CHAPTER 10

The Performance of Empirical Likelihood and its Generalizations

Guido W. Imbens and Richard H. Spady

ABSTRACT

We calculate higher-order asymptotic biases and mean squared errors (MSE) for a simple model with a sequence of moment conditions. In this setup, generalized empirical likelihood (GEL) and infeasible optimal GMM (OGMM) have the same higher-order biases, with GEL apparently having an MSE that exceeds OGMM's by an additional term of order $(M-1)/N$, i.e., the degree of overidentification divided by sample size. In contrast, any two-step GMM estimator has an additional bias relative to OGMM of order $(M-1)/N$ and an additional MSE of order $(M-1)^2/N$. Consequently, GEL must be expected to dominate two-step GMM. In our simple model all GEL's have equivalent next higher order behavior because generalized third moments of moment conditions are assumed to be zero; we explore, in further analysis and simulations, the implications of dropping this assumption.

1. INTRODUCTION

This paper has two parts. In the first part, we calculate higher-order asymptotic biases and mean squared errors (MSE) for a simple model with a sequence of moment conditions. In this setup, generalized empirical likelihood (GEL) and infeasible optimal GMM (OGMM) have the same higher-order biases, with GEL having an MSE that *apparently* exceeds OGMM's by an additional term of order $(M-1)/N$, i.e., the degree of overidentification divided by sample size. In contrast, any two-step GMM estimator has an additional bias relative to OGMM of order $(M-1)/N$ and an additional MSE of order $(M-1)^2/N$. Although these features do depend on the simple framework we have adopted (and on the force of "apparently," cf. the discussion of Lemma 9) we cannot see how a more complicated framework will rescue two-step GMM (generalized method of moments) from these fundamental difficulties. Consequently, we conclude that GEL must be expected to dominate two-step GMM, and our interest shifts to distinguishing between variants of GEL and the closely related, (if not dual) empirical discrepancy (ED) estimators.

In our simple model all GEL's have equivalent next higher-order behavior because third moments of moment conditions (i.e., products of the form $\psi_j\psi_k\psi_\ell$, where any or all of j, k, and ℓ may be equal) are assumed to be zero. We explore, in further analysis and simulations, the implications of dropping this

assumption. Our analysis indicates that one variant of GEL/ED, which can be identified with continuously updated GMM, ignores third and higher-order cumulants of the moment conditions used in GMM estimation, effectively treating these cumulants as zero. Whether this is advantageous depends, heuristically, on whether higher-order cumulants of moment conditions can be usefully estimated to deviate from zero, which of course depends on sample size and the data generating process. We find that when third moments are unimportant, all variants of GEL/ED provide virtually identical performance in our simple experiments. However, in cases where there is substantial skew, continuously updated GMM is usually inferior to other variants, including exponential tilting (ET) and empirical likelihood (EL), at small and moderate sample sizes. In some of these cases, ET is superior to EL; and in no case is there substantial loss in applying ET/EL in preference to the continuously updated estimator (CUE).

In their recent unpublished work Newey and Smith (2002) have demonstrated that bias-corrected EL is second-order efficient within the class of bias-corrected GEL estimators. The line of argument in some ways parallels the argument that bias corrected parametric ML is second-order efficient in the parametric case. As such, it does not explicitly calculate the higher order approximations to the mean squared error that are computed here, albeit for special cases. These calculations help us frame and examine cases where we conjecture (correctly) that ET outperforms EL in small and moderately sized samples. The practical relevance of the analogy of empirical likelihood to parametric likelihood in higher-order asymptotic behavior is problematic: for example, although the EL likelihood ratio test (ELRT) is Bartlett correctable in the absence of nuisance parameters, the behaviors of the ELRT and its Bartlett correction do not resemble their parametric counterparts (cf. Corcoran, Davison, and Spady, 1995; Imbens, Spady, and Johnson, 1998). Consequently, we feel that caution is warranted in choosing between members of the GEL class when moment conditions have nonzero third and higher order cumulants.

2. FRAMEWORK

Consider a sequence of independent and identically distributed pairs of random vectors $\{(v_i, w_i)\}_{i=1}^N$. The dimension of v_i and w_i is $M \geq 1$. We are interested in a scalar parameter θ, satisfying

$$E[\psi(v_i, w_i, \theta)] = 0,$$

for $i = 1, \ldots, N$, where

$$\psi(v_i, w_i, \theta) = (v_i + e_1)\cdot\theta - w_i = \begin{pmatrix} (v_{i1}+1)\cdot\theta - w_{i1} \\ v_{i2}\cdot\theta - w_{i2} \\ \vdots \\ v_{iM}\cdot\theta - w_{iM} \end{pmatrix}.$$

and e_1 is an $M-$vector with the first element equal to one and the other elements equal to zero.

We are interested in the properties of various estimators for θ as the degree of overidentification $(M-1)$, increases. Following Donald and Newey (2004) who look at the behavior of various instrumental variables estimators as the number of instruments increases, and Newey and Smith (2001) who look at bias of GEL and GMM estimators, we look at the leading terms in the asymptotic expansion of the estimators and consider the rate at which the moments of these terms increase with M.

We make the following simplifying assumptions. The pairs (v_{im}, w_{im}) and (v_{jn}, w_{jn}) are independent if either $i \neq j$ or $n \neq m$ (or both), and have the same distribution. Let $\mu_{rp} = E[v_{im}^r \cdot w_{im}^p]$ denote the moments of this distribution. Moments up to order $p+r \leq 6$ are assumed to be finite. Without essential loss of generality, let $\mu_{10} = \mu_{01} = 0$, implying the true value of θ is $\theta^* = 0$, let $\mu_{20} = \mu_{02} = 1$, and let $\mu_{11} = \rho$ be the correlation coefficient of v_{im} and w_{im}.

With these assumptions the system of moment conditions, in fact, contains no identifying information after the first moment, although this would not be known to an investigator. Since in general a system of M moment conditions being used to estimate a scalar parameter can be renormalized to a system with one efficient moment condition that depends on the parameter of interest and $(M-1)$ moment conditions that are uncorrelated with it, the above system models the situation in which successive moment conditions are increasingly less informative. One purpose of our analytical investigation is to demonstrate that some estimators are better able to resist the deterioration of efficiency caused by the addition of irrelevant moment conditions.

Let

$$\overline{v} = \frac{1}{N}\sum_{i=1}^{N} v_i,$$

$$\overline{w} = \frac{1}{N}\sum_{i=1}^{N} w_i,$$

$$\overline{ww'} = \frac{1}{N}\sum_{i=1}^{N} w_i w_i',$$

denote sample averages, let $\overline{v}_j$ and $\overline{w}_j$ denote the jth element of $\overline{v}$ and $\overline{w}$, respectively, and let $\overline{ww'}_{ij}$ denote the (i, j)th element of $\overline{ww'}$.

Denote the optimal, infeasible, GMM estimator by

$$\hat{\theta}_{\text{opt}} = \overline{w}_1/(1+\overline{v}_1).$$

This is the estimator based on using the optimal linear combination of the moments. Since only the first moment is informative, this implies using only the first moment, thus estimating θ by solving

$$\sum_{i=1}^{N}(v_{i1}+1)\cdot\theta - w_{i1}.$$

This estimator is not feasible because the researcher does not know the optimal linear combination of the moments, but it provides a useful benchmark against which to judge feasible estimators. Note that increasing M does not affect this estimator as all the additional moments are ignored.

First we expand this estimator up to terms of order $O_p(N^{-3/2})$.

Lemma 2.1. (Expansion Optimal GMM Estimator)

$$\hat{\theta}_{\text{opt}} = \overline{w}_1 - \overline{w}_1\overline{v}_1 + \overline{w}_1\overline{v}_1^2 + o_p(N^{-3/2}).$$

Proof. See Appendix.

Define

(i) $R^{\text{opt}} = \overline{w}_1 = O_p(N^{-1/2})$,
(ii) $S^{\text{opt}} = -\overline{w}_1\overline{v}_1 = O_p(N^{-1})$,
(iii) $T^{\text{opt}} = \overline{w}_1\overline{v}_1^2 = O_p(N^{-3/2})$,

so that $\hat{\theta}_{\text{opt}} = R^{\text{opt}} + S^{\text{opt}} + T^{\text{opt}} + o_p(N^{-3/2})$.

Lemma 2.2. (Bias of $\hat{\theta}_{\text{opt}}$)
The bias of the leading terms is

$$E[R^{\text{opt}} + S^{\text{opt}} + T^{\text{opt}} - \theta^*] = -\rho/N + \mu_{21}/N^2.$$

Proof. See Appendix.

Lemma 2.3. (Mean Squared Error of $\hat{\theta}_{\text{opt}}$)
The mean squared error of the leading terms is

$$E[(R^{\text{opt}} + S^{\text{opt}} + T^{\text{opt}} - \theta^*)^2] = 1/N - 2\mu_{12}/N^2 + 3(2\rho^2+1)/N^2 + o(1/N^2).$$

Proof. See Appendix.

3. TWO-STEP GMM ESTIMATOR

The first estimator we consider is the standard two-step generalized method of moments (GMM) estimator, due to Hansen (1982). Consider a generic GMM estimator, defined as the minimand of

$$\left(\frac{1}{N}\sum_{i=1}^{N}\psi(v_i, w_i, \theta)\right)' \cdot C \cdot \left(\frac{1}{N}\sum_{i=1}^{N}\psi(v_i, w_i, \theta)\right).$$

We focus here on the efficient GMM estimator, with the choice for the weight matrix C equal to

$$\left(\frac{1}{N}\sum_{i=1}^{N}\psi(v_i, w_i, \theta^*)\cdot\psi(v_i, w_i, \theta^*)'\right)^{-1} = (\overline{ww'})^{-1},$$

so that the GMM "weight" matrix is estimated at the true value of θ. Thus, the GMM objective function is

$$\left(\frac{1}{N}\sum_{i=1}^{N}\psi(v_i, w_i, \theta)\right)' \cdot (\overline{ww'})^{-1} \cdot \left(\frac{1}{N}\sum_{i=1}^{N}\psi(v_i, w_i, \theta)\right)$$
$$= ((\overline{v}+e_1)\cdot\theta - \overline{w})' \cdot (\overline{ww'})^{-1} \cdot ((\overline{v}+e_1)\cdot\theta - \overline{w}).$$

The first order condition for the GMM estimator is

$$0 = 2(\overline{v}+e_1)' \cdot (\overline{ww'})^{-1} \cdot ((\overline{v}+e_1)\cdot\theta - \overline{w}),$$

with the solution for the GMM estimator equal to

$$\hat{\theta}_{\text{gmm}} = \left((\overline{v}+e_1)' \cdot (\overline{ww'})^{-1} \cdot (\overline{v}+e_1)\right)^{-1} \cdot \left((\overline{v}+e_1)' \cdot (\overline{ww'})^{-1} \cdot \overline{w}\right).$$

The goal is to approximate this estimator up to terms of order $O_p(1/N)$ and evaluate the mean squared error of this approximation. In particular, the terms whose moments depend on M are of interest, and specifically how fast the mean squared error increases with the number of excess moments.

