Asymptotic Bias and Equivalence of GMM and GEL Estimators*

Whitney K. Newey
Department of Economics
M.I.T.

Richard J. Smith
Department of Economics
University of Bristol

First Draft: January 2000
This Revision: July 2000
Discussion Paper No. 01/517

Abstract

A number of recent studies have concluded that the most commonly used efficient two-step generalized method of moments (GMM) estimator may have large bias in applications. This problem has motivated the search for alternative efficient estimators with smaller bias. Estimators which have received some attention in the literature are the continuous updating estimator (CUE), the empirical likelihood estimator, and the exponential tilting estimator. We show that the continuous updating estimator is a member of a generalized empirical likelihood (GEL) class that also includes exponential tilting and empirical likelihood. Also, we derive stochastic expansions for each of the estimators along with asymptotic bias expressions. We find that an important part of the bias of GEL estimators will tend to be less than that of GMM with many moment conditions, and that the bias

*This research was initiated when the first author was Benjamin Meaker Visiting Professor at the University of Bristol. Helpful comments were provided by G. Imbens and seminar participants at Bristol. The NSF provided partial financial support for the first author and the research of the second author was supported by the ESRC under Grant No. R000237386.

of empirical likelihood is the same as a GMM estimator where the optimal linear combination coefficients are known.

JEL Classification: C13, C30
Keywords: GMM, Continuous Updating, Empirical Likelihood, Bias, Stochastic Expansions.

1 Introduction

A number of recent studies have concluded that the most commonly used efficient two-step generalized method of moments (GMM; Hansen, 1982) estimator may have large biases for the sample sizes typically encountered in applications. See, for example, the Special Section, July 1996, of the *Journal of Business and Economic Statistics.* This problem has motivated the search for alternative efficient estimators with smaller bias. Hansen, Heaton, and Yaron (1996) suggested the continuous updating estimator (CUE) and showed that it had smaller bias than multi-step GMM estimators in some Monte Carlo examples. Stock and Wright (2000) show that the CUE has better properties than the two-step GMM estimator under weak instrument asymptotic theory. Donald and Newey (2000) give a jackknife interpretation of the CUE that explains its small bias. Other estimators include the empirical likelihood (EL) estimator of Imbens (1997) and Qin and Lawless (1994), and the exponential tilting (ET) estimator of Imbens, Spady and Johnson (1998) and Kitamura and Stutzer (1997). The EL and ET estimators are included in the class of generalized empirical likelihood (GEL) estimators considered in Smith (1997, 1999), that use an objective function like that of Brown, Newey, and May (1997).

All of these estimators are asymptotically normal and mutually asymptotically equivalent. Thus, some alternative to the usual first order asymptotic theory must be used to compare their properties. We give one exact result by showing that the CUE is included in the class of GEL estimators. We also consider stochastic expansions like those of Nagar (1959) and Robinson

(1988). As a by-product we also give precise consistency and asymptotic normality results for any GEL estimator.

We show higher-order asymptotic equivalence of the CUE and all GEL estimators under zero expectation of third powers of the moment indicators. We also derive asymptotic bias expressions for each of the estimators, obtained as the expectation of leading terms in stochastic expansions. We show that each GEL estimator removes the bias of GMM that is associated with estimation of the Jacobian term that appears in the optimal linear combination of moments. This bias can be especially large when there are many moment conditions. This result provides an asymptotic justification for the relatively small bias for the CUE found by Hansen, Heaton, and Yaron (1996). We also find that the moment vector will tend to be biased when the estimator is biased. This result predicts that bias of the coefficients tends to be associated with excessive rejection for the test of over identifying restrictions.

We also show that empirical likelihood removes the bias due to estimation of the weight matrix in the optimal linear combination of moments, even when third moments are nonzero. Strikingly, we find that for empirical likelihood the bias is the same as for an (unfeasible) estimator based on the optimal linear combination of moment restrictions.

Section 2 describes the GMM, CUE, and GEL estimators and demonstrates that the CUE is a GEL estimator. Section 3 reports the stochastic expansions and asymptotic biases. Section 4 gives regularity conditions and a precise consistency and asymptotic normality result for GEL estimators. The Appendix contains some intermediate results and proofs.

2 The Estimators

To describe the estimators, let z_i, $(i = 1, ..., n)$, be i.i.d. observations on a data vector z. Consider the moment indicator $g(z, \beta)$, an m-vector of functions of the data observation z and the p-vector β of unknown parameters which are the object of inferential interest, where $m \geq p$. It is assumed that the true parameter vector β_0 satisfies the moment condition

$$E[g(z, \beta_0)] = 0,$$

where $E[.]$ denotes expectation taken with respect to the distribution of z. Let $g_i(\beta) \equiv g(z_i, \beta)$, $\hat{g}(\beta) \equiv n^{-1}\sum_{i=1}^{n} g_i(\beta)$, and $\hat{\Omega}(\beta) \equiv n^{-1}\sum_{i=1}^{n} g_i(\beta)g_i(\beta)'$. A two-step GMM estimator is one that satisfies, for some preliminary consistent estimator $\tilde{\beta}$ for β_0,

$$\hat{\beta}_{2S} = \arg\min_{\beta\in\mathcal{B}} \hat{g}(\beta)'\hat{\Omega}(\tilde{\beta})^{-1}\hat{g}(\beta), \quad (2.1)$$

where $\mathcal{B}$ denotes the parameter space. The CUE is obtained by simultaneously minimizing over β in $\hat{\Omega}(\beta)^{-1}$, that is

$$\hat{\beta}_{CUE} = \arg\min_{\beta\varepsilon\mathcal{B}} \hat{g}(\beta)'\hat{\Omega}(\beta)^{-}\hat{g}(\beta), \quad (2.2)$$

where A^- denotes a generalized inverse of a matrix A, satisfying $AA^-A = A$. As we show below, the objective function is invariant to the choice of generalized inverse.

GEL estimators are obtained as the solution to a saddle point problem. Let $\rho(.)$ be a function that is concave its domain, which is an open interval containing zero, λ be a m-vector of auxiliary parameters, and

$$\hat{P}(\beta, \lambda) = n^{-1}\sum_{i=1}^{n} \rho(\lambda' g_i(\beta)). \quad (2.3)$$

Also, let Γ_n be a set such that $\lambda' g_i(\beta)$ is in the domain of $\rho(.)$ for all $\lambda \in \Gamma_n$, $\beta \in \mathcal{B}$, and $i \leq n$. As specified in Section 4, it will suffice for the theory here that Γ_n place bounds on λ that shrink with n slower than $n^{-1/2}$. A GEL estimator is obtained as the solution

$$\hat{\beta}_{GEL} = \arg\min_{\beta \in \mathcal{B}} \sup_{\lambda \in \Gamma_n} \hat{P}(\beta, \lambda). \tag{2.4}$$

The EL estimator is a GEL estimator with $\rho(v) = \ln(1+v)$, see Imbens (1997) and Qin and Lawless (1994), and ET is also GEL with $\rho(v) = -\exp(v)$, see Imbens, Spady, and Johnson (1998) and Kitamura and Stutzer (1997).

Our first result shows that the CUE is also a GEL estimator. Let $\rho_v(v)$ and $\rho_{vv}(v)$ denote first and second derivatives of $\rho(v)$, respectively.

Theorem 1 *If $\rho(v)$ is quadratic, $\rho_v(0) \neq 0$ and $\rho_{vv}(0) < 0$, then $\hat{\beta}_{GEL} = \hat{\beta}_{CUE}$.*

GEL estimators are also related to estimators considered by Corcoran (1998). To describe these estimators, let $h(\pi)$ be a convex function of a scalar π that measures the discrepancy between π and the empirical probability $1/n$ of a single observation, that can depend on n. Consider the optimization problem

$$\min_{\pi_1,\ldots,\pi_n,\beta} \sum_{i=1}^{n} h(\pi_i), s.t. \sum_{i=1}^{n} \pi_i g_i(\beta) = 0, \sum_{i=1}^{n} \pi_i = 1. \tag{2.5}$$

When the solutions $\hat{\pi}_1, ..., \hat{\pi}_n$ of this problem are nonnegative, these can be interpreted as probabilities that minimize the discrepancy with the empirical measure subject to the moment conditions. We refer to the solution $\hat{\beta}_{MD}$ to this minimization problem as a minimum discrepancy (MD) estimator.

