Combining Empirical Likelihood and Generalized Method of Moments Estimators: Asymptotics and Higher Order Bias*

Roni Israelov[†]

Steven Lugauer[‡]

December 2010

Abstract

This paper proposes an estimator combining Empirical Likelihood (EL) and the Generalized Method of Moments (GMM) by allowing the sample average moment vector to deviate from zero and the sample weights to deviate from n^{-1}. The new estimator may be adjusted through free parameter $\delta \in (0, 1)$ with GMM behavior attained as $\delta \longrightarrow 0$ and EL as $\delta \longrightarrow 1$. When the sample size is small and the number of moment conditions is large, the parameter space under which the EL estimator is defined may be restricted at or near the population parameter value. The support of the parameter space for the new estimator may be adjusted through δ. The new estimator performs well in Monte Carlo simulations.

Keywords: Generalized Method of Moments, Empirical Likelihood

JEL Classification: C

*Portions of this paper were part of Roni Israelov's doctoral dissertation at Carnegie Mellon University.

†AQR Capital Management, LLC.

‡Corresponding author. University of Notre Dame, Department of Economics, 719 Flanner Hall, Notre Dame, IN 46556; Phone: 574-631-1432; Email: slugauer@nd.edu; Web Site: www.nd.edu/~slugauer.

1 Introduction

We propose an estimator combining the flexibility of the Generalized Method of Moments (GMM) estimator, which allows the moment vector to be non-zero, with that of Empirical Likelihood (EL), which allows the weighting vector to be non-fixed.[1] The new Penalized Method of Moments (PMM) estimator lets the sample average moment vector deviate from zero, but the deviation is costly through a GMM-type quadratic penalty function. The weighting vector can deviate from n^{-1}, but the deviation is costly through EL's Kullback-Leibler Information Criterion (KLIC) penalty function. The PMM objective function combines the GMM and EL objective functions. By including parameter $\delta \in (0,1)$ PMM becomes a continuum of estimators with behavior approaching GMM when δ approaches zero and with behavior approaching EL when δ approaches one.

The next section presents the PMM estimator, lists PMM's asymptotic properties, shows how PMM adopts features of GMM and EL, and details a specification test based on overidentifying restrictions. In practice, EL and PMM could be undefined at or near the population parameter values because their objective functions are undefined for non-positive sample weights. This possibility increases as the sample size decreases, the number of moments increases, or when the model is misspecified. However, Section 3 demonstrates how the PMM estimator allows the econometrician to increase the support of the parameter space by adjusting δ to shrink the optimal weights towards GMM's fixed weights. Section 3 also summarizes the results of a series of Monte Carlo simulation experiments on a Hall and Horowitz (1996) style model. As theory predicts, the restricted parameter space affects the estimates if the number of moment conditions is large relative to the sample size. In this case, PMM's estimates are more accurate on average, display far less variance over the simulations, and have less mis-sized hypothesis tests than EL. The complete Monte Carlo simulation results and all derivations and proofs of PMM's properties can be found in the online Appendix.

2 The Penalized Method of Moments Estimator

Let x_i be a sequence of iid random vectors taking values in $\mathcal{X} \subset \mathbb{R}^p$. Let θ denote the parameter vector of interest belonging in the space Θ, a compact subset of $\mathbb{R}^k$. The vector of equations $g(x_i, \theta)$ takes values in $\mathbb{R}^m$ with $m \geq k$ and identifies the population parameter vector θ_0

$$E\left[g(x_i, \theta_0)\right] = 0 \qquad E\left[g(x_i, \theta_0) g(x_i, \theta_0)'\right] = \Sigma.$$

[1] Several researchers have designed estimators to improve on the small-sample properties of GMM. The Exponential Tilting estimator of Kitamura and Stutzer (1997) and Imbens, Johnson, and Spady (1998), the Continuous Updating GMM estimator of Hansen, Heaton, and Yaron (1996), and the EL estimator of Owen (1988), Qin and Lawless (1994), and Imbens (1997) have received special attention. We focus on EL because it is the most widely used alternative to GMM. In related work, Ragusa (2008) and Corocoran (1998) examine the asymptotics for the class of minimum discrepancy estimators.