Lemma 3.4. (EXPANSION OF $\hat{\theta}_{\text{gmm}}$)

$$\begin{aligned}\hat{\theta}_{\text{gmm}} = {} & \overline{w}_1 - 2\overline{v}_1\overline{w}_1 + (\overline{ww'}_{11} - 1)\cdot\overline{w}_1 - e_1'(\overline{ww'} - \mathcal{I}_M)\overline{w} + \overline{v}'\overline{w} \\ & - \overline{v}(\overline{ww'} - \mathcal{I}_M)\overline{w} + e_1(\overline{ww'} - \mathcal{I}_M)(\overline{ww'} - \mathcal{I}_M)\overline{w} - 2\overline{v}_1\overline{v}'\overline{w} \\ & + 2\overline{v}_1 e_1'(\overline{ww'} - \mathcal{I}_M)\overline{w} + 2\overline{w}_1 e_1(\overline{ww'} - \mathcal{I}_M)\overline{v} \\ & - \overline{w}_1 e_1'(\overline{ww'} - \mathcal{I}_M)(\overline{ww'} - \mathcal{I}_M)e_1 + 4\overline{v}_1^2\overline{w}_1 \\ & + \overline{w}_1(\overline{ww'}_{11} - 1)(\overline{ww'}_{11} - 1) - 4\overline{v}_1\overline{w}_1(\overline{ww'}_{11} - 1) \\ & + (\overline{ww'}_{11} - 1)\overline{v}'\overline{w} - (\overline{ww'}_{11} - 1)e_1'(\overline{ww'} - \mathcal{I}_M)\overline{w} \\ & - \overline{w}_1\overline{v}'\overline{v} + o_p\left(N^{-3/2}\right).\end{aligned}$$

Proof. See Appendix.

Now define

(i) $R^{\text{gmm}} = \overline{w}_1 = O_p(N^{-1/2})$,
(ii) $S^{\text{gmm}} = -2\overline{v}_1\overline{w}_1 + (\overline{ww'}_{11} - 1)\cdot\overline{w}_1 - e_1'\left(\overline{ww'} - \mathcal{I}_M\right)\overline{w} + \overline{v}'\overline{w} = O_p(N^{-1})$,
(iii) $T^{\text{gmm}} = -\overline{v}(\overline{ww'} - \mathcal{I}_M)\overline{w} + e_1(\overline{ww'} - \mathcal{I}_M)(\overline{ww'} - \mathcal{I}_M)\overline{w} - 2\overline{v}_1\overline{v}'\overline{w} + 2\overline{v}_1 e_1'(\overline{ww'} - \mathcal{I}_M)\overline{w} + 2\overline{w}_1 e_1(\overline{ww'} - \mathcal{I}_M)\overline{v} - \overline{w}_1 e_1'(\overline{ww'} - \mathcal{I}_M)(\overline{ww'} - \mathcal{I}_M)e_1 + 4\overline{v}_1^2\overline{w}_1 + \overline{w}_1(\overline{ww'}_{11} - 1) \times (\overline{ww'}_{11} - 1) - 4\overline{v}_1\overline{w}_1(\overline{ww'}_{11} - 1) + (\overline{ww'}_{11} - 1)\overline{v}'\overline{w} - (\overline{ww'}_{11} - 1) e_1'(\overline{ww'} - \mathcal{I}_M)\overline{w} - \overline{w}_1\overline{v}'\overline{v} = O_p(N^{-3/2})$, so that

$$\hat{\theta}_{\text{gmm}} = R^{\text{gmm}} + S^{\text{gmm}} + T^{\text{gmm}} + o_p\left(N^{-3/2}\right).$$

For the bias of the GMM estimator we therefore investigate the moments of $R^{\text{gmm}} + S^{\text{gmm}} - \theta^*$.

Lemma 3.5. (BIAS OF $\hat{\theta}_{\text{gmm}}$)
The expectation of the leading terms is

$$E[R^{\text{gmm}} + S^{\text{gmm}} - \theta^*] = -\rho/N + \rho(M-1)/N + o(N^{-1}).$$

Proof. See Appendix.

Lemma 3.6. (MEAN SQUARED ERROR OF $\hat{\theta}_{\text{gmm}}$)
The mean squared error of the leading terms is

$$\begin{aligned}E&\left[(R^{\text{gmm}} + S^{\text{gmm}} + T^{\text{gmm}} - \theta^*)^2\right] \\ &= 1/N - 2\mu_{12}/N^2 + 3(2\rho^2+1)/N^2 + \rho^2(M-1)^2/N^2 \\ &\quad + (M-1)(3\rho^2 + 1 + 2\rho\mu_{03})/N^2 + o(1/N^2).\end{aligned}$$

Proof. See Appendix.

Note that the difference between the MSE for $\hat{\theta}_{\text{gmm}}$ and $\hat{\theta}_{\text{opt}}$ is in the last two terms. The first of these is proportional to $(M-1)^2$ and is the reason for the poor performance of the two-step GMM estimator when the degree of overidentification is high. If $M = 1$, the two extra terms vanish as the optimal GMM estimator and feasible GMM estimator coincide.

4. GENERALIZED EMPIRICAL LIKELIHOOD ESTIMATORS

In this section we consider alternatives to the standard two-step GMM estimators. The estimators considered include empirical likelihood (Qin and Lawless, 1994; Imbens, 1997), exponential tilting (Imbens et al., 1998;

Kitamura and Stutzer, 1997; Rothenberg, 1999), and the continuously updating estimator (Hansen, Heaton, and Yaron, 1996). The specific class of estimators we consider is related to that of the Cressie–Read family (cf. Baggerly, 1998; Corcoran, 1995), as well to the generalized empirical likelihood estimators, introduced by Smith (1997). For a given function $g(a)$, normalized to satisfy $g(0) = 1$, $g'(0) = 1$, and $g''(0) = \lambda$, the estimator for θ is defined through the system of equations

$$0 = \sum_{i=1}^{N} \psi(v_i, w_i, \theta) \cdot g(t'\psi(v_i, w_i, \theta)),$$

$$0 = \sum_{i=1}^{N} t' \frac{\partial \psi}{\partial \theta'}(v_i, w_i, \theta) \cdot g(t'\psi(v_i, w_i, \theta)),$$

solved as a function of θ and t. The leading choices for $g(a)$ are $g(a) = 1/(1-a)$ (empirical likelihood), $g(a) = \exp(a)$ (exponential tilting), and $g(a) = 1 + a$ (continuously updating).

Under standard conditions, the solution for t, denoted by $\hat{t}_g$ converges to a vector of zeros; $\hat{\theta}_g$, the solution for θ, converges to θ^*, and

$$\hat{t}_g = O_p(1/\sqrt{N}),$$

$$\hat{\theta}_g = O_p(1/\sqrt{N}).$$

The choice of $g(a)$ does not matter for the standard large sample distribution, and

$$\hat{t}_{g_1} - \hat{t}_{g_2} = o_p(1/\sqrt{N}),$$

$$\hat{\theta}_g - \hat{\theta}_{\text{opt}} = o_p(1/\sqrt{N}).$$

Lemma 4.7. (EXPANSION FOR $\hat{\theta}_g$)

$$\hat{\theta}_g = \overline{w}_1 + (\overline{ww'}_{11} - 1)\overline{w}_1 - e_1'(\overline{ww'} - \mathcal{I}_M)\overline{w}$$
$$+ \overline{w}'\overline{v} - 2\overline{w}_1\overline{v}_1 - \rho\overline{w}'\overline{w} + \rho\overline{w}_1^2 + o_p(1/N).$$

Proof. See Appendix.

Note that the choice of g in the family of generalized empirical likelihood estimators does not matter for the $O(1/N^2)$ term. This is special to our case. It relies on the fact that the first and other moments are independent. In general, with a scalar parameter one can always renormalize the moments in such a way that only the derivative of the first moment depends on the parameter of interest, and that in addition the other moments are uncorrelated with the first one. This, however, does not make the first and other moments independent and the equivalence result here depends on the cross moments of the type $E[\psi_1\psi_2^2]$ being equal to zero.

Now define

(i) $T_1 = \overline{w}_1$,
(ii) $R_1 = -2\overline{v}_1\overline{w}_1$,
(iii) $R_2 = \overline{w}_1(\overline{ww'}_{11} - 1)$,
(iv) $R_3 = -\overline{w}'(\overline{ww'} - \mathcal{I}_M)e_1$,
(v) $R_4 = \overline{w}'\overline{v}$,
(vi) $R_5 = \rho\overline{w}_1^2$,
(vii) $R_6 = -\rho\overline{w}'\overline{w}$.

Lemma 4.8. (BIAS OF $\hat{\theta}_g$)
The expectation of the leading terms is

$$E[T_1 + R_1 + R_2 + R_3 + R_4 + R_5 + R_6 - \theta^*] = -\rho/N + o_p(1/N).$$

Proof. See Appendix.

Lemma 4.9. (MEAN SQUARED ERROR OF $\hat{\theta}_g$)
The mean squared error due to terms of magnitude $O_p(N^{-1})$ and greater (i.e., those given explicitly in Lemma 4.7) are:

$$E\left[(T_1 + R_1 + R_2 + R_3 + R_4 + r_5 + R_6 - \theta^*)^2\right]$$
$$= 1/N - 2\mu_{12}/N^2 + (1 + 2\rho^2)/N^2$$
$$+ \rho^2(M-1)/N^2 + 2(M-1)/N^2 + o_p(1/N^2).$$

Proof. See Appendix.

Lemmas 2.3 and 3.6 show that the MSE of feasible GMM exceeds that of OGMM by a term that is of order $(M-1)^2/N^2$. This term is due to an order $(M-1)/N$ bias in feasible GMM that is not in OGMM. This term is not present in the bias of GEL either. Lemma 4.9 differs from Lemmas 2.3 and 3.6 by not considering $O_p(N^{-3/2}) \cdot O_p(N^{-1/2}) = O_p(N^{-2})$ terms as they arise from the expansion of the estimator; but these terms do not give rise to the key $(M-1)^2/N^2$ order term in Lemma 3.6, which is the product of two $O_p(N^{-1})$ terms that are fully reflected in Lemma 4.9.

Consequently, we conclude that feasible GMM acquires MSE at rate $(M-1)^2/N^2$. This is basically due to the fact that the bias of feasible GMM grows at rate $(M-1)/N$, GEL's bias does not. This is not an artifact of our special situation and is consistent with the more general bias argument in Newey and Smith (2000).