We can relate MD and GEL estimators by comparing their first-order conditions. For an m-vector of Lagrange multipliers $\hat{\alpha}_{MD}$ associated with the first constraint and a scalar $\hat{\mu}_{MD}$ for the second in (2.5), the MD first order conditions for $\hat{\pi}_i$ are $h_\pi(\hat{\pi}_i) = -\hat{\alpha}'_{MD} g_i(\hat{\beta}_{MD}) + \hat{\mu}_{MD}$. Let $G_i(\beta) \equiv \partial g_i(\beta)/\partial \beta'$. Assuming $h_\pi(.)$ is one-to-one, solving for $\hat{\pi}_i$ and substituting into the first-order conditions for $\hat{\beta}_{MD}$, $\hat{\alpha}_{MD}$ and $\hat{\mu}_{MD}$ gives, for MD,

$$\begin{aligned} \sum_{i=1}^{n} h_\pi^{-1}(-\hat{\alpha}'_{MD} g_i(\hat{\beta}_{MD}) + \hat{\mu}_{MD}) G_i(\hat{\beta}_{MD})' \hat{\alpha}_{MD} &= 0, \\ \sum_{i=1}^{n} h_\pi^{-1}(-\hat{\alpha}'_{MD} g_i(\hat{\beta}_{MD}) + \hat{\mu}_{MD}) g_i(\hat{\beta}_{MD}) &= 0, \end{aligned} \tag{2.6}$$

and $\sum_{i=1}^{n} h_\pi^{-1}(-\hat{\alpha}'_{MD} g_i(\hat{\beta}_{MD}) + \hat{\mu}_{MD}) = 1$. For comparison, the GEL first-order conditions are

$$\begin{aligned} \sum_{i=1}^{n} \rho_v(\hat{\lambda}'_{GEL} g_i(\hat{\beta}_{GEL})) G_i(\hat{\beta}_{GEL})' \hat{\lambda}_{GEL} &= 0, \\ \sum_{i=1}^{n} \rho_v(\hat{\lambda}'_{GEL} g_i(\hat{\beta}_{GEL})) g_i(\hat{\beta}_{GEL}) &= 0. \end{aligned} \tag{2.7}$$

In general, the first order conditions for GEL and MD are different, and hence so are the estimators of β.

There is a relationship between ML and GEL when $h(\pi)$ is a member of the Cressie-Read (1984) power family, where $h(\pi) = \pi^{\gamma+1}/[\gamma(\gamma+1)]$. In this case $h_\pi^{-1}(.)$ is homogenous, so that $\hat{\mu}_{MD}$ can be factored out of (2.6). Then the first-order conditions eq. (2.6) and eq. (2.7) coincide for

$$\rho(v) = -\gamma^2(1+v)^{(\gamma+1)/\gamma}/(\gamma+1), \hat{\lambda}_{GEL} = \hat{\alpha}_{MD}/\hat{\mu}_{MD}. \tag{2.8}$$

In this case the GEL saddle point problem is a dual of the MD one, in the sense that $\hat{\lambda}_{GEL}$ is a ratio of Lagrange multipliers from MD. When

$h(\pi)$ is not a member of the Cressie-Read family, $h_{\pi}^{-1}(.)$ is not homogenous. It then appears to be impossible to factor out $\hat{\mu}_{MD}$, and that the MD and GEL estimators are different. We focus on the GEL class because for large n they are obtained from a much smaller dimensional optimization problem that MD. Also, the ability to estimate the distribution of the data using $\hat{\pi}_i$, $(1, ..., n)$, is not lost for a GEL estimator. As shown in Brown, Newey, and May (1997), the discrete distribution with $\Pr(z = z_i) = \rho_v(\hat{\lambda}_{GEL}' g_i(\hat{\beta}_{GEL})) / \sum_{j=1}^{n} \rho_v(\hat{\lambda}_{GEL}' g_j(\hat{\beta}_{GEL}))$ is an efficient estimator of the distribution of a single observation.

3 Higher Order Asymptotic Properties

We obtain stochastic expansions for the parameter vector $\theta = (\beta', \lambda')'$ that includes the auxiliary parameters λ as well as β. Including λ seems to simplify calculations and leads to interesting results for $\hat{\lambda}$. For the CUE we will find T_n such that

$$\hat{\theta}_{CUE} = \theta_0 + T_n + O_p(n^{-3/2}), \tag{3.1}$$

where T_n is quadratic in sample averages. As discussed in Rothenberg (1984), $E[T_n]$ will then be an approximate bias, even though moments of the CUE may not exist. Under certain regularity conditions $E[T_n]$ will coincide with the expectation of an Edgeworth approximation to the distribution of the estimator.

For another estimator $\tilde{\theta}$, either GEL or GMM, we will find R_n that is quadratic in sample moments such that

$$\tilde{\theta} = \hat{\theta}_{CUE} + R_n + O_p(n^{-3/2}), R_n = O_p(n^{-1}). \tag{3.2}$$

Thus, these two estimators are asymptotically equivalent, meaning $\sqrt{n}(\tilde{\theta} - \hat{\theta}_{CUE}) \xrightarrow{p} 0$, and R_n is the dominant remainder in the difference between them. We will be particularly interested in the case $R_n = 0$, where the estimator will be higher-order asymptotically equivalent to the CUE (meaning $n(\tilde{\theta} - \hat{\theta}_{CUE}) \xrightarrow{p} 0$). Also, when they are not higher-order equivalent, by summing equations (3.1) and (3.2) we find that $\tilde{\theta}$ also satisfies equation (3.1) with $T_n + R_n$ replacing T_n. Thus, the approximate bias of $\tilde{\theta}$ will be $E[T_n + R_n]$, which can be compared with the CUE bias $E[T_n]$. Throughout the rest of the paper, we will refer to these expectations as "bias," with the understanding that they are only approximate.

In this Section we report the form of T_n and R_n and their expectations, and compare them. In the next Section we give the regularity conditions for these expansions. To describe the results, let $g_i \equiv g_i(\beta_0)$, $G_i \equiv G_i(\beta_0)$, and

$$\begin{aligned} \Omega &\equiv E[g_i g_i'], G \equiv E[G_i], V \equiv (G'\Omega^{-1}G)^{-1}, \\ H &\equiv VG'\Omega^{-1}, P \equiv \Omega^{-1} - \Omega^{-1}GVG'\Omega^{-1}. \end{aligned}$$

Also, let a be an m-vector such that

$$a_j \equiv tr(VE[\partial^2 g_{ij}(\beta_0)/\partial\beta\partial\beta'])/2, (j = 1, ..., m), \tag{3.3}$$

where g_{ij} denotes the jth element of g_i, $(j = 1, ..., m)$.

Theorem 2 *If Assumptions 4.1-4.4 are satisfied then* $\hat{\theta}_{CUE} = \theta_0 + T_n + O_p(n^{-3/2})$ *where*

$$E[T_n] = -[H', P]'(a - E[G_i H g_i] - E[g_i g_i' P g_i])/n. \tag{3.4}$$

The bias for $\hat{\beta}_{CUE}$ is the first p elements of $E[T_n]$,

$$Bias(\hat{\beta}_{CUE}) = -Ha/n + E[HG_iHg_i]/n + HE[g_ig_i'Pg_i]/n$$

Each of the terms in the bias has an interpretation. The sum of the first two terms, $-Ha/n + E[HG_iHg_i]/n$, is precisely the bias for a GMM estimator with moment vector $G'\Omega^{-1}g(z,\beta)$ (see Lemma A.3). As shown by Hansen (1982), this moment vector is the optimal (asymptotic variance minimizing) linear combination of $g(z,\beta)$. Thus, this expression is the bias for the (infeasible) optimal GMM estimator where the optimal linear combination coefficients $G'\Omega^{-1}$ did not have to be estimated. The first term Ha arises from nonlinearity of the moments, and will be zero when they are linear in parameters. The second term is generally not zero whenever there is endogeneity, but it will tend not to be very big. For instrumental variables, where $g(z,\beta)$ is a product of instruments and residuals, the combined expression is essentially the bias with exact identification, where the instruments are the optimal linear combination (with true coefficients). It is well known that this bias will tend to be small, with the main problem being caused by over-identification. In particular, this term will generally not grow with the number of moments m.

The third term in the bias is due to estimation of Ω. In some cases this may be a significant source of bias. For example, Altonji and Segal (1996) show that when g_i involves variance and covariance indicators, this kind of bias can be important. This source of bias would not be present if a fixed weighting matrix W were used rather than the estimate $\hat{\Omega}(\hat{\beta}_{CUE})$. Also, it will be zero if the third moments of g_i are zero. For example, this will occur if g_i consists of products of instruments with a residual that is conditionally symmetric (e.g. Gaussian) given the instruments. Also, as we see below, for EL this term will disappear, even without symmetry.

In summary, for the CUE, the bias is the sum of the bias for an estimator using the optimal linear combination of instruments and a term arising from estimation of Ω. Thus, the bias does not include a term for estimation of G. As we now show, such a term does show up for GMM, and is an important source of bias for that estimator.

The bias terms for $\hat{\lambda}_{CUE}$ are similar to those for $\hat{\beta}_{CUE}$. These terms impact the properties of tests of over-identification. For instance, from the first-order conditions for CUE (obtained from eq. (2.7) with $\rho_v(v) = 1 + v$), we see that over-identification test statistic is

$$n\hat{g}(\hat{\beta}_{GEL})'\hat{\Omega}(\hat{\beta}_{GEL})^{-1}\hat{g}(\hat{\beta}_{GEL}) = n\hat{\lambda}'_{GEL}\hat{\Omega}(\hat{\beta}_{GEL})\hat{\lambda}_{GEL}.$$

Thus, when the biases are large, so that the distribution of $\hat{\lambda}_{GEL}$ is not centered near zero, the overidentification test will tend to be large, leading to over-rejection.