Let n denote the sample size and let all summations be over $1, \ldots, n$. Let ω denote a vector of weights on the $(n-1)$ dimensional unit simplex given by $\Psi = \{\omega = (\omega_1, \ldots, \omega_n) | \omega_i \geq 0, \sum_i \omega_i = 1\}$. Let $g_i(\theta)$ represent $g(x_i, \theta)$ and define $m_i(\theta) \equiv \partial g_i(\theta)/\partial\theta'$, $v_i(\theta) \equiv g_i(\theta)g_i(\theta)'$, and

$$
\begin{array}{ll}
G_n(\theta) = \frac{1}{n}\sum_i g_i(\theta) & \mathcal{G}_n(\theta) \equiv \mathcal{G}_n(\omega, \theta) = \sum_i \omega_i g_i(\theta) \\
M_n(\theta) = \frac{1}{n}\sum_i m_i(\theta) & \mathcal{M}_n(\theta) \equiv \mathcal{M}_n(\omega, \theta) = \sum_i \omega_i m_i(\theta) \\
V_n(\theta) = \frac{1}{n}\sum_i v_i(\theta) & \mathcal{V}_n(\theta) \equiv \mathcal{V}_n(\omega, \theta) = \sum_i \omega_i v_i(\theta) - \mathcal{G}_n(\theta)\mathcal{G}_n(\theta)'.
\end{array}
$$

In addition G, $\mathcal{G}$, M, $\mathcal{M}$, V, and $\mathcal{V}$ refer to the limiting value as $n \to \infty$ of the respective functions evaluated at the true parameter value θ_0. Let W_n refer to an $m \times m$ full rank symmetric positive definite weighting matrix with limiting value W as $n \to \infty$.

2.1 PMM Defined

In the limit, the equal-weight sample average moment vector converges to its expected value of zero. However, the sample average moments are a random vector, and in small samples adding flexibility by allowing some deviation from the expected values may provide desirable estimation properties. For example, GMM sets k sample moments to zero, and $m-k$ sample moments are not restricted and may deviate from zero; GMM restricts the weights on individual observations to n^{-1}. In contrast, EL allows the weights to deviate from n^{-1}, but restricts all m sample moments to zero. We propose an estimator that allows both the weights and the moment vector to deviate from n^{-1} and zero respectively. Deviation of the weighted-average moment vector is costly through a quadratic penalty function and deviation of the sample weights is costly through a KLIC penalty function. We define the Penalized Method of Moments (PMM) estimator as follows.

PMM Estimator

$$
\hat{\theta} = \underset{\substack{\theta\in\Theta \\ \omega\in\Psi}}{\arg\min} \quad \underset{\text{PMM}}{Q(\omega, \theta)} = \frac{1}{\delta(1-\delta)}\left[\delta n \mathcal{G}_n(\theta)' W_n \mathcal{G}_n(\theta) - 2(1-\delta)\sum_i \ln(n\omega_i)\right] \tag{1}
$$

The sample weights must sum to one ($\sum_i \omega_i = 1$), and $\delta \in (0,1)$ allows for the relative importance of the quadratic penalty versus the KLIC penalty to be adjusted. The division by $\delta(1-\delta)$ ensures proper scaling for $Q(\hat{\omega}, \hat{\theta})$ to serve as a test statistic of model specification based on the overidentifying restrictions.

We follow Qin and Lawless (1994) and use Lagrange multipliers to solve the optimization problem. Let

$$
\mathcal{L} = \frac{1}{\delta(1-\delta)}\left[\delta n \mathcal{G}_n(\theta)' W_n \mathcal{G}_n(\theta) - (1-\delta) 2\sum_i \ln(n\omega_i)\right] - \lambda\left(\sum_i \omega_i - 1\right), \tag{2}
$$

where λ is the Lagrange multiplier. Differentiating with respect to ω_i gives the first set of first-order conditions

$$\frac{\partial \mathcal{L}}{\partial \omega_i} = \frac{2}{\delta(1-\delta)}\left[\delta n g_i(\theta)' W_n \mathcal{G}_n(\theta) - \frac{1-\delta}{\omega_i}\right] - \lambda = 0. \tag{3}$$

Multiplying by ω_i and summing over i provides

$$\sum_i \omega_i \frac{\partial \mathcal{L}}{\partial \omega_i} = \frac{2}{\delta(1-\delta)}\left[\delta n \mathcal{G}_n(\theta)' W_n \mathcal{G}_n(\theta) - n(1-\delta)\right] - \lambda = 0. \tag{4}$$