5. EMPIRICAL DISCREPANCY THEORY

Having argued that generalized empirical likelihood offers "asymptotic resistance" to the deterioration of estimation efficiency as moment conditions are added, we turn to analyzing *differences* between members of this class. To do this, we interpret these estimators from the point of view of empirical discrepancy (ED, also sometimes called minimum discrepancy) theory, as found in the statistics literature in Corcoran (1995, 1998) and Baggerly (1998). In this section we show that the probabilities or reweighting implicit in GEL estimators can be expressed as functions of dot or inner products of Lagrange multipliers with moment conditions. In the following section we show that leading GEL estimators can be explicitly characterized as having reweighting functions that are polynomials in these dot products. Consequently, different GEL estimators weigh the higher-order moments of moment conditions differently, affecting their stability and ability to incorporate such higher-order information into estimation and inference.

We modify and extend some of the previous notation in order to deal with this more general context. A random variable z is i.i.d. according to $F(\cdot)$ and we have a sample $z_1, z_2, \dots, z_n$. In addition, for unknown θ of dimension k there is a (known) function $\psi(z, \theta)$ such that $\operatorname{E}\psi(z, \theta) = 0$. $\psi(z, \theta)$ is of dimension $m \geq k$. Empirical discrepancy theory considers choosing θ and probabilities $p_1, \dots, p_n$ on each of the data points such that

$$\sum_{i=1}^{n} h\left(p_i, \frac{1}{n}\right) \text{ is minimized subject to } \operatorname*{E}_{p}\psi(z, \theta) = 0 \text{ and}$$

$$\sum p_i = 1$$

where $h(\cdot, \cdot)$ is a measure of the discrepancy between two discrete measures, with the property that $h(\frac{1}{n}, \frac{1}{n}) = 0$; there are also some technical conditions on $h(\cdot, \cdot)$'s partial derivative with respect to its first argument.

Thus, empirical discrepancy theory chooses θ and a reweighting of the data so that the moment conditions hold and a discrepancy measure is minimized.

$$Q(\theta, p) = \sum_{i=1}^{n} h\left(p_i, \frac{1}{n}\right) + \alpha\left(\sum p_i - 1\right) + t\sum_{i=1}^{n} p_i\psi_i(\theta). \qquad (5.1)$$

Consider the determination of p first:

$$\frac{\partial Q(\theta, p)}{\partial p_i} = \frac{\partial h}{\partial p_i} + \alpha + t\psi_i(\theta) = 0 \qquad (5.2)$$

$$\sum_{i=1}^{n}\left\{\frac{\partial Q(\theta, p)}{\partial p_i}p_i\right\} = \sum\frac{\partial h}{\partial p_i}p_i + \alpha\sum p_i + t\sum_{i=1}^{n} p_i\psi_i(\theta)$$

$$= \sum\frac{\partial h}{\partial p_i}p_i + \alpha \qquad + 0$$

So $\alpha = -\sum\frac{\partial h}{\partial p_i}p_i$; substituting into (5.2):

$$\frac{\partial Q(\theta, p)}{\partial p_i} = \frac{\partial h}{\partial p_i} - \sum\frac{\partial h}{\partial p_i}p_i + t\psi_i(\theta) = 0.$$

Note:

$$\frac{\partial h}{\partial p_i} = -t\psi_i(\theta) \text{ is a solution.}$$

Note that t is an m-dimensional Lagrange multiplier of the original problem.

Remaining with the problem of constructing p for given θ, there are three common choices for $h(\cdot, \cdot)$:

- $h(p_i, \frac{1}{n}) = p_i(p_i - \frac{1}{n})$, or effectively $\sum h(p_i, \frac{1}{n}) = \sum p_i^2$, in which case $p_i = k(1 + t\psi_i(\theta))$ and $t = -(\sum\psi_i\psi_i')^{-1}(\sum\psi_i)$; this is often called Euclidean likelihood.
- $h(p_i, \frac{1}{n}) = \frac{1}{n}\{\log(\frac{1}{n}) - \log p_i\}$, or effectively $-\sum h(p_i, \frac{1}{n}) = \sum\log p_i$; $p_i = k\frac{1}{1+t\psi_i(\theta)}$; this is Owen's (1988) empirical likelihood (EL).
- $h(p_i, \frac{1}{n}) = p_i\{\log(\frac{1}{n}) - \log p_i\}$, or effectively $-\sum h(p_i, \frac{1}{n}) = \sum p_i\log p_i$; $p_i = ke^{t\cdot\psi_i(\theta)}$; this is called exponential tilting (ET).

Empirical likelihood and exponential tilting exchange the role of the empirical measure and the measure p that is under construction: ET finds the p to which the empirical measure is "KLIC-closest," while EL finds the p that is KLIC-closest to the empirical measure. Thus, ET "imagines" that the data generating process is p, (which obeys $E(\psi) = 0$) while EL imagines the DGP as a repetition of the observed data, which does not obey the specified moment conditions. To us, this suggests ET should be superior to EL. But EL has some higher-order asymptotic properties that mimic those of parametric likelihood, such as Bartlett correctability of its likelihood ratio test (but only when there are no nuisance parameters) and higher-order efficiency (when bias corrected) within the bias-corrected GEL class.

The preceding three cases are all members of the Cressie–Read (Cressie and Read, 1984) family, with

$$h\left(p_i, \frac{1}{n}\right) = \left(\frac{p_i}{1/n}\right)^{\lambda} - 1$$

$$p_i = k\left(\frac{1}{1 + t\psi_i(\theta)}\right)^{1/(\lambda+1)}$$

for $\lambda \in [-2, 1]$ so that $\lambda = -2$ is Euclidean likelihood, $\lambda = -1$ is ET, and $\lambda = 0$ is EL.

Turning now to the problem of estimating θ, the minimum discrepancy estimate is obtained by differentiating (5.1) with respect to θ to obtain:

$$\frac{\partial Q(\theta, p)}{\partial \theta} = t \sum_{i=1}^{n} p_i \frac{\partial \psi_i(\theta)}{\partial \theta} = 0,$$

a system of equations in k elements of θ. Thus, the entire system of $(m + k)$ equations can be written simply as:

$$\mathop{\mathrm{E}}_{p}\, \psi(\theta) = 0 \qquad (m \text{ equations}) \tag{5.3a}$$

$$t \cdot \mathop{\mathrm{E}}_{p} \frac{\partial \psi(\theta)}{\partial \theta} = 0 \qquad (k \text{ equations}). \tag{5.3b}$$

One way to think of these equations is that, having fixed θ and a formula for p (by choice of $h(\cdot, \cdot)$), the first m equations determine t. Similarly, for a fixed t and p, the remaining k equations determine θ.[1]

The duality of GEL and ED is examined in two papers of Newey and Smith (2000, 2001). Writing the GEL estimator as:

$$\hat{\theta}_{\mathrm{GEL}} = \arg\min_{\theta \in \Theta} \sup_{t \in T} n^{-1} g(t \cdot \psi_i(\theta))$$

the GEL estimator's estimating equations coincide with (5.3a) in cases where the derivative of $g(\cdot)$, denoted g', can be interpreted as being proportional to a probability. This can be done for the three cases under consideration here, as well as for all members of the Cressie–Read family. Newey and Smith (2000, 2001) show that for Euclidean likelihood, $g(t\psi_i) = -t\psi_i - (t\psi_i)^2/2$ and the resulting GEL estimator coincides with the continuously updated GMM estimator of Hansen et al. (1996). Consequently, we will denote the three estimators as $\hat{\theta}_{\mathrm{CUE}}$, $\hat{\theta}_{\mathrm{ET}}$, and $\hat{\theta}_{\mathrm{EL}}$.

6. A FURTHER CHARACTERIZATION OF ED/GEL ESTIMATORS

Rewriting the first equation of system (5.3a) as

$$\sum_{i=1}^{n} p(t \cdot \psi_i(\theta))\psi_i(\theta) = 0 \tag{6.4}$$

[1] This schema cannot be used to define a simple iterative procedure to compute θ, for in fact the saddlepoint nature of these equations makes the naive iterative procedure of (1) fix θ; (2) calculate t; (3) calculate new θ; unstable in a neighborhood of the solution θ_* of $\theta_* - \theta(t(\theta_*)) = 0$.

we can express the probabilities associated with CUE, ET, and EL (after absorbing some sign changes into k) as

$$\begin{aligned} p_i[\mathrm{CUE}] &= k_{\mathrm{CUE}}(1 + t\psi_i(\theta)) \\ p_i[\mathrm{ET}] &= k_{\mathrm{ET}}(e^{t\psi_i(\theta)}) \\ p_i[\mathrm{EL}] &= k_{\mathrm{EL}}\left(\frac{1}{1 - t\psi_i(\theta)}\right). \end{aligned}$$

Taking a Taylor series expansion of $p_i[ET]$ we can define a sequence of p functions:

$$\begin{aligned} p_i[\mathrm{ET}, 1] &= k_{\mathrm{ET},1}(1 + t\psi_i(\theta)) = p_i[\mathrm{CUE}] \\ p_i[\mathrm{ET}, 2] &= k_{\mathrm{ET},2}\left(1 + t\psi_i(\theta) + \frac{(t\psi_i(\theta))^2}{2}\right) \\ p_i[\mathrm{ET}, 3] &= k_{\mathrm{ET},3}\left(1 + t\psi_i(\theta) + \frac{(t\psi_i(\theta))^2}{2} + \frac{(t\psi_i(\theta))^3}{6}\right) \\ &\vdots \\ p_i[\mathrm{ET}, \infty] &= k_{\mathrm{ET},\infty}\left(1 + t\psi_i(\theta) + \frac{(t\psi_i(\theta))^2}{2} + \frac{(t\psi_i(\theta))^3}{6} + \dots\right) \\ &\quad k_{\infty} e^{t \cdot \psi_i(\theta)} = p_i[\mathrm{ET}]. \end{aligned}$$

And similarly for $p_i[\mathrm{EL}]$ we have

$$\begin{aligned} p_i[\mathrm{EL}, 1] &= k_{\mathrm{EL},1}(1 + t\psi_i(\theta)) = p_i[\mathrm{CUE}] \\ p_i[\mathrm{EL}, 2] &= k_{\mathrm{EL},2}(1 + t\psi_i(\theta) + (t\psi_i(\theta))^2) \\ p_i[\mathrm{EL}, 3] &= k_{\mathrm{EL},3}(1 + t\psi_i(\theta) + (t\psi_i(\theta))^2 + (t\psi_i(\theta))^3) \\ &\vdots \\ p_i[\mathrm{EL}, \infty] &= k_{\mathrm{EL},\infty}(1 + t\psi_i(\theta) + (t\psi_i(\theta))^2 + (t\psi_i(\theta))^3 + \dots) \\ &= k_{\mathrm{EL},\infty}\left(\tfrac{1}{1 - t\psi_i(\theta)}\right) = p_i[\mathrm{EL}]. \end{aligned}$$