To describe the higher-order relationship between GMM and CUE and derive the bias for GMM, it is useful to introduce for GMM an auxiliary parameter estimator $\hat{\lambda}_{2S}$ analogous to that for GEL. Consider the equations

$$\begin{aligned} \hat{G}(\hat{\beta}_{2S})'\hat{\lambda}_{2S} &= 0, \\ \hat{g}(\hat{\beta}_{2S}) + \hat{\Omega}(\tilde{\beta})\hat{\lambda}_{2S} &= 0. \end{aligned} \qquad (3.5)$$

The first-order conditions for GMM are obtained by solving the second equation for $\hat{\lambda}_{2S}$ and substituting it into the first. Formulating GMM in this way makes comparison with GEL easier. For instance, for the iterated GMM estimator of Hanson, Heaton, and Yaron (1996) where $\tilde{\beta} = \hat{\beta}_{2S}$, we see by comparing this equation with eq. (2.7), for $\rho_v(v) = -(1+v)$, that the GMM and CUE first order conditions are identical except that the CUE first-order condition includes the additional term $\sum_{j=1}^m \hat{\lambda}'_{CUE,j}[\sum_{i=1}^n g_{ij}(\hat{\beta}_{CUE})G_i(\hat{\beta}_{CUE})'/n]\hat{\lambda}_{CUE}$. As shown in Donald and Newey (2000) and as we discuss below, it is exactly this term that helps to reduce the bias for CUE relative to GMM.

Let $\tilde{\lambda} = -P\hat{g}(\beta_0)$ and $G_{i,j\bullet}$ denote the jth row of G_i, $(j = 1, ..., m)$. We have the following result:

Theorem 3 *If Assumptions 4.1-4.4 are satisfied and* $\tilde{\beta} = \arg\min_\beta \hat{g}(\beta)'\hat{W}\hat{g}(\beta)$ *for* $\hat{W} = W + O_p(n^{-1/2})$ *and* W *positive-definite, then for* $H_W = (G'WG)^{-1}G'W$, $\hat{\theta}_{2S} = \hat{\theta}_{CUE} + R_n + O_p(n^{-3/2})$ *where*

$$\begin{aligned} R_n &= -[V, -H]' \sum_{j=1}^m \tilde{\lambda}_j E[g_{ij}G_i'] \tilde{\lambda} \\ &\quad + [H', P]' \sum_{j=1}^m \tilde{\lambda}_j \{E[g_{ij}G_i] + E[g_i G_{i,j\bullet}]\}(H - H_W)\hat{g}(\beta_0), \\ E[R_n] &= [-V, H]' E[G_i' P g_i]/n. \end{aligned} \tag{3.6}$$

The second term in R_n is due to the use of a non-optimal weighting matrix $\hat{W}$ in constructing the initial estimator $\tilde{\beta}$. It will be zero if $W = \Omega$, as for the iterated GMM estimator considered by Hansen, Heaton, and Yaron (1996). More generally, only one iteration is required for its disappearance. Also, even if it is nonzero, it contributes nothing towards the bias, because $\tilde{\lambda}$ and $(H - H_W)\hat{g}(\beta_0)$ are uncorrelated.

The first term in R_n is due to the estimation of G in the linear combination coefficients $G'\Omega^{-1}$ for the optimal GMM estimator. Its presence means that the bias of GMM estimator $\hat{\beta}_{2S}$ will be

$$Bias(\hat{\beta}_{2S}) = Bias(\hat{\beta}_{CUE}) - VE[G_i'Pg_i]/n.$$

The last term can be quite large, especially when there are many moment restrictions. In particular, note that $E[G_i'Pg_i]$ will tend to grow with the number of moment restrictions.

We can obtain precise expressions in the important example of a homoskedastic linear simultaneous equation. For simplicity we will consider the case with one right-hand side variable and a symmetric disturbance.

Example. Let x denote a m-vector of instrumental variables and y and w be endogenous variables satisfying

$$\begin{aligned} y &= \beta_0 w + \varepsilon, E[\varepsilon|x] = 0, E[\varepsilon^3|x] = 0, \\ \sigma_\varepsilon^2 &= var(\varepsilon|x), \sigma_{\varepsilon w} = cov(\varepsilon, w|x). \end{aligned}$$

Then for $g(z, \beta) = x(y - \beta w)$ and $Q = E[x_i x_i']$, it follows that $\Omega = \sigma_\varepsilon^2 Q$, $G = -E[x_i w_i]$, and $E[g_i G_i'] = -E[x_i \varepsilon_i w_i x_i'] = -\sigma_{\varepsilon w} Q$. Here the estimators are asymptotically equivalent to two-stage least squares. Note that

$$\begin{aligned} E[G_i'Pg_i] &= -E[\varepsilon_i w_i x_i'Px_i] = -\sigma_{\varepsilon w} E[x_i'Px_i] = -\sigma_{\varepsilon w} tr(QP) \\ &= -(\sigma_{\varepsilon w}/\sigma_\varepsilon^2)(m-1). \end{aligned}$$

Also,

$$HE[G_i g_i']H' = -\sigma_{\varepsilon w} HQH' = -(\sigma_{\varepsilon w}/\sigma_\varepsilon^2)V.$$

Then we have the bias expressions for β corresponding to the CUE and GMM respectively,

$$Bias(\hat{\beta}_{CUE}) = -(\sigma_{\varepsilon w}/\sigma_\varepsilon^2)V/n, Bias(\hat{\beta}_{2S}) = (m-2)(\sigma_{\varepsilon w}/\sigma_\varepsilon^2)V/n.$$

It is interesting to note that these are exactly the limited information maximum likelihood and two-stage least squares approximate biases found by Rothenberg (1983) and Nagar (1959) respectively. In this example the bias of the two step GMM estimator increases linearly with m while the bias of the CUE estimator does not depend on m.

To derive the relationship between the CUE and GEL estimators it is convenient to renormalize by replacing $\rho(v)$ with $\rho([\rho_v(0)/\rho_{vv}(0)]v)$. This replacement has no effect on $\hat{\beta}_{GEL}$ but makes the scale of $\hat{\lambda}_{GEL}$ comparable for different $\rho(v)$. Then we have the following result:

Theorem 4 *If Assumptions 4.1-4.4 are satisfied then* $\hat{\theta}_{GEL} = \hat{\theta}_{CUE} + R_n + O_p(n^{-3/2})$ *for*

$$\begin{aligned} R_n &= -[H', P]'\{\rho_v(0)\rho_{vvv}(0)/2\rho_{vv}(0)^2\}\sum_{j=1}^{m}\tilde{\lambda}_j E[g_{ij}g_i g_i']\tilde{\lambda}, \\ E[R_n] &= -[H', P]'\{\rho_v(0)\rho_{vvv}(0)/2\rho_{vv}(0)^2\}E[g_i g_i' P g_i]/n. \end{aligned}$$

The remainder term is a linear combination of quadratic forms in $\tilde{\lambda}$. It will be zero, when for each j, $E[g_{ij}g_i g_i'] = 0$. As previously noted, this condition may not be satisfied, but will in the important case of instrumental variables estimation with a symmetric disturbance. In that case CUE and GEL estimators are higher-order asymptotically equivalent. Also, both will have bias that is the same as the GMM estimator with moment functions $G'\Omega^{-1}g(z, \beta)$.

In general,

$$Bias(\hat{\beta}_{GEL}) = -Ha/n + E[HG_iHg_i]/n + [1 - \rho_v(0)\rho_{vvv}(0)/2\rho_{vv}(0)^2]HE[g_ig_i'Pg_i]/n \quad (3.7)$$

The last term disappears when $\rho_v(0)\rho_{vvv}(0) = 2\rho_{vv}(0)^2$. Since for EL, where $\rho(v) = \ln(1+v)$, this equality is satisfied, we find that

$$Bias(\hat{\beta}_{EL}) = -Ha/n + E[HG_iHg_i]/n.$$

Thus, the EL estimator of β has a higher-order bias that is exactly the same as for an estimator with moment functions $G'\Omega^{-1}g(z,\beta)$. For the EL estimator, there is no bias from estimation of either G or Ω. This same property would be shared by any GEL estimator with same value for $\rho_v(0)\rho_{vvv}(0) = 2\rho_{vv}(0)^2$, although by Theorem 3.3 it follows that any such estimator would be higher-order equivalent to EL. For instance, in the Cressie-Read family $\rho_v(0)\rho_{vvv}(0)/2\rho_{vv}(0)^2 = (1-\gamma)/2$, so the EL estimator (obtained as $\gamma \to -1$) is the only member of this family that has this zero bias property.