Solving for λ

$$\lambda = \frac{2}{\delta(1-\delta)}\left[n\delta \mathcal{G}_n(\theta) W_n \mathcal{G}_n(\theta) - n(1-\delta)\right]. \tag{5}$$

Substituting λ from equation (5) into (3) and rearranging terms results in the following system of n equations to identify the n probability weights

$$\omega_i = \frac{1}{n}\left(\frac{1}{1 + \frac{\delta}{1-\delta}\left(g_i(\theta) - \mathcal{G}_n(\theta)\right)' W_n \mathcal{G}_n(\theta)}\right). \tag{6}$$

Equation (6) does not emit a closed form solution for the weights because $\mathcal{G}_n(\theta)$ is a function of the weight vector. However, the n equations imply a solution to the n unknowns conditional on θ, and the implied weights, $\omega(\theta)$, are a function of the parameter vector. In the Appendix, we apply the implicit function theorem to verify an implicit function $\omega(\theta)$ exists asymptotically. For PMM, all three components, $\mathcal{G}_n(\theta)$, $\mathcal{M}_n(\theta)$, and $\mathcal{V}_n(\theta)$, use the same vector of weights.

Because the optimal PMM weights are an implicit function of the parameter vector, the optimally weighted objective function may be written as a function of only the parameter vector

$$\underset{\text{PMM}}{Q}(\hat{\omega}, \theta) = \underset{\text{PMM}}{Q}(\omega(\theta), \theta) = \underset{\text{PMM}}{Q}(\theta).$$

So far, only the optimal weights have been defined. Differentiating (2) with respect to θ gives the second set of first-order conditions that identify $\hat{\theta}$

$$\mathcal{M}_n(\hat{\theta})' W_n \mathcal{G}_n(\hat{\theta}) = 0. \tag{7}$$

Equation (7) is a system of k equations in k unknowns. The first-order conditions are a function of the optimal weights, but the optimal weights are implicitly a function of the parameter vector. Hence, equations (6) and (7) can be solved simultaneously for the n optimal weights and the k optimal parameters. The first-order conditions for PMM have the familiar form in which a linear combination of an estimate of the orthogonality condition is set to zero. Next, we list PMM's asymptotic properties.

2.2 PMM Asymptotic Properties

PMM inherits most of the asymptotic properties of GMM and EL. We postpone the discussion of higher-order bias until Section 2.4.

Let $f(n) = \mathcal{O}(g(n))$ denote an asymptotic upper bound

$$f(n) = \mathcal{O}(g(n)) \quad \Leftrightarrow \quad \limsup_{n\to\infty} \left| \frac{f(n)}{g(n)} \right| < \infty,$$

and $f(n) = o(g(n))$ denote asymptotic negligibility

$$f(n) = o(g(n)) \quad \Leftrightarrow \quad \lim_{n\to\infty} \frac{f(n)}{g(n)} = 0.$$

Theorem 1 ***(Limiting Behavior)***

Suppose $\delta \in (0,1)$, W_n is a full rank positive definite symmetric $m \times m$ matrix, and $\omega(\theta)$ is the vector of weights implied by the first order conditions of equation (2) and defined by (6). In addition, suppose that $W_n = \Sigma^{-1} + \mathcal{O}\left(n^{-\frac{1}{2}}\right)$ is symmetric positive semi-definite. Then,

$$\textit{(i)} \quad \omega_i(\theta_0) - n^{-1} = \mathcal{O}\left(n^{-\frac{3}{2}}\right) \qquad \forall i, \tag{8}$$

$$\textit{(ii)} \quad \mathcal{G}_n(\theta_0) = (1-\delta)G_n(\theta_0) + \mathcal{O}\left(n^{-1}\right), \tag{9}$$

$$\textit{(iii)} \quad \mathcal{V}_n(\theta_0) = \Sigma + \mathcal{O}\left(n^{-\frac{1}{2}}\right), \tag{10}$$

$$\textit{(iv)} \quad \mathcal{M}_n(\theta_0) = M_n(\theta_0) + \mathcal{O}\left(n^{-\frac{1}{2}}\right). \tag{11}$$

Theorem 1 describes the limiting behavior of the probability weights, moment vector, second moment matrix, and Jacobian term. The weights converge to fixed weights, and the weighted-average Jacobian and second-moment matrices converge to the limiting value of their respective equal-weight counterparts. The weighted average moment vector, on the other hand, converges to the sample average scaled by $(1-\delta)$. As δ approaches one and PMM's first-order conditions approach those of EL, $\mathcal{G}_n(\theta_0) \to 0$, which is the constraint imposed by EL estimation.