Thus, all three p functions have the same first-order Taylor series expansion, coinciding exactly with $p[\mathrm{CUE}]$. Then $p[\mathrm{ET}]$ and $p[\mathrm{EL}]$ include higher powers of $(t \cdot \psi_i)$, the former having factorially declining weights or coefficients and the latter the coefficients $\{1, 1, \dots 1\}$. Since t is an $O_p(n^{-1/2})$ object, the difference in the treatment of t^2 terms induces differences of $O_p(n^{-1})$ in $\hat{\theta}_{\mathrm{CUE}}$, $\hat{\theta}_{\mathrm{ET}}$, and $\hat{\theta}_{\mathrm{EL}}$ and consequently their MSE behavior differs at $O_p(n^{-2})$. (This will be true for all members of the Cressie–Read family.[2])

[2] The Cressie–Read expansion is $p[\mathrm{CR}] = 1 + (t \cdot \psi_i)/(1 + \lambda) + (2 + \lambda)(t \cdot \psi_i)^2/2(1 + \lambda)^2 + (2 + \lambda)(3 + 2\lambda)(t \cdot \psi_i)^3/6(1 + \lambda)^3 + \cdots$

To see the effect of these differences, let us consider the difference between the first two elements of the sequence of ET functions for the equation setting the expectation of the jth component of ψ:

$$\sum_{i=1}^{n} p(t \cdot \psi_i(\theta))\psi_{ij}(\theta) = 0 \tag{6.5}$$

$$\sum_{i=1}^{n} k_{\mathrm{ET},1}(1 + t\psi_i(\theta))\psi_{ij}(\theta) = 0 \tag{6.6}$$

$$\sum_{i=1}^{n} k_{\mathrm{ET},2}\left(1 + t\psi_i(\theta) + \frac{(t\psi_i(\theta))^2}{2}\right)\psi_{ij}(\theta) = 0 \tag{6.7}$$

Supressing the i subscript momentarily, the extra terms in (6.7) (relative to (6.6)) are of the form:

$$.5 * (t_1\psi_1 + t_2\psi_2 + \cdots t_m\psi_m)^2\psi_j,$$

so that sums of these involve third moments of ψ. Consequently, in problems where third moments are zero, notably those in which ψ is symmetric, these terms will be converging rapidly to zero and thus have no effect even at $O_p(n^{-2})$.

7. A DETAILED ANALYSIS OF SOME SIMPLE EXAMPLES

To examine further the relation between the choice of a GEL/ED and the higher order moments of the underlying data, consider the estimation of the scalar parameter θ from a scalar random variable x where it is known that x has mean θ and variance 2θ. Thus $\psi(x, \theta)$ is given by:

$$\begin{aligned} x - \theta &= 0 \\ x^2 - \theta^2 - 2\theta &= 0 \end{aligned} \tag{7.8}$$

Writing the second moment condition in the way indicated (rather than $(x-\theta)^2 - 2\theta = 0$) does not change the numerical values of the resulting estimates of θ, but it does simplify $\frac{\partial\psi}{\partial\theta}$ to:

$$\frac{\partial\psi}{\partial\theta} = \begin{cases} -1 \\ -2\theta - 2 \end{cases}$$

Consequently, $\frac{\partial\psi}{\partial\theta}$ does not depend on the data so $\mathrm{E}\,\frac{\partial\psi}{\partial\theta}$ does not depend on p. Using (5.3b) this means $\hat{\theta}$ can be determined from

$$t \cdot \underset{p}{\mathrm{E}}\frac{\partial\psi}{\partial\theta} = 0$$

$$t_1(-1) + t_2(-2\theta - 2) = 0$$

$$\theta = \frac{-t_1}{2t_2} - 1.$$

Table 10.1. *Symmetric distributions with $E(x) = 1$ and $V(x) = 2$ used in the Monte Carlo experiments*

Case	Distribution	First four cumulants
1	$N(1, 2)$	$\{1, 2, 0, 0\}$
2	Symmetric mixture of normals: $.5N(0, 1)\tilde{}.5N(2, 1)$	$\{1, 2, 0, -2\}$
3	$t(df = 4)$	$\{1, 2, 0, \infty\}$
4	Uniform (on $1 - \sqrt{6}, 1 + \sqrt{6}$)	$\{1, 2, 0, -4.8\}$

It is apparent that our three estimators will differ, in this special case, only in their choice of t. For the CUE estimator, $t = (\psi'\psi)^{-1}\overline{\psi}$, that is, the coefficients of the regression of a column of 1's on ψ, and so θ is determined by the fixed point of a function of five moment functions of ψ: the means of the two moment functions (expressed as functions of θ) and the corresponding three variances and covariances. CUE is thus committed to local (to θ) sufficiency of five statistics and will ignore, for example, differences in skew between elements of the sample space. In cases where skew is zero, we can expect the difference between CUE and ET or EL to be negligible, whereas for nonzero skew, we might expect ET and/or EL to prove superior to CUE, but only at sample sizes at which $O_p(n^{-2})$ effects are operative.

To demonstrate these effects, we construct several data-generating processes which satisfy the moment conditions in (7.8) but have different properties for their higher-order moments. For each case we compute MSE and bias, and do this for CUE, ET, and EL. In addition, for ET and EL we compute p according to successive terms in the relevant Taylor series expansion, so that $p_i[\mathrm{EL}, 1] = k_{\mathrm{EL},1}(1 + t\psi_i(\theta)) = p_i[\mathrm{CUE}]$, $p_i[\mathrm{EL}, 3] = k_{\mathrm{EL},3}(1 + t\psi_i(\theta) + (t\psi_i(\theta))^2 + (t\psi_i(\theta))^3)$, and so forth.

In this way we can see whether the advantages, if any, of ET and EL over CUE set in after taking into account only a relatively small number of additional higher moments, and similarly if differences between EL and ET require the full limiting case of including some information about all higher-order moments.

We consider eight data generating processes: the first four of these are symmetric distributions with mean 1 and variance 2; the second four are asymmetric, also with mean 1 and variance 2. A short description of these, together with the first four cumulants of the distributions, is given in the following two tables (Tables 10.1 and 10.2).

Appendices B and C (not included here, but available at http://www.faculty.econ.northwestern.edu/spady/imbens-spady) contain tables for all eight cases, tabulating the MSE and the bias, respectively, of the parameter θ (which in all cases is 1) for a variety of sample sizes and 6,000 replications. In each of the tables, we report the value of the MSE (Appendix B) or bias (Appendix C) in 6,000 replications for CUE, ET, and EL, together with the MSE's or bias of

Table 10.2. *Asymmetric distributions with* $E(x) = 1$ *and* $V(x) = 2$ *used in the Monte Carlo experiments*

Case	Distribution	First four cumulants
5	$\chi^2(1)$	{1, 2, 8, 48}
6	Asymmetric mixture of normals: $.25N(-1, 2/3)^{*}.75N(5/3, 2/3)$	{1, 2, −1.777, −1.185}
7	Lognormal: $\theta = 0, \sigma^2 = \log(3)$	{1, 2, 20, 624}
8	Inverse gaussian: $\mu = 1, \lambda = .5$	{1, 2, 24, 84.85}

Table 10.3. *Estimates of MSE of* $\hat{\theta}$ *and jackknifed standard error estimates from 6,000 simulations of system (7.8) with normal errors*

			Case 1: $N(1, 2)$ $n = 50$			
Taylor degree	1 = CUE	3	5	7	ET(Infinity)	Sample mean
MSE	0.020821	0.020702	0.020691	0.020688	0.020687	0.039753
s.e.	0.000366	0.000365	0.000365	0.000365	0.000365	0.000705
Taylor degree	1 = CUE	3	5	7	EL(Infinity)	Sample mean
MSE	0.020821	0.020675	0.020671	0.020656	0.020586	0.039753
s.e.	0.000366	0.000365	0.000366	0.000367	0.000367	0.000705
			Case 1: $N(1, 2)$ $n = 100$			
Taylor degree	1 = CUE	3	5	7	ET(Infinity)	Sample mean
MSE	0.010304	0.010256	0.010255	0.010254	0.010254	0.019945
s.e.	0.000189	0.000188	0.000188	0.000188	0.000188	0.000353
Taylor degree	1 = CUE	3	5	7	EL(Infinity)	Sample mean
MSE	0.010304	0.010235	0.010226	0.010224	0.010218	0.019945
s.e.	0.000189	0.000188	0.000188	0.000188	0.000188	0.000353

the estimators based on the Taylor series expansion of degrees 3, 5, and 7, and the simple mean. (CUE corresponds to a Taylor series expansion of degree 1, ET and EL to degree ∞). For each entry we report a jackknife estimate of the standard error, this is given in the row labeled 's.e.'.

The results from cases 1 through 4 are easily summarized; specimen results for MSE in case 1 at $n = 50$ and 100 are given in Table 10.3.

The main points are (1) there is no important difference in the MSE performance of any of the estimators; (2) the apparent asymptotic efficiency gain from exploiting the second moment condition (as measured by the ratio of the MSE of any of the GEL estimators to the MSE of the sample mean) is achieved by $n = 50$ in case 1 and within the sample sizes presented in Appendix B in other cases. EL and ET typically offer a (very) small improvement over CUE and there is no case where employing EL or ET presents any real cost relative to CUE. Thus, taking the higher-order moments into account, as do ET and EL, does not generate an unstable estimator in these cases. This is true even in case 3 (t with $df = 4$) where the fourth cumulant of the first moment condition is infinite.