The reduction in bias from GEL estimators does not come for free. Monte Carlo evidence of Hansen, Heaton, and Yaron (1996) shows that it is accompanied by substantial increases in variance. This generally leads to more accurate coverage probabilities for confidence intervals, but can substantially lengthen those intervals. It would be useful to compare the higher-order variances also, in order to understand this trade-off. This comparison is beyond the scope of this paper, but is the subject of ongoing research.

4 Asymptotic Theory for GEL

To obtain precise results it is necessary to specify several regularity conditions.

Assumption 4.1 **(a)** *$\beta_0 \in \mathcal{B}$ is the unique solution to $E[g(z,\beta)] = 0$;* **(b)** *$\mathcal{B}$ is compact;* **(c)** *$g(z,\beta)$ is continuous at each $\beta \in \mathcal{B}$ with probability one;* **(d)** *$E\left[\sup_{\beta\in\mathcal{B}} \|g(z,\beta)\|^{\alpha}\right] < \infty$ for some $\alpha > 2$;* **(e)** *Ω is nonsingular.*

Most of these regularity conditions are standard for GMM (e.g. Newey and McFadden, 1994). Existence of higher than second moments, however, is important for GEL estimators.

Assumption 4.2 **(a)** *$\rho(v)$ is twice continuously differentiable and concave on its domain, an open interval V containing 0, $\rho_v(0) \neq 0$, and $\rho_{vv}(0) < 0$;* **(b)** *either* **(i)** *$\Gamma_n = \{\lambda : \|\lambda\| \leq Dn^{-\gamma}\}$ with $\frac{1}{2} > \gamma > \frac{1}{\alpha}$, or* **(ii)** *$V = \Re$, $\Gamma_n = \Re^m$;* **(c)** *for some neighborhood $\mathcal{N}$ of β_0, $\varepsilon > 0$, and for all λ, $E[\sup_{\beta\in\mathcal{N}} |\rho(\lambda' g(z,\beta))|] < \infty$, $E[\sup_{\beta\in\mathcal{N},\|\lambda\|\leq\varepsilon} \|g(z,\beta)\|^2 |\rho_{vv}(\lambda' g(z,\beta))|] < \infty$.*

This assumption specifies that either the domain of $\rho(v)$ is the real line or bounds are placed on λ. When combined with the existence of higher than second moments in the previous assumption, this condition leads to $\lambda' g_i(\beta)$ being in the domain V of $\rho(v)$ for all β and $i \leq n$.

These conditions lead to a consistency result, and more.

Theorem 5 *If Assumptions 4.1 and 4.2 are satisfied then $\hat{\beta}_{GEL} \xrightarrow{p} \beta_0$. Also, $\|\hat{\lambda}_{GEL}\| = O_p(n^{-1/2})$, and $\|\hat{g}(\hat{\beta}_{GEL})\| = O_p(n^{-1/2})$.*

This result also gives a convergence rate for $\hat{\lambda}_{GEL}$ and $\hat{g}(\hat{\beta}_{GEL})$ as a by-product of consistency. A similar result holds for the two-step GMM estimator under these regularity conditions.

For asymptotic normality we need additional regularity conditions.

Assumption 4.3 **(a)** *$\beta_0 \in int(\mathcal{B})$;* **(b)** *$g(z, \beta)$ is differentiable in a neighborhood $\mathcal{N}$ of β_0 and $E[\sup_{\beta \in \mathcal{N}} \|\partial g(z, \beta)/\partial \beta'\|] < \infty$;* **(c)** *$rank(G) = p$.*

This is a standard smoothness condition for the asymptotic normality of GMM that could be relaxed. Under this condition it follows that the GEL estimator is asymptotically normal:

Theorem 6 *If Assumptions 4.1 - 4.3 are satisfied then*

$$\sqrt{n}(\hat{\beta}_{GEL} - \beta_0) \xrightarrow{d} N(0, V), \sqrt{n}\hat{\lambda}_{GEL} \xrightarrow{d} N(0, P),$$
$$n[2\rho_{vv}(0)/\rho_v(0)^2][\rho(0) - \hat{P}(\hat{\beta}_{GEL}, \hat{\lambda}_{GEL})] \xrightarrow{d} \chi^2(m - p).$$

This result shows asymptotic normality of the GEL estimators $\hat{\beta}_{GEL}$ and $\hat{\lambda}_{GEL}$, and that, properly normalized, the saddle-point objective function has a limiting chi-squared distribution. This statistic is a version of the GMM test statistic of overidentifying restrictions. Indeed, as noted above, for the CUE where $\rho(v)$ is quadratic, it is identical to the GMM statistic. Moreover, the proof of Theorem 4.2 reveals that the GEL estimators $\hat{\beta}_{GEL}$ and $\hat{\lambda}_{GEL}$ are asymptotically uncorrelated.

More smoothness is needed for validity of the asymptotic expansions reported in Section 3. Assumption 4.4 below strengthens aspects of Assumptions 4.1-4.3 which together are then sufficient for the validity of the expansions of the previous section.

Assumption 4.4 **(a)** $\rho(.)$ *is four times continuously differentiable in a neighborhood of* 0*;* **(b)** $E[\sup_{\beta\in\mathcal{N}} \|\partial g(z,\beta)/\partial\beta'\|^{\alpha/(\alpha-2)}] < \infty$*;* **(c)** $\alpha \geq 4$*;* **(d)** $E[\|g(z,\beta)\|^6] < \infty$*;* **(e)** $g(z,\beta)$ *is four times continuously differentiable on* $\mathcal{N}$ *and for any element of* $g(z,\beta)$ *and any partial derivative* $\Delta(z,\beta)$ *up to order four,* $E[\sup_{\beta\in\mathcal{N}} |\Delta(z,\beta)|^2] < \infty$*.*

With this assumption in hand the validity of the expansions given in Theorems 3.1-3.3 can be shown, as is done in the following Appendix.

Appendix: Proofs

Proof of Theorem 2.1: Let $A = [g_1(\beta), ..., g_n(\beta)]'/\sqrt{n}$ and $\iota = (1, ..., 1)'$ be an n-vector of units. Thus, $\hat{g}(\beta) = A'\iota/\sqrt{n}$ and $\hat{\Omega}(\beta) = A'A$. By Rao (1973, 1b.5(vi),(viii)), $A(A'A)^{-}A'$ is invariant to the choice of generalized inverse and $A'A(A'A)^{-}A' = A$ for any generalized inverse. Since the CUE objective function is $\iota'A(A'A)^{-}A'\iota/n$, it is invariant to the generalized inverse. By $\rho(v)$ quadratic, a second-order Taylor expansion is exact, giving

$$\hat{P}(\beta, \lambda) = \rho(0) + \rho_v(0)\hat{g}(\beta)'\lambda + \frac{1}{2}\rho_{vv}(0)\lambda'\hat{\Omega}(\beta)\lambda. \tag{A.1}$$

By concavity of $\hat{P}(\beta, \lambda)$ in λ, any solution $\hat{\lambda}(\beta)$ to the first-order conditions

$$0 = \rho_v(0)\hat{g}(\beta) + \rho_{vv}(0)\hat{\Omega}(\beta)\lambda$$

will maximize $\hat{P}(\beta, \lambda)$ with respect to λ holding β fixed. Then, $\hat{\Omega}(\beta)\hat{\Omega}(\beta)^{-}\hat{g}(\beta) = A'A(A'A)^{-}A'\iota/\sqrt{n} = \hat{g}(\beta)$, so that $\hat{\lambda}(\beta) = -[\rho_v(0)/\rho_{vv}(0)]\hat{\Omega}(\beta)^{-}\hat{g}(\beta)$ solves the first-order conditions. Substituting $\hat{\lambda}(\beta)$ back in (A.1) gives

$$\hat{P}(\beta, \hat{\lambda}(\beta)) = \rho(0) - [\rho_v(0)^2/2\rho_{vv}(0)]\hat{g}(\beta)'\hat{\Omega}(\beta)^{-}\hat{g}(\beta). \tag{A.2}$$

Since $\rho_{vv}(0) < 0$ by concavity of $\rho(v)$, $\hat{P}(\beta, \hat{\lambda}(\beta))$ is a monotonic increasing transformation of the CUE objective function. Therefore, CUE minimizes $\hat{P}(\beta, \hat{\lambda}(\beta))$. Furthermore, from the saddle point form of eq. (2.4), the GEL estimator also minimizes this function. Therefore, the set of GEL estimators coincides with the set of CUE. Q.E.D.

Throughout the Appendix, C will denote a generic positive constant that may be different in different uses. Also, with probability approaching one will be abbreviated as w.p.a.1, positive semi-definite as p.s.d., UWL will denote a uniform weak law of large numbers such as Lemma 2.4 of Newey and McFadden (1994), and CLT will refer to the Lindbergh-Levy central limit theorem. Before proving Theorem 4.1 we give two preliminary lemmas.

