olkin, I., Ed.; Stanford University Press: Stanford, CA, USA, 1960.
14. Huber, P.J. *Robust Statistics*; John Wiley & Sons: New York, NY, USA, 1981.
15. Hampel, F.R.; Ronchetti, E.M.; Rousseeuw, P.J.; Stahel, W.A. *Robust Statistics: The Approach Based on Influence Functions*; John Wiley & Sons: New York, NY, USA, 1986.
16. Koenker, R.; Machado, J.A.F.; Keels, C.L.S.; Welsh, A.H. Momentary lapses: Moment expansions and the robustness of minimum distance estimation. *Econom. Theory* **1994**, *10*, 172–197.
17. Kitamura, Y.; Stutzer, M. An information-theoretic alternative to generalized method of moments estimation. *Econometrica* **1997**, *65*, 861–874.
18. Imbens, G.; Spady, R.; Johnson, P. Information theoretic approaches to inference in moment condition models. *Econometrica* **1998**, *66*, 333–357.
19. Van Akkeren, M.; Judge, G.G.; Mittelhammer, R.C. Generalized moment based estimation and inference. *J. Econom.* **2002**, *107*, 127–148.
20. Mittelhammer, R.; Judge, G. Endogeneity and Moment Based Estimation under Squared Error Loss. In *Handbook of Applied Econometrics and Statistics*; Wan, A., Ullah, A., Chaturvedi, A., Eds.; Marcel Dekker: New York, NY, USA, 2001.
21. Mittelhammer, R.C.; Judge, G.; Miller, D. *Econometric Foundations*; Cambridge University Press: New York, NY, USA, 2000.
22. Marsh, T.L.; Mittelhammer, R.C. Generalized Maximum Entropy Estimation of a First Order Spatial Autoregressive Model. In *Advances in Econometrics, Spatial and Spatiotemporal Econometrics*; LeSage, J.P., Ed; Elsevier: New York, USA, 2004; Volume 18.
23. Ciavolino, E.; Dahlgaard, J.J. Simultaneous Equation Model based on the generalized maximum entropy for studying the effect of management factors on enterprise performance. *J. Appl. Stat.* **2009**, *36*, 801–815.
24. Papalia, R.B.; Ciavolino, E. GME estimation of spatial structural equations models. *J. Classif.* **2011**, *28*, 126–141.
25. Zellner, A. Estimation of regression relationships containing unobservable independent variables. *Int. Econ. Rev.* **1970**, *11*, 441–454.
26. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.
27. Kullback, S. *Information Theory and Statistics*; John Wiley & Sons: New York, NY, USA, 1959.
28. Pompe, B. On some entropy measures in data analysis. *Chaos Solitons Fractals* **1994**, *4*, 83–96.

29. Zellner, A. Bayesian and Non-Bayesian Estimation Using Balanced Loss Functions. In *Statistical Decision Theory and Related Topics*; Gupta, S., Berger, J., Eds.; Springer Verlag: New York, NY, USA, 1994.
30. Dreze, J.H.; Richard, J.F. Bayesian Analysis of Simultaneous Equations Systems. In *Handbook of Econometrics*; Griliches, Z., Intriligator, M.D., Eds.; Elsevier: New York, NY, USA, 1983.
31. Malinvaud, E. *Statistical Methods of Econometrics*, 3rd Ed; North-Holland: Amsterdam, The Netherlands, 1980.
32. Mittelhammer, R.C.; Cardell, N.S.; Marsh, T.L. The data-constrained generalized maximum entropy estimator of the GLM: Asymptotic theory and inference. *Entropy* **2013**, *15*, 1756–1775.
33. Pukelsheim, F. The three sigma rule. *Am. Stat.* **1994**, *48*, 88–91.
34. Davidson, R.; MacKinnon, J.G. *Estimation and Inference in Econometrics*; Oxford: New York, NY, USA, 1993.
35. Mittelhammer, R.C. *Mathematical Statistics for Economics and Business*; Springer: New York, NY, USA, 1996.
36. Cragg, J.G. On the relative small sample properties of several structural-equation estimators. *Econometrica* **1967**, *35*, 89–110.
37. Tsurumi, H. Comparing Bayesian and Non-Bayesian Limited Information Estimators. In *Bayesian and Likelihood Methods in Statistics and Econometrics*; Geisser, S., Hodges, J.S., Press, S.J., Zellner, A., Eds.; North Holland Publishing: Amsterdam, The Netherlands, 1990.
38. Klein, L.R. *Economic Fluctuations in the United States, 1921–1941*; John Wiley & Sons: New York, NY, USA, 1950.
39. Rao, C.R. *Linear Statistical Inference and Its Applications*, 2nd ed; John Wiley & Sons: New York, NY, USA, 1973.
40. Hoadley, B. Asymptotic properties of maximum likelihood estimators for the independent but not identically distributed case. *Ann. Math. Stat.* **1971**, *42*, 1977–1991.
41. White, H. *Asymptotic Theory for Econometricians*; Academic Press: New York, NY, USA 1984.