Table 10.4. *Estimates of MSE of* $\hat{\theta}$ *from 6,000 simulations of system (7.8) with* $\chi^2(1)$ *errors*

			Case 5: $\chi^2(1)$ $n = 50$			
Taylor degreee	1 = CUE	3	5	7	ET(Infinity)	Sample mean
MSE	0.047774	0.041204	0.039950	0.039545	0.037596	0.040168
s.e.	0.001024	0.000918	0.000886	0.000870	0.000740	0.000752
Taylor degree	1 = CUE	3	5	7	EL(Infinity)	Sample mean
MSE	0.047774	0.041705	0.040883	0.040684	0.033915	0.040168
s.e.	0.001024	0.000948	0.000941	0.000941	0.000679	0.000752
			Case 5: $\chi^2(1)$ $n = 100$			
Taylor degree	1 = CUE	3	5	7	ET(Infinity)	Sample mean
MSE	0.018503	0.015828	0.015564	0.015521	0.015508	0.020042
s.e.	0.000400	0.000333	0.000321	0.000317	0.000315	0.000365
Taylor degree	1 = CUE	3	5	7	EL(Infinity)	Sample mean
MSE	0.018503	0.015656	0.015386	0.015327	0.014741	0.020042
s.e.	0.000400	0.000339	0.000332	0.000330	0.000294	0.000365
			Case 5: $\chi^2(1)$ $n = 800$			
Taylor degree	1 = CUE	3	5	7	ET(Infinity)	Sample mean
MSE	0.001827	0.001832	0.001832	0.001832	0.001833	0.002444
s.e.	0.000033	0.000033	0.000033	0.000033	0.000033	0.000045
Taylor degree	1 = CUE	3	5	7	EL(Infinity)	Sample mean
MSE	0.001827	0.001843	0.001845	0.001846	0.001848	0.002444
s.e.	0.000033	0.000033	0.000033	0.000033	0.000033	0.000045

More interesting and varied results are obtained in the presence of skew. For case 5 (Table 10.4), in which x is $\chi^2(1)$, we see for CUE at $n = 50$ that the effect of adding an additional moment is to produce an estimator that is worse than the sample mean; the MSE's of the estimators at Taylor degrees 3, 5, and 7 are about the same as the sample mean; and that ET and EL are better than their corresponding degree 7 estimators and also the sample mean; both of these effects are greater for EL than ET. The superiority of EL and ET to CUE (and of EL to ET) continues through some of the larger sample sizes, but by $n = 800$ this ranking has reversed itself, though the differences are now no longer greater than estimated standard errors.

Thus, in this example with skew, we find results in accord with our earlier conjecture that EL and ET can be expected to outperform CUE because the former reflect skew and higher moments in the construction of t (and thus in general in the distribution estimates embodied in p) in a way that CUE does not. These effects must eventually disappear as the sample size grows, because all the estimators in question reach the same GMM efficiency bound.

Case 6 (an asymmetric normal mixture) is unremarkable except for the fact that the second moment condition is extremely informative in this case: adding it reduces MSE for all the estimators to the (apparent) asymptotic relative bound of 0.25 times the sample mean's MSE.

Table 10.5. *Estimates of MSE of $\hat{\theta}$ from 6,000 simulations of system (7.8) with lognormal errors*

			Case 7: lognormal $n = 50$			
Taylor degree	1 = CUE	3	5	7	ET(Infinity)	Sample mean
MSE	0.068870	0.061752	0.058380	0.056165	0.033665	0.039825
s.e.	0.001650	0.001567	0.001517	0.001481	0.000601	0.001013
Taylor degree	1 = CUE	3	5	7	EL(Infinity)	Sample mean
MSE	0.068870	0.069696	0.072404	0.073894	0.041134	0.039825
s.e.	0.001650	0.001672	0.001699	0.001719	0.000668	0.001013
			Case 7: lognormal $n = 100$			
Taylor degree	1 = CUE	3	5	7	ET(Infinity)	Sample mean
MSE	0.022288	0.018707	0.017705	0.017130	0.015554	0.020098
s.e.	0.000623	0.000508	0.000470	0.000446	0.000310	0.000578
Taylor degree	1 = CUE	3	5	7	EL(Infinity)	Sample mean
MSE	0.022288	0.022046	0.023572	0.024439	0.022448	0.020098
s.e.	0.000623	0.000569	0.000574	0.000583	0.000378	0.000578
			Case 7: lognormal $n = 400$			
Taylor degree	1 = CUE	3	5	7	ET(Infinity)	Sample mean
MSE	0.003789	0.004283	0.004310	0.004309	0.004307	0.005090
s.e.	0.000071	0.000077	0.000077	0.000077	0.000077	0.000099
Taylor degree	1 = CUE	3	5	7	EL(Infinity)	Sample mean
MSE	0.003789	0.004713	0.004858	0.004904	0.004943	0.005090
s.e.	0.000071	0.000085	0.000089	0.000091	0.000093	0.000099

Cases 7 and 8, lognormality and the inverse Gaussian distribution, have greater skew and kurtosis than case 5, the $\chi^2(1)$ example. Results for MSE in these cases for $n = 50$, 100, and 400 are shown in Tables 10.5 and 10.6. As with case 5, at small sample sizes CUE does worse than the sample mean; ET does better than the sample mean for the lognormal case, as do both ET and EL in the inverse Gaussian case. Unlike the $\chi^2(1)$ case, at small and moderate sample sizes, ET outperforms EL. In addition, as the Taylor degree is expanded from CUE to ET, ET shows continuous improvement (at those small and moderate sample sizes in which ET outperforms CUE); this is not the case for EL in these two examples.

In Appendix C, we present biases for the cases considered here and in Appendix B. Quite notably, bias does not generally make a substantial difference to the MSE. This could perhaps be expected from the fact that the correlation between ψ and $\frac{\partial \psi}{\partial \theta}$ is zero in this example, because the latter does not depend on x. This suggests that the (sometimes erratic) effects typically seen in Appendix C are $O(n^{-2})$ or higher.

The cases where bias is most evident and potentially important to the MSE ranking of the estimators, given in Tables 10.7 and 10.8, occur when x is lognormal or inverse Gaussian; the biases tend to be largest for EL. In view of the result of Newey and Smith (2000) demonstrating the higher-order efficiency of bias-corrected EL, (and EL alone among the GEL class), it is interesting to note that even after bias-correction, EL continues to have a larger MSE than ET in these particular examples.

Table 10.6. *Estimates of MSE of $\hat{\theta}$ from 6,000 simulations of system (7.8) with inverse Gaussian errors*

			Case 8: inverse Gaussian $n = 50$			
Taylor degree	1 = CUE	3	5	7	ET(Infinity)	Sample mean
MSE	0.049724	0.043575	0.041270	0.039837	0.027857	0.038833
s.e.	0.001349	0.001271	0.001228	0.001197	0.000551	0.000821
Taylor degree	1 = CUE	3	5	7	EL(Infinity)	Sample mean
MSE	0.049724	0.048685	0.050107	0.051142	0.032708	0.038833
s.e.	0.001349	0.001361	0.001382	0.001399	0.000599	0.000821
			Case 8: inverse Gaussian $n = 100$			
Taylor degree	1 = CUE	3	5	7	ET(Infinity)	Sample mean
MSE	0.016042	0.013755	0.013263	0.013023	0.012420	0.019573
s.e.	0.000434	0.000356	0.000334	0.000321	0.000246	0.000387
Taylor degree	1 = CUE	3	5	7	EL(Infinity)	Sample mean
MSE	0.016042	0.015417	0.016079	0.016453	0.015510	0.019573
s.e.	0.000434	0.000392	0.000394	0.000398	0.000281	0.000387
			Case 8: inverse Gaussian $n = 400$			
Taylor degree	1 = CUE	3	5	7	ET(Infinity)	Sample mean
MSE	0.003160	0.003246	0.003247	0.003246	0.003243	0.004931
s.e.	0.000059	0.000059	0.000059	0.000059	0.000059	0.000092
Taylor degree	1 = CUE	3	5	7	EL(Infinity)	Sample mean
MSE	0.003160	0.003381	0.003419	0.003432	0.003428	0.004931
s.e.	0.000059	0.000062	0.000062	0.000063	0.000063	0.000092

8. SUMMARY

Higher-order asymptotic arguments suggest that GEL/ED/ "one-step efficient" estimates of overidentified moment models will prove superior to two-step GMM, since the MSE of two-step GMM grows at rate $O((M-1)^2/N^2)$ where $(M-1)$ is the degree of overidentification, whereas the GEL class apparently has (in the special case considered) an MSE that grows at rate $O((M-1)/N^2)$. Consequently, interest shifts to distinguishing between elements of the GEL family on the basis of estimation performance. With a simple argument and example, it appears that the simplest GEL variant, the continuously updated or Euclidean likelihood estimator, is dominated by the more elaborate ET and EL estimators. The difference between these two variants can be seen to lie in their treatment of third and higher-order moments of moment conditions, with EL weighing these more heavily than ET.

Table 10.7. *Estimates of bias of $\hat{\theta}$ from 6,000 simulations of system (7.8) with lognormal errors*

			Case 7: lognormal $n = 50$			
Taylor degree	1 = CUE	3	5	7	ET(Infinity)	Sample mean
Bias	−0.073549	−0.028002	−0.018381	−0.014462	0.011123	−0.000712
s.e.	0.003252	0.003188	0.003111	0.003054	0.002365	0.002577
Taylor degree	1 = CUE	3	5	7	EL(Infinity)	Sample mean
Bias	−0.073549	−0.023550	−0.012970	−0.009200	0.051964	−0.000712
s.e.	0.003252	0.003395	0.003470	0.003508	0.002531	0.002577
			Case 7: lognormal $n = 100$			
Taylor degree	1 = CUE	3	5	7	ET(Infinity)	Sample mean
Bias	−0.004507	0.034959	0.040531	0.041955	0.043940	0.002031
s.e.	0.001927	0.001707	0.001636	0.001601	0.001507	0.001830
Taylor degree	1 = CUE	3	5	7	EL(Infinity)	Sample mean
Bias	−0.004507	0.045751	0.057451	0.061930	0.080875	0.002031
s.e.	0.001927	0.001824	0.001838	0.001853	0.001628	0.001830
			Case 7: lognormal $n = 400$			
Taylor degree	1 = CUE	3	5	7	ET(Infinity)	Sample mean
Bias	0.023147	0.030086	0.030184	0.030119	0.030000	0.000490
s.e.	0.000736	0.000750	0.000753	0.000753	0.000754	0.000921
Taylor degree	1 = CUE	3	5	7	EL(Infinity)	Sample mean
Bias	0.023147	0.033090	0.033761	0.033928	0.032701	0.000490
s.e.	0.000736	0.000777	0.000787	0.000791	0.000804	0.000921

Table 10.8. *Estimates of bias of $\hat{\theta}$ from 6,000 simulations of system (7.8) with inverse Gaussian errors*

			Case 8: inverse Gaussian $n = 50$			
Taylor degree	1 = CUE	3	5	7	ET(Infinity)	Sample mean
Bias	−0.049814	−0.012106	−0.005470	−0.002999	0.009935	0.000036
s.e.	0.002806	0.002691	0.002622	0.002577	0.002151	0.002544
Taylor degree	1 = CUE	3	5	7	EL(Infinity)	Sample mean
Bias	−0.049814	−0.005333	0.004693	0.007965	0.046092	0.000036
s.e.	0.002806	0.002848	0.002889	0.002918	0.002258	0.002544
			Case 8: inverse Gaussian $n = 100$			
Taylor degree	1 = CUE	3	5	7	ET(Infinity)	Sample mean
Bias	−0.002652	0.022747	0.025227	0.025676	0.026322	−0.000678
s.e.	0.001635	0.001485	0.001451	0.001436	0.001398	0.001806
Taylor degree	1 = CUE	3	5	7	EL(Infinity)	Sample mean
Bias	−0.002652	0.032346	0.039543	0.042147	0.049622	−0.000678
s.e.	0.001635	0.001548	0.001556	0.001564	0.001475	0.001806
			Case 8: inverse Gaussian $n = 400$			
Taylor degree	1 = CUE	3	5	7	ET(Infinity)	Sample mean
Bias	0.009872	0.012438	0.012302	0.012226	0.012154	0.000467
s.e.	0.000714	0.000718	0.000718	0.000718	0.000718	0.000907
Taylor degree	1 = CUE	3	5	7	EL(Infinity)	Sample mean
Bias	0.009872	0.014237	0.014430	0.014455	0.013003	0.000467
s.e.	0.000714	0.000728	0.000732	0.000733	0.000737	0.000907

While EL has an array of higher-order theoretical properties that are in some ways similar to those of parametric likelihood, our analysis shows that it weighs higher-order moments of moment conditions much more heavily than does ET. Consequently, in contexts where these higher-order moments are likely both to be important and poorly defined in the sample sizes of interest, EL may prove to have a more erratic behavior than ET. This is borne out rather clearly in a few examples considered in this paper; this in turn suggests that no member of the GEL class will dominate the field unambiguously.