Lemma 7 *If Assumptions 4.1 and 4.2 are satisfied then there is C such that w.p.a.1,*

$$\sup_{\lambda \in \Gamma_n} \hat{P}(\beta_0, \lambda) \leq \rho(0) + C\|\hat{g}(\beta_0)\|^2.$$

Proof. Firstly, consider Assumption 4.2(b)(ii) in which the domain of $\rho(v)$ is the real line. For $g_i \equiv g_i(\beta_0)$, we have $P_0(\lambda) \equiv E[\rho(\lambda' g_i)]$ existing for each λ by Assumption 4.2(c). By global concavity of $\rho(v)$ and strict concavity on a neighborhood of zero (by $\rho_{vv}(0) < 0$),

$$\rho(\lambda' g_i) \leq \rho(0) + \rho_v(0)\lambda' g_i, \tag{A.3}$$

with strict inequality for $\lambda' g_i \neq 0$. Also, by Ω nonsingular, for any $\lambda \neq 0$, $\Pr(\lambda' g_i \neq 0) > 0$, so that $P_0(\lambda) < P_0(0) = \rho(0)$. Then it follows by standard consistency results for concave objective functions (e.g. Newey and McFadden, 1994, Theorem 2.7) that $\bar{\lambda} = \arg\max_{\lambda \in \Re^m} \hat{P}(\beta_0, \lambda)$ exists w.p.a.1 and $\bar{\lambda} \xrightarrow{p} 0$. Then, by Assumption 4.2(b)(ii) and UWL, for any $\dot{\lambda} = \tau\bar{\lambda}$, $0 \leq \tau \leq 1$, $\sum_i \rho_{vv}(\dot{\lambda}' g_i) g_i g_i'/n \xrightarrow{p} \rho_{vv}(0)\Omega$.

Next, under Assumption 4.2(b)(i), let $b_i = \sup_{\beta \in \mathcal{B}} \|g_i(\beta)\|$. Then by Assumption 4.1 and the Markov inequality, $n^{-1}\sum_{i=1}^n b_i^\alpha = O_p(1)$, so that

$$\max_{1 \leq i \leq n} b_i = \left(\max_{1 \leq i \leq n} b_i^\alpha\right)^{1/\alpha} \leq n^{1/\alpha}\left(n^{-1}\sum_{i=1}^n b_i^\alpha\right)^{1/\alpha} = O_p(n^{1/\alpha}). \tag{A.4}$$

It follows from (A.4) and Assumption 4.2(b)(i) that

$$\sup_{\beta\in\mathcal{B}}\sup_{\lambda\in\Gamma_n}\max_{1\le i\le n}|\lambda' g_i(\beta)| \le Dn^{-\gamma}\max_{1\le i\le n} b_i \xrightarrow{p} 0.$$

Therefore, w.p.a.1, $\lambda' g_i(\beta) \in V$ for all $i \le n$, $\lambda \in \Gamma_n$ and $\beta \in \mathcal{B}$. For this case, let $\bar{\lambda} = \arg\max_{\lambda\in\Gamma_n} \hat{P}(\beta_0, \lambda)$ and $\dot{\lambda} = \tau\bar{\lambda}$, $0 \le \tau \le 1$. Then $\max_{i\le n}|\rho_{vv}(\dot{\lambda}' g_i) - \rho_{vv}(0)| \xrightarrow{p} 0$, so that $\sum_i \rho_{vv}(\dot{\lambda}' g_i) g_i g_i'/n \xrightarrow{p} \rho_{vv}(0)\Omega$.

Next, note that in either case, it follows by $\rho_{vv}(0)\Omega$ negative definite, that there exists $C > 0$ such that w.p.a.1, $\sum_i \rho_{vv}(\dot{\lambda}' g_i) g_i g_i'/n \le -CI$ in the p.s.d. sense. Then by a second-order Taylor expansion with Lagrange remainder, we have w.p.a.1,

$$\begin{aligned}\sup_{\lambda\in\Gamma_n}\hat{P}(\beta_0,\lambda) &= \hat{P}(\beta_0,\bar{\lambda}) = \rho(0) + \rho_v(0)\bar{\lambda}'\hat{g}(\beta_0) + \bar{\lambda}'\left[\sum_{i=1}^n \rho_{vv}(\dot{\lambda}' g_i) g_i g_i'/n\right]\bar{\lambda}/2 \\ &\le \rho(0) + \rho_v(0)\bar{\lambda}'\hat{g}(\beta_0) - C\bar{\lambda}'\bar{\lambda}/2 \\ &\le \rho(0) + \sup_{\lambda}[\rho_v(0)\lambda'\hat{g}(\beta_0) - C\lambda'\lambda/2] = \rho(0) + [\rho_v(0)^2/2C]\|\hat{g}(\beta_0)\|^2.\end{aligned}$$

Q.E.D.

Lemma 8 *If Assumptions 4.1 and 4.2 are satisfied, then* $\|\hat{g}(\hat{\beta}_{GEL})\| = O_p(n^{-1/2})$.

Proof. Let $\hat{\beta} = \hat{\beta}_{GEL}$, $\hat{g}_i = g_i(\hat{\beta})$ and $\delta_n = Dn^{-\gamma}$ for γ and D as in Assumption 4.2(b)(i) and $\bar{\lambda} = \hat{g}(\hat{\beta})\delta_n sgn(\rho_v(0))/\|\hat{g}(\hat{\beta})\|$. It follows similarly to the proof of Lemma A.1 that $\max_{i\le n}|\bar{\lambda}'\hat{g}_i| \xrightarrow{p} 0$, and hence for any $\dot{\lambda}$ as above, $\sum_i \rho_{vv}(\dot{\lambda}'\hat{g}_i)\hat{g}_i\hat{g}_i'/n - \rho_{vv}(0)\hat{\Omega}(\hat{\beta}) \xrightarrow{p} 0$. Also, $\hat{\Omega}(\hat{\beta}) = O_p(1)$. Therefore, w.p.a.1 $\sum_i \rho_{vv}(\dot{\lambda}'\hat{g}_i)\hat{g}_i\hat{g}_i'/n \ge -CI$ in the p.s.d. sense, so by a second-order Taylor expansion

$$\begin{aligned}\hat{P}(\hat{\beta},\bar{\lambda}) &\ge \rho(0) + \rho_v(0)\bar{\lambda}'\hat{g}(\hat{\beta}) - C\bar{\lambda}'\bar{\lambda} \qquad \text{(A.5)} \\ &= \rho(0) + |\rho_v(0)|\|\hat{g}(\hat{\beta})\|\delta_n - C\delta_n^2.\end{aligned}$$

w.p.a.1. Noting that $\hat{P}(\hat{\beta}, \bar{\lambda}) \leq \sup_{\lambda \in \Gamma_n} \hat{P}(\hat{\beta}, \lambda) \leq \sup_{\lambda \in \Gamma_n} \hat{P}(\beta_0, \lambda)$, it follows by Lemma A.1 and eq. (A.5) that w.p.a.1, $|\rho_v(0)| \|\hat{g}(\hat{\beta})\| \delta_n - C\delta_n^2 \leq C\|\hat{g}(\beta_0)\|^2$. Solving for $\|\hat{g}(\hat{\beta})\|$ then gives

$$\|\hat{g}(\hat{\beta})\| \leq C\|\hat{g}(\beta_0)\|^2/\delta_n + C\delta_n = O_p(\delta_n), \tag{A.6}$$

as $\|\hat{g}(\beta_0)\|^2 = O_p(n^{-1})$ by CLT. Now, for any $\varepsilon_n \to 0$, redefine $\bar{\lambda} = \varepsilon_n \hat{g}(\hat{\beta}) sgn(\rho_v(0))$. Note that $\bar{\lambda} = o_p(\delta_n)$ by eq. (A.6), so that $\bar{\lambda} \in \Gamma_n$ w.p.a.1. Then, by a similar argument to that above,

$$\varepsilon_n \|\hat{g}(\hat{\beta})\|^2 (|\rho_v(0)| - \varepsilon_n C) \leq C\|\hat{g}(\beta_0)\|^2 = O_p(n^{-1}).$$

Since, for all n large enough, $|\rho_v(0)| - \varepsilon_n C$ is bounded away from zero, it follows that $\varepsilon_n \|\hat{g}(\hat{\beta})\|^2 = O_p(n^{-1})$. The conclusion then follows by a standard result from probability theory, that if $\varepsilon_n Y_n = O_p(n^{-1})$ for all $\varepsilon_n \to 0$, then $Y_n = O_p(n^{-1})$. Q.E.D.

Proof of Theorem 4.1: Let $\hat{\beta} = \hat{\beta}_{GEL}$, $\hat{\lambda} = \hat{\lambda}_{GEL}$ and $g(\beta) = E[g(z, \beta)]$. By Lemma A.2, $\hat{g}(\hat{\beta}) \xrightarrow{p} 0$, and by UWL, $\sup_{\beta \in \mathcal{B}} \|\hat{g}(\beta) - g(\beta)\| \xrightarrow{p} 0$ and $g(\beta)$ is continuous. The triangle inequality then gives $g(\hat{\beta}) \xrightarrow{p} 0$. Since $g(\beta) = 0$ has a unique zero β_0, $\|g(\beta)\|$ must be bounded away from zero outside any neighborhood of β_0. Therefore, $\hat{\beta}$ must be inside any neighborhood of β_0 w.p.a.1, i.e. $\hat{\beta} \xrightarrow{p} \beta_0$.

Next, by eq. (A.3) and UWL, $P(\beta, \lambda) = E[\rho(\lambda' g(z, \beta))]$ is continuous in β for given λ and $\sup_{\beta \in \mathcal{B}} |\hat{P}(\beta, \lambda) - P(\beta, \lambda)| \xrightarrow{p} 0$. Then by the consistency of $\hat{\beta}$, $\hat{P}(\hat{\beta}, \lambda) \xrightarrow{p} P_0(\lambda)$ for all λ. Let $\hat{g}_i = g_i(\hat{\beta})$. It then follows as in the proof of Lemma A.1 that $\hat{\lambda} \xrightarrow{p} 0$ and by UWL under Assumption 4.2(b)(ii) that for any $\dot{\lambda} = \tau\hat{\lambda}$, $0 \le \tau \le 1$, $\sum_i \rho_{vv}(\dot{\lambda}'\hat{g}_i)\hat{g}_i\hat{g}_i'/n \xrightarrow{p} \rho_{vv}(0)\Omega$. Also, it follows similarly to the proof of Lemma A.1 that this same condition holds under Assumption 4.2(b)(i). Then by a second-order Taylor expansion,

$$\begin{aligned}
\rho(0) &= \hat{P}(\hat{\beta}, 0) \le \hat{P}(\hat{\beta}, \hat{\lambda}) \\
&= \rho(0) + \rho_v(0)\hat{\lambda}'\hat{g}(\hat{\beta}) + \hat{\lambda}'[\sum_i \rho_{vv}(\dot{\lambda}'\hat{g}_i)\hat{g}_i\hat{g}_i'/n]\hat{\lambda} \\
&\le \rho(0) + |\rho_v(0)|\|\hat{\lambda}\|\|\hat{g}(\hat{\beta})\| - C\|\hat{\lambda}\|^2.
\end{aligned}$$

Dividing through by $\|\hat{\lambda}\|$ and solving gives $\|\hat{\lambda}\| \le C\|\hat{g}(\hat{\beta})\| = O_p(n^{-1/2})$, giving the second conclusion. Q.E.D.

Proof of Theorem 4.2: By Theorem 4.1, w.p.a.1 the constraint on λ, if present, is not binding, and by β_0 in the interior of $\mathcal{B}$ neither is the constraint $\beta \in \mathcal{B}$. Therefore, the first order conditions of eq. (2.7) are satisfied w.p.a.1 with $\rho(kv)$ replacing $\rho(v)$ for $k = \rho_v(0)/\rho_{vv}(0)$. Then by a mean-value expansion of the second part of these first order conditions we have, for $\hat{\theta}[= (\hat{\beta}', \hat{\lambda}')'] = (\hat{\beta}_{GEL}', \hat{\lambda}_{GEL}')'$, $\theta_0 = (\beta_0', 0')'$, $\hat{g}_i = g_i(\hat{\beta})$,

$$\begin{aligned}
0 &= \begin{pmatrix} 0 \\ \rho_v(0)\hat{g}(\beta_0) \end{pmatrix} + \bar{M}(\hat{\theta} - \theta_0), && \text{(A.7)} \\
\bar{M} &= \begin{pmatrix} 0 & \sum_{i=1}^n \rho_v(k\hat{\lambda}'\hat{g}_i)G_i(\hat{\beta})'/n \\ \sum_{i=1}^n \rho_v(k\bar{\lambda}'\hat{g}_i)G_i(\bar{\beta})/n & \sum_{i=1}^n k\rho_{vv}(k\bar{\lambda}'\hat{g}_i)g_i(\bar{\beta})\hat{g}_i'/n \end{pmatrix},
\end{aligned}$$

where $\bar{\beta}$ and $\bar{\lambda}$ are mean-values that actually differ from row to row of the matrix $\bar{M}$. By $\bar{\lambda} = O_p(n^{-1/2})$, it follows by an arguments like those of the proof of Theorem 4.1 that for $\tilde{\lambda}$ equal to $\hat{\lambda}$ or $\bar{\lambda}$,

$$\max_{i \le n} |\tilde{\lambda}'\hat{g}_i| \le \|\tilde{\lambda}\| \max_{i \le n} \|\hat{g}_i\| = O_p(n^{-1/2}n^{1/\alpha}) \xrightarrow{p} 0.$$

Therefore,

$$\max_{i\le n}|\rho_v(k\tilde{\lambda}'\hat{g}_i)-\rho_v(0)|\xrightarrow{p}0,\max_{i\le n}|\rho_{vv}(k\bar{\lambda}'\hat{g}_i)-\rho_{vv}(0)|\xrightarrow{p}0.$$

It then follows from UWL that $\bar{M}\xrightarrow{p}\rho_v(0)M$, where

$$M=\begin{pmatrix}0 & G'\\ G & \Omega\end{pmatrix}.$$

Inverting and solving in eq. (A.7) then gives

$$\begin{aligned}\sqrt{n}(\hat{\theta}-\theta_0) &= -\bar{M}^{-1}(0,\rho_v(0)\sqrt{n}\hat{g}(\beta_0)')'=-M^{-1}(0,\sqrt{n}\hat{g}(\beta_0)')'+o_p(1) \quad \text{(A.8)}\\ &= -(H',P)'\sqrt{n}\hat{g}(\beta_0)+o_p(1).\end{aligned}$$

The first conclusion follows from this equation and the CLT. The second conclusion follows similarly. For the third conclusion, note that an expansion and eq. (A.8) give

$$\hat{g}(\hat{\beta})=\hat{g}(\beta_0)-GH\hat{g}(\beta_0)+o_p(n^{-1/2})=-\Omega\hat{\lambda}+o_p(n^{-1/2}).$$

Expanding,

$$\begin{aligned}\hat{P}(\hat{\beta},\hat{\lambda}) &= \rho(0)+k\rho_v(0)\hat{\lambda}'\hat{g}(\hat{\beta})+k^2\hat{\lambda}'[\sum_{i=1}^{n}\rho_{vv}(k\bar{\lambda}'\hat{g}_i)\hat{g}_i\hat{g}_i'/n]\hat{\lambda}/2 \quad \text{(A.9)}\\ &= \rho(0)+k\rho_v(0)\hat{\lambda}'\hat{g}(\hat{\beta})+k^2\rho_{vv}(0)\hat{\lambda}'\Omega\hat{\lambda}/2+o_p(n^{-1})\\ &= \rho(0)-[\rho_v(0)^2/2\rho_{vv}(0)]\hat{g}(\hat{\beta})'\Omega^{-1}\hat{g}(\hat{\beta})+o_p(n^{-1}).\end{aligned}$$

It follows as in Hansen (1982) that $n\hat{g}(\hat{\beta})'\Omega^{-1}\hat{g}(\hat{\beta})\xrightarrow{d}\chi^2(m-p)$, so the conclusion follows from eq. (A.9). Q.E.D.

The next result is used to obtain the higher-order expansion for the CUE. Let $q=p+m$.

Lemma 9 *Consider an estimator* $\hat{\theta}$ *such that* $\hat{\theta} \xrightarrow{p} \theta_0$ *and (a)* $\hat{m}(\hat{\theta}) = 0$ *and* $\hat{m}(\theta_0) = O_p(n^{-1/2})$*; (b)* $M = M(\theta_0)$ *is nonsingular and for any* $\bar{\theta} \xrightarrow{p} \theta_0$*,* $\hat{M}(\bar{\theta}) \xrightarrow{p} M$*, where* $\hat{M}(\theta) = \partial\hat{m}(\theta)/\partial\theta'$*; (c)* $\hat{m}(\theta)$ *is three times continuously differentiable; (d)* $\hat{M}(\theta_0) = M + O_p(n^{-1/2})$ *and* $\partial\hat{M}(\theta_0)/\partial\theta_j = M_{\theta j} + O_p(n^{-1/2})$*,* $(j = 1, ..., q)$*; (e) in a neighborhood* $\mathcal{N}$ *of* θ_0*,* $\sup_{\theta\in\mathcal{N}} \|\partial^2\hat{M}(\theta)/\partial\theta_j\partial\theta_k\| = O_p(1)$*. Then for* $\hat{U} = -M^{-1}\hat{m}(\theta_0)$ *and* $\hat{M} = \hat{M}(\theta_0)$*,*

$$\hat{\theta} = \theta_0 + \hat{U} - M^{-1}\{[\hat{M} - M]\hat{U} + \sum_{j=1}^{q} \hat{U}_j M_{\theta j}\hat{U}/2\} + O_p(n^{-3/2}). \qquad \text{(A.10)}$$