APPENDIX A

Complete details of the more mechanical aspects of the proofs can be found on http://www.nuff.ox.ac.uk/users/spady/imbens-spady.pdf.

Lemma A.10. (EXPANSION OF MATRIX INVERSION)
Let A, B, and C be $M \times M$ symmetric matrices of order $O_p(1)$, with A invertible. Then

(i) $(A + B/\sqrt{N})^{-1} = A^{-1} + o_p(1)$,
(ii) $(A + B/\sqrt{N})^{-1} = A^{-1} - A^{-1}BA^{-1}/\sqrt{N} + o_p(1/\sqrt{N})$,
(iii) $(A + B/\sqrt{N})^{-1} = A^{-1} - A^{-1}BA^{-1}/\sqrt{N} + A^{-1}BA^{-1}BA^{-1}/N + o_p(1/N)$,
(iv) $(A + B/\sqrt{N} + C/N)^{-1} = A^{-1} - A^{-1}BA^{-1}/\sqrt{N} - A^{-1}CA^{-1}/N + A^{-1}BA^{-1}BA^{-1}/N + o_p(1/N)$.

Proof of Lemma A.10. See the web page.

Proof of Lemma 2.2
We show the following three results, which then imply the main result:

(i) $E[R^{opt}] = E[\overline{w}_1] = \theta^*$,
(ii) $E[S^{opt}] = -E[\overline{v}_1\overline{w}_1] = -\rho/N$,
(iii) $E[T^{opt}] = E[\overline{w}_1\overline{v}_1\overline{v}_1] = \mu_{21}/N^2$.

(i) This is immediate.
(ii) $E[S^{opt}] = -E[\overline{v}_1\overline{w}_1] = -E\left[\frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N v_{i1}w_{j1}\right] = -E\left[\frac{1}{N^2}\sum_{i=1}^N v_{i1}w_{i1}\right] = -\rho/N$.
(iii) $E[T^{opt}] = E[\overline{w}_1\overline{v}_1\overline{v}_1] = E\left[\frac{1}{N^3}\sum_{i=1}^N\sum_{j=1}^N\sum_{k=1}^N w_{i1}v_{j1}v_{k1}\right] = E\left[\frac{1}{N^3}\sum_{i=1}^N w_{i1}v_{i1}v_{i1}\right] = \mu_{21}/N^2$.

Proof of Lemma 2.3. We first show the following results:

(i) $E[R^{opt}R^{opt}] = E[\overline{w}_1\overline{w}_1] = 1/N$,
(ii) $E[R^{opt}S^{opt}] = -E[\overline{w}_1\overline{v}_w\overline{w}_1] = -\mu_{12}/N^2$,
(iii) $E[S^{opt}S^{opt}] = E[\overline{v}_1\overline{w}_1\overline{v}_1\overline{w}_1] = (1 + 2\rho^2)/N^2 + o(1/N^2)$
(iv) $E[R^{opt}T^{opt}] = E[\overline{w}_1\overline{w}_1\overline{v}_1\overline{v}_1] = (2\rho^2 + 1)/N^2$.

In the following, let $\delta_{mn} = 1$ if $m = n$ and zero otherwise.

(i) $E[R^{opt}R^{opt}] = E[\overline{w}_1\overline{w}_1] = E[\sum_{i=1}^N\sum_{j=1}^N w_{i1}w_{j1}] = E[\sum_{i=1}^N w_{i1}^2] = 1/N$.
(ii) $E[R^{opt}S^{opt} \cdot R_1] = -E[\overline{w}_1\overline{w}_1\overline{v}_1] = -E[\frac{1}{N^3}\sum_{i=1}^N\sum_{j=1}^N\sum_{k=1}^N w_{i1}w_{j1}v_{k1}] = -E[\frac{1}{N^3}\sum_{i=1}^N w_{i1}^2 v_{i1}] = -\mu_{12}/N^2$.
(iii) $E[S^{opt}S^{opt}] = E[(\overline{v}_1\overline{w}_1)^2] = E[(\frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N v_{i1}w_{j1})^2] = E[\frac{1}{N^4}\sum_{i=1}^N\sum_{j=1}^N\sum_{k=1}^N\sum_{l=1}^N v_{i1}w_{j1}v_{k1}w_{l1}]$.

Because the (v_{im}, w_{im}) is independent of (v_{jn}, w_{jn}) if either $i \neq j$ or $m \neq n$, we can ignore all terms where one of the four indices i, j, k, and l, does not match up with at least one of the others. Ignoring also the N terms with all four indices matching up because they are of lower order, we only consider terms with $(i = j, k = l, i \neq k)$, $(i = l, j = k, i \neq j)$, or $(i = k, j = l, i \neq j)$, leading to

$$E[S^{opt}S^{opt}] = \frac{1}{N^4}E\left[\sum_{i=1}^N\sum_{k\neq i} v_{i1}w_{i1}v_{k1}w_{k1} + \sum_{i=1}^N\sum_{j\neq i} v_{i1}w_{j1}v_{j1}w_{i1} + \sum_{i=1}^N\sum_{j\neq i} v_{i1}w_{j1}v_{i1}w_{j1}\right] + o(1/N^2)$$
$$= \frac{1}{N^2}(\rho^2 + \rho^2 + 1) + o(1/N^2) = (2\rho^2 + 1)/N^2 + o(1/N^2).$$

(iv)

$$E[R^{opt}T^{opt}] = E[\overline{w}_1\overline{w}_1\overline{v}_1\overline{v}_1]$$
$$= E\left[\frac{1}{N^4}\sum_{i=1}^N\sum_{j=1}^N\sum_{k=1}^N\sum_{l=1}^N w_{i1}w_{j1}v_{i1}v_{k1}\right]$$
$$= (2\rho^2 + 1)/N^2 + o(1/N^2),$$

by the same argument as in Lemma 2.2(iii).

Then, adding up the three components

$$E[(R^{opt} + S^{opt} + T^{opt} - \theta^*)^2)]$$
$$= E[R^{opt}R^{opt} + 2R^{opt}S^{opt} + S^{opt}S^{opt} + 2R^{opt}T^{opt}] + o(N^{-2})$$
$$= 1/N - 2\mu_{12}/N^2 + 3(1 + 2\rho^2)/N^2 + o(N^{-2}).$$

Proof of Lemma 3.4. First we expand $\overline{ww'}^{-1}$ using Lemma A.10:

$$\overline{ww'}^{-1} = \mathcal{I}_M - (\overline{ww'} - \mathcal{I}_M) + (\overline{ww'} - \mathcal{I}_M)(\overline{ww'} - \mathcal{I}_M) + o_p(N^{-1}).$$

Second, we expand $(\bar v + e_1)'\overline{ww'}^{-1}(\bar v + e_1)$:

$$\begin{aligned}
&(\bar v+e_1)'\overline{ww'}^{-1}(\bar v+e_1)\\
&\quad = (\bar v+e_1)'\left(\mathcal{I}_M - (\overline{ww'}-\mathcal{I}_M) + (\overline{ww'}-\mathcal{I}_M)(\overline{ww'}-\mathcal{I}_M)\right)\\
&\qquad \times (\bar v+e_1) + o_p(N^{-1})\\
&\quad = 1 + 2\bar v_1 - (\overline{ww'}_{11}-1) + \bar v'\bar v - 2e_1'(\overline{ww'}-\mathcal{I}_M)\bar v\\
&\qquad + e_1(\overline{ww'}-\mathcal{I}_M)(\overline{ww'}-\mathcal{I}_M)e_1 + o_p(N^{-1}).
\end{aligned}$$

Next, we invert this expression, again using Lemma A.10:

$$\begin{aligned}
&\left((\bar v+e_1)'\overline{ww'}^{-1}(\bar v+e_1)\right)^{-1}\\
&\quad = 1 - 2\bar v_1 + (\overline{ww'}_{11}-1) + 2e_1(\overline{ww'}-\mathcal{I}_M)\bar v\\
&\qquad - e_1(\overline{ww'}-\mathcal{I}_M)(\overline{ww'}-\mathcal{I}_M)e_1 - \bar v'\bar v + 4\bar v_1^2\\
&\qquad + (\overline{ww'}_{11}-1)^2 - 4\bar v_1(\overline{ww'}_{11}-1) + o_p(N^{-1}).
\end{aligned}$$

Fourth, we expand $(\bar v + e_1)'\overline{ww'}^{-1}\bar w$:

$$\begin{aligned}
&(\bar v+e_1)'\overline{ww'}^{-1}\bar w\\
&\quad = (\bar v+e_1)'\left(\mathcal{I}_M - (\overline{ww'}-\mathcal{I}_M) + (\overline{ww'}-\mathcal{I}_M)(\overline{ww'}-\mathcal{I}_M)\right)\bar w\\
&\qquad + o_p(N^{-3/2})\\
&\quad = \bar w_1 + \bar v'\bar w - e_1'(\overline{ww'}-\mathcal{I}_M)\bar w - \bar v'(\overline{ww'}-\mathcal{I}_M)\bar w\\
&\qquad + e_1'(\overline{ww'}-\mathcal{I}_M)(\overline{ww'}-\mathcal{I}_M)\bar w + o_p(N^{-3/2}).
\end{aligned}$$