Proof. A mean-value expansion then gives $\hat{\theta} - \theta_0 = -\hat{M}(\bar{\theta})^{-1}\hat{m}(\theta_0)$ w.p.a.1, so that $\hat{\theta} = \theta_0 + O_p(n^{-1/2})$, and hence $\hat{\theta} - \theta_0 = \hat{U} + O_p(n^{-1})$. Expanding,

$$0 = \hat{m}(\theta_0) + \hat{M}(\hat{\theta} - \theta_0) + \sum_{j=1}^{q}(\hat{\theta}_j - \theta_{j0})[\partial\hat{M}(\bar{\theta})/\partial\theta_j](\hat{\theta} - \theta_0)/2.$$

Note that replacing $\partial\hat{M}(\bar{\theta})/\partial\theta_j$ by $M_{\theta j}$ in the last term leads to an error that is $O_p(n^{-3/2})$ by hypotheses (d) and (e). Then adding and subtracting $M(\hat{\theta} - \theta_0)$ and solving gives

$$\begin{aligned}\hat{\theta} &= \theta_0 - M^{-1}[\hat{m}(\theta_0) + (\hat{M} - M)(\hat{\theta} - \theta_0) + \sum_{j=1}^{q}(\hat{\theta}_j - \theta_{j0})M_{\theta j}(\hat{\theta} - \theta_0)/2] + O_p(n^{-3/2}) \\ &= \theta_0 + \hat{U} - M^{-1}[(\hat{M} - M)\hat{U} + \sum_{j=1}^{q}\hat{U}_j M_{\theta j}\hat{U}/2] + O_p(n^{-3/2}).\end{aligned}$$

Q.E.D.

Proof of Theorem 3.1: Let

$$\begin{aligned}\hat{m}(\theta) &= \sum_{i=1}^{n}(1 + \lambda' g_i(\beta))(\lambda' G_i(\beta), g_i(\beta)')'/n, \\ \hat{M}(\theta) &= \partial\hat{m}(\theta)/\partial\theta' = \sum_{i=1}^{n}\begin{pmatrix} G_i(\beta)'\lambda \\ g_i(\beta)\end{pmatrix}(\lambda' G_i(\beta), g_i(\beta)')/n\end{aligned}$$

$$+\sum_{i=1}^{n}(1+\lambda' g_i(\beta))\begin{pmatrix} \partial[G_i(\beta)'\lambda]/\partial\beta' & G_i(\beta)' \\ G_i(\beta) & 0 \end{pmatrix}/n.$$

It follows as in the proof of Theorem 4.2 that hypotheses (a) and (b) of Lemma A.3 are satisfied with

$$M = \begin{pmatrix} 0 & G' \\ G & \Omega \end{pmatrix}, M^{-1} = \begin{pmatrix} -V & H \\ H' & P \end{pmatrix}.$$

Also, condition (c) holds by Assumption 4.4. Let $G_{i,\bullet j}$ be the j^{th} column of G_i, and $G_{i,j\bullet}$ be the j^{th} row of G_i. Then Assumption 4.4 and the CLT give hypothesis (d) of Lemma A.3 with

$$M_{\theta j} = \begin{pmatrix} 0 & E[\partial G_i(\beta_0)'/\partial\beta_j] \\ E[\partial G_i(\beta_0)/\partial\beta_j] & E[G_{i,\bullet j}g_i'] + E[g_i G_{i,\bullet j}'] \end{pmatrix}, (j \le p),$$

$$M_{\theta,j+p} = \begin{pmatrix} E[\partial G_{i,\bullet j}(\beta_0)'/\partial\beta] & E[g_{ij}G_i' + G_{i,j\bullet}' g_i'] \\ E[g_{ij}G_i + g_i G_{i,j\bullet}] & 0 \end{pmatrix}, (j \ge 1).$$

Condition (e) of Lemma A.3 follows from Assumption 4.4 and UWL. Note that from the definition in Lemma A.3, $\hat{m}(\theta_0) = (0, \hat{g}(\beta_0)')'$ and $\hat{U}[= (\hat{U}_\beta', \hat{U}_\lambda')'] = -[H', P]'\hat{g}(\beta_0)$, so that $E[\hat{U}\hat{U}'] = diag[V, P]/n$. Hence,

$$E[\sum_{j=1}^{q}\hat{U}_j M_{\theta j}\hat{U}/2] = E[\sum_{j=1}^{p}\hat{U}_{\beta_j}\begin{pmatrix} 0 \\ E[\partial G_i(\beta_0)/\partial\beta_j] \end{pmatrix}\hat{U}_\beta]/2$$

$$+E[\sum_{j=1}^{m}\hat{U}_{\lambda_j}\begin{pmatrix} E[g_{ij}G_i' + G_{i,j\bullet}' g_i'] \\ 0 \end{pmatrix}\hat{U}_\lambda]/2.$$

Now, denoting the jth column of P by P_j, $(j = 1, ..., m)$,

$$\sum_{j=1}^{m} E[\hat{U}_{\lambda_j}\{E[g_{ij}G_i'] + E[G_{i,j\bullet}' g_i']\}\hat{U}_\lambda]/2 = \sum_{j=1}^{m} tr(E[g_{ij}G_i']E[\hat{U}_\lambda\hat{U}_{\lambda_j}])$$

$$= \sum_{j=1}^{m} tr(E[g_{ij}G_i']P_j)/n = \sum_{j=1}^{m} E[G_i'P_j g_{ij}]/n$$
$$= E[G_i'Pg_i]/n.$$

Therefore, noting $E[\hat{U}_\beta' E[\partial^2 g_{ij}(\beta_0)/\partial\beta\partial\beta']\hat{U}^\beta]/2 = a_j/n$, it then follows that

$$E[\sum_{j=1}^{q} \hat{U}_j M_{\theta j}\hat{U}/2] = (0, a'/n)' + (E[G_i'Pg_i]'/n, 0)'.$$

Also,

$$E[(\hat{M} - M)\hat{U}] = E[\hat{M}\hat{U}] = -\begin{pmatrix} E[G_i'Pg_i]/n \\ E[G_iHg_i]/n + E[g_ig_i'Pg_i]/n \end{pmatrix}.$$

The conclusion then immediately follows from (A.10) of Lemma A.3. Q.E.D.

The next result is used to obtain the asymptotic expansions given in Section 3.

Lemma 10 *Consider two estimators $\hat{\theta}_j$ such that (a) $\hat{\theta}_j = \theta_0 + O_p(n^{-1/2})$ and $\hat{m}_j(\hat{\theta}_j) = 0$, $(j = 1, 2)$; (b) there is a nonsingular M such that for any $\bar{\theta} = \theta_0 + O_p(n^{-1/2})$, $\partial\hat{m}_1(\bar{\theta})/\partial\theta = M + O_p(n^{-1/2})$; (c) $\hat{m}_2(\hat{\theta}_2) = \hat{m}_1(\hat{\theta}_2) + R_n + O_p(n^{-3/2})$, $R_n = O_p(n^{-1})$. Then*

$$\hat{\theta}_2 = \hat{\theta}_1 - M^{-1}R_n + O_p(n^{-3/2}).$$

Proof. A mean-value expansion gives $0 = \hat{m}_2(\hat{\theta}_2) = \hat{m}_1(\hat{\theta}_1) = \hat{m}_1(\hat{\theta}_2) + [\partial\hat{m}_1(\bar{\theta})/\partial\theta'](\hat{\theta}_1 - \hat{\theta}_2)$, where $\bar{\theta}$ is the mean-value that actually differs from row to row. Note that by smoothness of a matrix inverse at any nonsingular point, $[\partial\hat{m}_1(\bar{\theta})/\partial\theta']^{-1} = M^{-1} + O_p(n^{-1/2})$. Then solving gives

$$\hat{\theta}_2 - \hat{\theta}_1 = -[\partial\hat{m}_1(\bar{\theta})/\partial\theta']^{-1}[\hat{m}_2(\hat{\theta}_2) - \hat{m}_1(\hat{\theta}_2)]$$
$$= -M^{-1}[R_n + O_p(n^{-3/2})] + O_p(n^{-1/2})[R_n + O_p(n^{-3/2})].$$

Q.E.D.