Finally, we consider the product:

$$\begin{aligned}
&\left((\bar v+e_1)'\overline{ww'}^{-1}(\bar v+e_1)\right)^{-1}(\bar v+e_1)'\overline{ww'}^{-1}\bar w\\
&\quad = \bar w_1 + \bar v'\bar w - e_1'(\overline{ww'}-\mathcal{I}_M)\bar w - 2\bar v_1\bar w_1 + (\overline{ww'}_{11}-1)\bar w_1\\
&\qquad - \bar v(\overline{ww'}-\mathcal{I}_M)\bar w + e_1(\overline{ww'}-\mathcal{I}_M)(\overline{ww'}-\mathcal{I}_M)\bar w\\
&\qquad - 2\bar v_1\bar v'\bar w + 2\bar v_1 e_1'(\overline{ww'}-\mathcal{I}_M)\bar w\\
&\qquad + 2\bar w_1 e_1(\overline{ww'}-\mathcal{I}_M)\bar v - \bar w_1 e_1'(\overline{ww'}-\mathcal{I}_M)(\overline{ww'}-\mathcal{I}_M)e_1\\
&\qquad + 4\bar v_1^2\bar w_1 + \bar w_1(\overline{ww'}_{11}-1)(\overline{ww'}_{11}-1)\\
&\qquad - 4\bar v_1\bar w_1(\overline{ww'}_{11}-1) + (\overline{ww'}_{11}-1)\bar v'\bar w\\
&\qquad - (\overline{ww'}_{11}-1)e_1'(\overline{ww'}-\mathcal{I}_M)\bar w - \bar w_1\bar v'\bar v + o_p(N^{-3/2}).
\end{aligned}$$

Proof of Lemma 3.5. Define:

(i) $S_1^{\text{gmm}} = -2\bar v_1\bar w_1$,
(ii) $S_2^{\text{gmm}} = (\overline{ww'}_{11}-1)\cdot\bar w_1$,
(iii) $S_3^{\text{gmm}} = -e_1'\left(\overline{ww'}-\mathcal{I}_M\right)\bar w$,
(iv) $S_4^{\text{gmm}} = \bar v'\bar w$.

We show the following results:

(i) $E[R^{\text{gmm}}] = E[\bar w_1] = \theta^*$,
(ii) $E[S_1^{\text{gmm}}] = E[-2\bar v_1\bar w_1] = -2\rho/N$,
(iii) $E[S_2^{\text{gmm}}] = E[(\overline{ww'}_{11}-1)\cdot\bar w_1] = \mu_{03}/N$,
(iv) $E[S_3^{\text{gmm}} = -e_1'\left(\overline{ww'}-\mathcal{I}_M\right)\bar w] = -\mu_{03}/N$,
(v) $E[S_4^{\text{gmm}}] = E[\bar v'\bar w] = M\rho/N$,

which then by adding up imply the result in Lemma 3.5. The details of the calculation are shown on the web page.

Proof of Lemma 3.6.

The proof proceeds by computing the expectations of eighteen separate terms, using methods similar to those of Lemma 3.5.The details are on the web page.

Before proving Lemma 4.7, it is useful to consider the solution for t given θ. Define $\hat t(\theta)$ implicitly through the first equation:

$$0 = \sum_{i=1}^N \psi(v_i, w_i, \theta)\cdot g(t(\theta)'\psi(v_i, w_i, \theta)).$$

Lemma A.11. (EXPANSION FOR $\hat t(\theta)$)
If $\theta = \hat\theta_{\text{opt}} + o_p(1/\sqrt{N})$, then

$$\begin{aligned}
\hat t(\theta) = {} & -e_1\theta + \bar w - \bar v\bar w_1 - \overline{ww'(\bar w - e_1\bar w_1)w'(\bar w - e_1\bar w_1)}\lambda/2\\
& + (\overline{ww'}-\mathcal{I}_M)e_1\bar w_1\\
& - (\overline{ww'}-\mathcal{I}_M)\bar w - 2\rho\bar w_1^2 e_1 + 2\rho\bar w_1\bar w + o_p(1/N).
\end{aligned}$$

Proof of Lemma A.11: Use a Taylor series expansion around zero for $g(a)$, $g(a) = g(0) + g'(0)a + g''(\tilde a)a^2/2 = 1 + a + g''(\tilde a)a^2/2$, to write the equation characterizing $\hat t(\theta)$ as

$$0 = \sum_{i=1}^N \psi(v_i, w_i, \theta)\cdot\left(1 + t(\theta)'\psi(v_i, w_i, \theta) + g''(a)\left(t(\theta)'\psi(v_i, w_i, \theta)\right)/2\right),$$

for some a between zero and $t(\theta)'\psi(v_i, w_i, \theta)$. Hence,

$$\begin{aligned}
\hat t(\theta) = & -\left(\frac{1}{N}\sum_{i=1}^N \psi(v_i, w_i, \theta)\psi(v_i, w_i, \theta)'\right)^{-1}\\
& \times\left(\sum_{i=1}^N \psi(v_i, w_i, \theta) + \psi(v_i, w_i, \theta)g''(a)\left(t(\theta)'\psi(v_i, w_i, \theta)\right)^2/2\right).
\end{aligned}$$

The second step is to show that

$$\frac{1}{N}\sum_{i=1}^{N}\psi(v_i,w_i,\theta)g''(a)\left(t(\theta)'\psi(v_i,w_i,\theta)\right)^2/2$$
$$=\overline{ww'(\overline{w}-e_1\overline{w}_1)w'}(\overline{w}-e_1\overline{w}_1)\lambda/2+o_p(1/N). \qquad \text{(A.9)}$$

To see this, first note that because $\theta=\overline{w}_1+o_p(1/\sqrt{N})$, we have $\hat{t}(\theta)=\overline{w}-e_1\overline{w}_1+o_p(1/\sqrt{N})$. Hence,

$$\frac{1}{N}\sum_{i=1}^{N}\psi(v_i,w_i,\theta)\left(t(\theta)'\psi(v_i,w_i,\theta)\right)^2$$
$$=\frac{1}{N}\sum_{i=1}^{N}\psi(v_i,w_i,\theta^*)\left(t(\theta)'\psi(v_i,w_i,\theta^*)\right)^2+o_p(1/N)$$
$$=\frac{1}{N}\sum_{i=1}^{N}w_i\left(t(\theta)'w_i\right)^2+o_p(1/N)$$
$$=\frac{1}{N}\sum_{i=1}^{N}w_i\left((\overline{w}-e_1\overline{w}_1)'w_i\right)^2+o_p(1/N)$$
$$=\overline{ww'(\overline{w}-e_1\overline{w}_1)w'}(\overline{w}-e_1\overline{w}_1)+o_p(1/N).$$

Since $t=o_p(1)$, $a=o_p(1)$, and $g''(a)=\lambda+o_p(1)$, so that the result in Equation (A.9) follows.

The third step is to show that, with $\theta=\overline{w}_1+o_p(1/\sqrt{N})$, we have

$$\left[\frac{1}{N}\sum_{i=1}^{N}\psi(v_i,w_i,\theta)\psi(v_i,w_i,\theta)'\right]^{-1}$$
$$=\mathcal{I}_M+(\overline{ww'}-\mathcal{I}_M)+2\rho\overline{w}_1\mathcal{I}_M+o_p(1/\sqrt{N}). \qquad \text{(A.10)}$$

To see this, first write

$$\frac{1}{N}\sum_{i=1}^{N}\psi(v_i,w_i,\theta)\psi(v_i,w_i,\theta)'$$
$$=\overline{ww'}-2\overline{vw'}\theta-2e_1\overline{w}'\theta+\theta^2\overline{(v+e_1)(v+e_1)'}$$
$$=\mathcal{I}_M+(\overline{ww'}-\mathcal{I}_M)-2\rho\mathcal{I}_M\theta+o_p(1/\sqrt{N})$$
$$=\mathcal{I}_M+(\overline{ww'}-\mathcal{I}_M)-2\rho\mathcal{I}_M\overline{w}_1+o_p(1/\sqrt{N}).$$

Hence, using Lemma A.10,

$$\left[\frac{1}{N}\sum_{i=1}^{N}\psi(v_i,w_i,\theta)\psi(v_i,w_i,\theta)'\right]^{-1}$$
$$=\mathcal{I}_M-(\overline{ww'}-\mathcal{I}_M)+2\rho\mathcal{I}_M\overline{w}_1+o_p(1/\sqrt{N}),$$

which proves the equality in Equation (A.10).

Then, using the fact that $\overline{\psi(v,w,\theta)}=(\overline{v}+e_1)\theta-\overline{w}$, we can approximate the expression for $\hat{t}(\theta)$ as

$$\hat{t}(\theta)=-\left(\mathcal{I}_M-(\overline{ww'}-\mathcal{I}_M)+2\rho\mathcal{I}_M\overline{w}_1\right)$$
$$\times\left((\overline{v}+e_1)\theta-\overline{w}+\overline{ww'(\overline{w}-e_1\overline{w}_1)w'}(\overline{w}-e_1\overline{w}_1)\lambda/2\right)$$
$$=-(\overline{v}+e_1)\theta+\overline{w}-\overline{ww'(\overline{w}-e_1\overline{w}_1)w'}(\overline{w}-e_1\overline{w}_1)\lambda/2$$
$$+(\overline{ww'}-\mathcal{I}_M)e_1\overline{w}_1-(\overline{ww'}-\mathcal{I}_M)\overline{w}-2\rho\overline{w}_1^2e_1$$
$$+2\rho\overline{w}_1\overline{w}+o_p(1/N).$$
$$=-e_1\theta+\overline{w}-\overline{w}_1\overline{v}-\overline{ww'(\overline{w}-e_1\overline{w}_1)w'}(\overline{w}-e_1\overline{w}_1)\lambda/2$$
$$+(\overline{ww'}-\mathcal{I}_M)e_1\overline{w}_1-(\overline{ww'}-\mathcal{I}_M)\overline{w}-2\rho\overline{w}_1^2e_1$$
$$+2\rho\overline{w}_1\overline{w}+o_p(1/N).$$

Proof of Lemma 4.7. The solution for $\hat{\theta}_g$ is characterized by the equation

$$0=\hat{t}(\theta)'\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\psi}{\partial\theta'}(v_i,w_i,\theta)\cdot g(\hat{t}(\theta)'\psi(v_i,w_i,\theta)).$$

We can write this as

$$0=\Big(\big[-e_1\theta+\overline{w}-\overline{w}_1\overline{v}-\overline{ww'(\overline{w}-e_1\overline{w}_1)w'}(\overline{w}-e_1\overline{w}_1)\lambda/2$$
$$+(\overline{ww'}-\mathcal{I}_M)e_1\overline{w}_1-(\overline{ww'}-\mathcal{I}_M)\overline{w}-2\rho\overline{w}_1^2e_1+2\rho\overline{w}_1\overline{w}\big]$$
$$+\hat{t}(\theta)-\big[e_1\theta+\overline{w}-\overline{w}_1\overline{v}-\overline{ww'(\overline{w}-e_1\overline{w}_1)w'}(\overline{w}-e_1\overline{w}_1)\lambda/2$$
$$+(\overline{ww'}-\mathcal{I}_M)e_1\overline{w}_1-(\overline{ww'}-\mathcal{I}_M)\overline{w}-2\rho\overline{w}_1^2e_1+2\rho\overline{w}_1\overline{w}\big]\Big)'$$
$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\psi}{\partial\theta'}(v_i,w_i,\theta)\cdot g(\hat{t}(\theta)'\psi(v_i,w_i,\theta)).$$