Proof of Theorem 3.2: Let $\hat{\theta}_1 = \hat{\theta}_{CUE}$ and $\hat{\theta}_2 = \hat{\theta}_{2S}$. Thus, $\hat{m}_1(\theta) = \hat{m}(\theta)$ is as specified in the proof of Theorem 3.1 and, from (3.5),

$$\hat{m}_2(\theta) = (\lambda'\hat{G}(\beta), \hat{g}(\beta)' + \lambda'\hat{\Omega}(\tilde{\beta}))'. \tag{A.11}$$

Let $\hat{g}_i = g_i(\hat{\beta}_{2S})$, $\hat{G}_i = G_i(\hat{\beta}_{2S})$ and note that

$$\hat{m}_2(\hat{\theta}_2) = \hat{m}_1(\hat{\theta}_2) + \begin{pmatrix} -\sum_{i=1}^n \hat{\lambda}_{2S}'\hat{g}_i\hat{G}_i'\hat{\lambda}_{2S}/n \\ [\hat{\Omega}(\tilde{\beta}) - \hat{\Omega}(\hat{\beta}_{2S})]\hat{\lambda}_{2S} \end{pmatrix}.$$

It follows similarly to previous results that $\sum_{i=1}^n \hat{g}_{ij}\hat{G}_i'/n = E[g_{ij}G_i]+O_p(n^{-1/2})$ and that $\hat{\lambda}_{2S} = \tilde{\lambda}+O_p(n^{-1})$, so that $\sum_{i=1}^n \hat{\lambda}_{2S}'\hat{g}_i\hat{G}_i'\hat{\lambda}_{2S}/n = \sum_{j=1}^m \tilde{\lambda}_j E[g_{ij}G_i']\tilde{\lambda}+ O_p(n^{-3/2})$. Similarly, it follows that $[\hat{\Omega}(\tilde{\beta}) - \hat{\Omega}(\hat{\beta})]\hat{\lambda}_{2S} = \sum_{j=1}^m \tilde{\lambda}_j\{E[g_{ij}G_i] + E[g_iG_{i,j\bullet}]\}(\tilde{\beta} - \hat{\beta}) + O_p(n^{-3/2})$. Therefore, hypothesis (c) of Lemma A.4 is satisfied for R_n as given in the statement of Theorem 3.2. It was shown in the proof of Theorem 3.1 that hypothesis (b) is satisfied, so the conclusion follows from Lemma A.3. Q.E.D.

Proof of Theorem 3.3: Let $\hat{\theta}_1$ be the CUE, $\hat{\theta}_2$ the GEL estimator, and $k = [\rho_v(0)/\rho_{vv}(0)]$. Note that they satisfy the first-order conditions $\hat{m}_j(\hat{\theta}_j) = 0$ for $\hat{m}_1(\theta)$ from the proof of Theorem 3.2, and

$$\hat{m}_2(\theta) = \sum_{i=1}^n \rho_v(k\lambda' g_i(\beta))[\lambda' G_i(\beta), g_i(\beta)']'/[\rho_v(0)n]. \tag{A.12}$$

Let $\hat{g}_i = g_i(\hat{\beta}_{GEL})$ and $\hat{G}_i = G_i(\hat{\beta}_{GEL})$. Expand $\hat{m}_2(\hat{\theta}_2)$ in $\hat{\lambda} = \hat{\lambda}_{GEL}$ to obtain

$$\begin{aligned} \hat{m}_2(\hat{\theta}_2) &= \hat{m}_1(\hat{\theta}_2) \\ &\quad +[k^2/2\rho_v(0)]\sum_{i=1}^n \rho_{vvv}(k\bar{\lambda}'\hat{g}_i)(\hat{\lambda}'\hat{g}_i)^2(\hat{\lambda}'\hat{G}_i, \hat{g}_i{}')'/n. \end{aligned} \tag{A.13}$$

Note that $\sum_{i=1}^n \|\hat{g}_i\|^2\|\hat{G}_i\|/n = O_p(1)$ using Hölder's inequality and Assumptions 4.1(d) and 4.4(b)(c). Also, by arguments similar to those of Lemma A.1, $\max_{i\leq n} |\rho_{vvv}(k\bar{\lambda}'\hat{g}_i)| = O_p(1)$. Therefore, as $\hat{\lambda} = O_p(n^{-1/2})$, it follows that $\sum_{i=1}^n \rho_{vvv}(k\bar{\lambda}'\hat{g}_i)(\hat{\lambda}'\hat{g}_i)^2\hat{G}_i'\hat{\lambda}/n = O_p(n^{-3/2})$. Similarly, by an additional expansion in $\bar{\lambda}$ and Assumption 4.4 (a) and (c),

$$\begin{aligned} & \|\sum_{i=1}^n [\rho_{vvv}(k\bar{\lambda}'\hat{g}_i) - \rho_{vvv}(0)](\hat{\lambda}'\hat{g}_i)^2\hat{g}_i/n\| \\ \leq & \ O_p(1)\|\hat{\lambda}\|^3 \sum_{i=1}^n \|\hat{g}_i\|^4/n = O_p(n^{-3/2}). \end{aligned}$$

Also, by an expansion in $\hat{\beta}_{GEL}$, Assumptions 4.1(d) and 4.4(b)(c)(d), and the CLT, for each j, we have $\sum_{i=1}^n \hat{g}_i\hat{g}_i'\hat{g}_{ij}/n = E[g_ig_i'g_{ij}] + O_p(n^{-1/2})$. Noting that the first-order conditions for $\hat{\lambda}$ can be expanded and solved to obtain $\hat{\lambda} = [\sum_{i=1}^n \rho_{vv}(k\bar{\lambda}'\hat{g}_i)\hat{g}_i\hat{g}_i'/n\rho_{vv}(0)]^{-1}\hat{g}(\hat{\beta}_{GEL})$, analogous arguments give $\hat{\lambda} = \tilde{\lambda} + O_p(n^{-1})$. Hence,

$$\begin{aligned} & [k^2/2\rho_v(0)] \sum_{i=1}^n \rho_{vvv}(k\bar{\lambda}'\hat{g}_i)(\hat{\lambda}'\hat{g}_i)^2\hat{g}_i/n \\ = & \ [\rho_v(0)\rho_{vvv}(0)/2\rho_{vv}(0)^2] \sum_{j=1}^m \tilde{\lambda}_j E[g_{ij}g_ig_i']\tilde{\lambda} + O_p(n^{-3/2}), \end{aligned}$$

with the first term $O_p(n^{-1})$. Thus hypothesis (c) of Lemma A.4 is satisfied for R_n as given in Theorem 3.3. Condition (b) of Lemma A.4 was proved in Theorem 3.1, so that the result follows by Lemma A.4. Q.E.D.

References

Brown, B.W., W.K. Newey and S. May (1997): “Efficient Bootstrapping for GMM”, *mimeo*, M.I.T.

Corcoran, S.A. (1998): “Bartlett Adjustment of Empirical Discrepancy Statistics”, *Biometrika*, 85, 967-972.

Cressie, N., and T. Read (1984): “Multinomial Goodness-of-Fit Tests”, *Journal of the Royal Statistical Society*, *Series B*, 46, 440-464.

Donald, S.G. and W.K. Newey (2000): ”A Jacknife Interpretation of the Continuous Updating Estimator,” *Economics Letters*, forthcoming.

Hansen, L. P. (1982): “Large Sample Properties of Generalized Method of Moments Estimators”, *Econometrica*, 50, 1029-1054.

Hansen, L.P., J. Heaton and A. Yaron (1996): “Finite-Sample Properties of Some Alternative GMM Estimators”, *Journal of Business and Economic Statistics*, 14, 262-280.

Imbens, G.W., R.H. Spady and P. Johnson (1998): ”Information Theoretic Approaches to Inference in Moment Condition Models,” *Econometrica,* 66, 333-357.

Kitamura, Y., and M. Stutzer (1997): “An Information-Theoretic Alternative to Generalized Method of Moments Estimation”, *Econometrica*, 65, 861-874.

Nagar, A.L. (1959) "The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations", *Econometrica* 27, 573-595.

Newey, W.K. and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in Engle, R. and D. McFadden, eds., *Handbook of Econometrics, Vol. 4,* New York: North Holland.

Qin, J. and Lawless, J. (1994): "Empirical Likelihood and General Estimating Equations", *Annals of Statistics*, 22, 300-325.

Rao, C.R. (1973): *Linear Statistical Inference and Its Applications*, New York: Wiley.

Robinson, P.M. (1988): "The Stochastic Difference Between Econometric Statistics," *Econometrica* 56, 531-548.

Rothenberg, T.J. (1983): "Asymptotic Properties of Some Estimators in Structural Models," in Karlin, S., T. Amemiya and L.A. Goodman, (eds.), *Studies in Econometrics, Time Series, and Multivariate Statistics.* New York: Academic Press.

Rothenberg, T.J (1984): "Approximating the Distributions of Econometric Estimators and Test Statistics," in Griliches, Z. and M.D. Intriligator, eds., *Handbook of Econometrics, Vol. 2*, New York: North-Holland.

Smith, R. J. (1997): “Alternative Semi-Parametric Likelihood Approaches to Generalized Method of Moments Estimation”, *Economic Journal*, 107, 503-519.

Smith, R. J. (1999): “Generalized Empirical Likelihood Methods for Moment Condition Models”, *mimeo*, University of Bristol. Revision of paper presented at Cowles Foundation Econometrics Conference on “New Developments in Time Series Econometrics”, Yale University, 1999.

Stock, J.H., and J. Wright (2000): ”GMM with Weak Identification”, *Econometrica*, forthcoming.