Hence,

$$\hat{\theta}_g=\left(e_1'\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\psi}{\partial\theta'}(v_i,w_i,\theta)\cdot g(\hat{t}(\theta)'\psi(v_i,w_i,\theta))\right)^{-1}$$
$$\times\Big(\big[\overline{w}-\overline{w}_1\overline{v}-\overline{ww'(\overline{w}-e_1\overline{w}_1)w'}(\overline{w}-e_1\overline{w}_1)\lambda/2$$
$$+(\overline{ww'}-\mathcal{I}_M)e_1\overline{w}_1-(\overline{ww'}-\mathcal{I}_M)\overline{w}-2\rho\overline{w}_1^2e_1+2\rho\overline{w}_1\overline{w}\big]$$
$$+\hat{t}(\theta)-\big[e_1\theta+\overline{w}-\overline{w}_1\overline{v}-\overline{ww'(\overline{w}-e_1\overline{w}_1)w'}(\overline{w}-e_1\overline{w}_1)\lambda/2$$
$$+(\overline{ww'}-\mathcal{I}_M)e_1\overline{w}_1-(\overline{ww'}-\mathcal{I}_M)\overline{w}-2\rho\overline{w}_1^2e_1+2\rho\overline{w}_1\overline{w}\big]\Big)'$$
$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\psi}{\partial\theta'}(v_i,w_i,\theta)\cdot g(\hat{t}(\theta)'\psi(v_i,w_i,\theta)).$$

We break this up in a couple of parts. First we show that

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\psi}{\partial\theta'}(v_i, w_i, \theta)\cdot g(\hat{t}(\theta)'\psi(v_i, w_i, \theta))$$
$$= e_1 + \bar{v} - \rho\bar{w} + \rho e_1\bar{w}_1 + o_p(1/\sqrt{N}). \qquad \text{(A.11)}$$

To see this, write out $\psi(v_i, w_i, \theta) = (v_i + e_1)\theta - w_i$ to get

$$\frac{1}{N}\sum_{i=1}^{N}(v_i + e_1)\cdot g(\hat{t}(\theta)'(v_i\theta + e_1\theta - w_i))$$
$$= \frac{1}{N}\sum_{i=1}^{N}(v_i + e_1)\cdot(1 + t(\theta)'(v_i\theta + e_1\theta - w_i)) + o_p(1/\sqrt{N})$$
$$= e_1 + \bar{v} - \frac{1}{N}\sum_{i=1}^{N} v_i\bar{w}'w_i + o_p(1/\sqrt{N})$$
$$= e_1 + \bar{v} - \rho\bar{w} + \rho e_1\bar{w}_1 + o_p(1/\sqrt{N}),$$

which proves the equality in Equation (A.11). A direct implication is that

$$\left(e_1'\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\psi}{\partial\theta'}(v_i, w_i, \theta)\cdot g(\hat{t}(\theta)'\psi(v_i, w_i, \theta))\right)^{-1}$$
$$= 1 - \bar{v}_1 + o_p(1/\sqrt{N}). \qquad \text{(A.12)}$$

Second, we show that

$$\left(\hat{t}(\theta) - \left[e_1\theta + \bar{w} - \bar{w}_1\bar{v} - \overline{ww'(\bar{w} - e_1\bar{w}_1)w'}(\bar{w} - e_1\bar{w}_1)\lambda/2\right.\right.$$
$$\left.\left. + (\overline{ww'} - \mathcal{I}_M)e_1\bar{w}_1 - (\overline{ww'} - \mathcal{I}_M)\bar{w} - 2\rho\bar{w}_1^2 e_1 + 2\rho\bar{w}_1\bar{w}\right]\right)'$$
$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\psi}{\partial\theta'}(v_i, w_i, \theta)\cdot g(\hat{t}(\theta)'\psi(v_i, w_i, \theta)) = o_p(1/N).$$

This follows from Lemma 4.7, which implies that the first factor is $o_p(1/N)$, combined with the fact that the left-hand side of Equation (A.11) is $O_p(1)$. Third, we show that

$$\left(\bar{w} - \bar{w}_1\bar{v} - \overline{ww'(\bar{w} - e_1\bar{w}_1)w'}(\bar{w} - e_1\bar{w}_1)\lambda/2 + (\overline{ww'}\right.$$
$$\left. - \mathcal{I}_M)e_1\bar{w}_1 - (\overline{ww'} - \mathcal{I}_M)\bar{w} - 2\rho\bar{w}_1^2 e_1 + 2\rho\bar{w}_1\bar{w}\right)'$$
$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\psi}{\partial\theta'}(v_i, w_i, \theta)\cdot g(\hat{t}(\theta)'\psi(v_i, w_i, \theta))$$
$$= \bar{w}_1 - e_1'\bar{w}_1\bar{v} - e_1'\overline{ww'(\bar{w} - e_1\bar{w}_1)w'}(\bar{w} - e_1\bar{w}_1)\lambda/2$$
$$+ e_1'(\overline{ww'} - \mathcal{I}_M)e_1\bar{w}_1 - e_1'(\overline{ww'} - \mathcal{I}_M)\bar{w} - 2\rho\bar{w}_1^2$$
$$+ 2\rho\bar{w}_1^2 + \bar{w}'\bar{v} - \rho\bar{w}'\bar{w} + \rho\bar{w}_1^2 + o_p(1/N).$$

$$= \bar{w}_1 - \bar{w}_1\bar{v}_1 - e_1'\overline{ww'(\bar{w} - e_1\bar{w}_1)w'}(\bar{w} - e_1\bar{w}_1)\lambda/2$$
$$+ e_1'(\overline{ww'} - \mathcal{I}_M)e_1\bar{w}_1 - e_1'(\overline{ww'} - \mathcal{I}_M)\bar{w}$$
$$+ \bar{w}'\bar{v} - \rho\bar{w}'\bar{w} + \rho\bar{w}_1^2 + o_p(1/N).$$

Now note that although $\overline{ww'(\bar{w} - e_1\bar{w}_1)w'}(\bar{w} - e_1\bar{w}_1) = O_p(1/N)$, $e_1'\overline{ww'(\bar{w} - e_1\bar{w}_1)w'}(\bar{w} - e_1\bar{w}_1) = o_p(1/N)$, because the subtraction of $e_1\bar{w}_1$ from $\bar{w}$ makes $(\bar{w} - e_1\bar{w}_1)$ independent of $e_1 w$. This relies on the full independence assumption we are using in the sequence of the moments. Because of this the term $e_1'\overline{ww'(\bar{w} - e_1\bar{w}_1)w'}(\bar{w} - e_1\bar{w}_1)$ is of lower order, and the above expression reduces to

$$\bar{w}_1 - \bar{w}_1\bar{v}_1 + e_1'(\overline{ww'} - \mathcal{I}_M)e_1\bar{w}_1 - e_1'(\overline{ww'} - \mathcal{I}_M)\bar{w}$$
$$+ \bar{w}'\bar{v} - \rho\bar{w}'\bar{w} + \rho\bar{w}_1^2 + o_p(1/N).$$

Finally bringing all the terms together, we get

$$\hat{\theta}_{\xi} = \bar{w}_1 + e_1'(\overline{ww'} - \mathcal{I}_M)e_1\bar{w}_1 - e_1'(\overline{ww'} - \mathcal{I}_M)\bar{w}$$
$$+ \bar{w}'\bar{v} - \rho\bar{w}'\bar{w} - 2\bar{w}_1\bar{v}_1 + \rho\bar{w}_1^2 + o_p(1/N).$$

Proof of Lemma 4.8. The proof proceeds by calculating:

(i) $E[T_1] = \theta^*$,
(ii) $E[R_3] = -\mu_{03}/N$,
(iii) $E[R_5] = \rho/N$,
(iv) $E[R_4] = \rho M/N$,
(v) $E[R_6] = -\rho M/N$,
(vi) $E[R_1] = -\rho/N$,
(vii) $E[R_2] = \mu_{03}/N$,
(viii) $E[R_7] = o_p(1/N)$.

The result then follows from adding up the expectations; details are on the web page.

Proof of Lemma 4.9. The expectations of twenty-eight component terms are defined and calculated; the result then follows from summing these components. The explicit calculation is given on the web page.

References

Baggerly, K. (1998), "Empirical Likelihood As a Goodness-of-Fit Measure," *Biometrika*, 85, 535–47.

Corcoran, S. A. (1995), "Empirical Discrepancy Theory," manuscript, Department of Statistics, University of Oxford.

Corcoran, S. A., A. C. Davison, and R. Spady (1995), "Reliable Inference From Empirical Likelihoods," Technical Report, Department of Statistics, Oxford University.

Corcoran, S. A. (1998), "Bartlett Adjustment of Empirical Discrepancy Statistics," *Biometrika*, 85, 967–72.

Cressie, N., and T. Read (1984), "Multinomial Goodness-of-Fit Tests," *Journal of the Royal Statistical Society, Series B*, 46, 440–64.

Donald, S., and W. K. Newey (2004), "Choosing the Number of Instruments," *Econometrica*, 69, 1161–91.

Hansen, L.-P. (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–54.

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. (1997), "One Step Estimators for Over-identified Generalized Method of Moments Models," *Review of Economic Studies*, 64, 359–83.

Imbens, G. W., R. H. Spady, and P. Johnson (1998), "Information Theoretic Approaches to Inference in Moment Condition Models," *Econometrica*, 66, 333–57.

Kitamura, Y., and M. Stutzer (1997), "An Information Theoretic Alternative to Generalized Method of Moments Estimation," *Econometrica*, 65, 861–74.

Owen, A. (1988), "Empirical Likelihood Ratio Confidence Intervals for a Single Functional," *Biometrika*, 75, 237–49.

Newey, W. K., and R. J. Smith, (2000), "Asymptotic Bias and Equivalence of GMM and GEL," mimeo, MIT and University of Bristol.

Newey, W. K., and R. J. Smith, (2001), "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators," mimeo, MIT and University of Bristol.

Qin, J., and J. Lawless (1994), "Empirical Likelihood and General Estimating Equations," *Annals of Statistics*, 22, 300–25.

Rothenberg, T. (1999), "Higher Order Properties of Empirical Likelihood for Simultaneous Equations," seminar notes, Harvard/MIT Econometrics Seminar, University of Bristol.

Smith, R. J. (1997), "Alternative Semi-Parametric Approaches to Generalized Method of Moments Estimation," *Economic Journal*, 107, 503–19.