Let $||\cdot||$ denote the Euclidean norm and define $\mathcal{S}_n(\theta)$ as

$$\mathcal{S}_n(\theta) \equiv \left(\delta \mathcal{V}_n(\theta) + (1-\delta)W_n^{-1}\right)^{-1}. \tag{12}$$

The matrix $\mathcal{S}_n(\theta)$ is a continuously updated weighting matrix. The inverse of $\mathcal{S}_n(\theta)$ is a convex combination of the efficiently estimated second moment matrix and the inverted weighting matrix provided by the econometrician.[2] PMM exhibits consistency according to the following theorem.

[2]Efficient GMM sets the inverted weighting matrix equal to a consistent estimate of Σ. If the PMM weighting matrix is also a consistent estimate, then $\mathcal{S}_n^{-1}(\hat{\theta})$ is an estimate of Σ.

Theorem 2 ***(Consistency)***

If $E[g_i(\theta_0)g_i(\theta_0)']$ is positive definite, $m_i(\theta)$ is continuous in the neighborhood of the true value θ_0, $\|m_i(\theta)\|$ and $\|g_i(\theta)\|^3$ are bounded by some integrable function $H(x)$ in this neighborhood, and $E[m_i(\theta_0)]$ has rank k. Then as $n \to \infty$, with probability one $Q(\theta)_{PMM}$ attains its minimum value at some point $\hat{\theta}$ in the interior of the ball $\|\theta - \theta_0\| \le n^{-\varphi}$ where $0 < \varphi < \frac{1}{2}$ and $\hat{\theta}$ satisfies $\mathcal{M}_n(\hat{\theta})'\mathcal{S}_n(\hat{\theta})G_n(\hat{\theta}) = 0$.

Theorem 3 provides the first order asymptotic distribution of the parameter estimate for any supplied symmetric positive definite weighting matrix W_n.

Theorem 3 ***(Asymptotic Normality)***

Suppose $\hat{\theta}$ satisfies equation (7) where $W_n \xrightarrow{p} W$, W_n is full rank and symmetric positive semi-definite, and $\hat{\theta} \xrightarrow{p} \theta_0$. Let $\mathcal{S}$ denote the limiting value of $\mathcal{S}_n(\theta_0)$ as $n \to \infty$, and assume

1. *$\theta_0 \in$ interior(Θ)*
2. *$G_n(\theta)$ and $\mathcal{G}_n(\theta)$ are continuously differentiable in a neighborhood $\mathfrak{N}$ of θ_0*
3. *$\sqrt{n}G_n(\theta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma)$*
4. *there exists $M(\theta)$ that is continuous at θ_0 and $\sup_{\theta \in \mathfrak{N}} \|M_n(\theta) - M(\theta)\| \xrightarrow{p} 0$*
5. *there exists $\mathcal{M}(\theta)$ that is continuous at θ_0 and $\sup_{\theta \in \mathfrak{N}} \|\mathcal{M}_n(\theta) - \mathcal{M}(\theta)\| \xrightarrow{p} 0$*
6. *for $M = M(\theta_0)$, $M'\mathcal{S}M$ is nonsingular.*

Then $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, (M'\mathcal{S}M)^{-1}M'\mathcal{S}\Sigma\mathcal{S}M(M'\mathcal{S}M)^{-1}\right)$.

The asymptotic variance is similar in form to that of non-efficient GMM, except that $\mathcal{S}$, a function of the weighting matrix W, takes the place of W in GMM.[3]

Theorem 4 ***(Efficiency)***

If all the conditions of Theorem 3 hold and $W_n \xrightarrow{p} \Sigma^{-1}$, then $\hat{\theta}$ is efficient in that it achieves the lowest first order asymptomatic variance and $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, (M'\Sigma^{-1}M)^{-1}\right)$.

PMM shares GMM's main drawback; PMM is a two-step estimator. Thus, PMM has the same efficiency condition as GMM – W_n must be a consistent estimate of Σ^{-1}.

[3] See equation (15) in the next section.

2.3 The Ingredients: GMM and EL

The PMM family of estimators merges the objective functions of GMM and EL. GMM, EL, and PMM estimation equations build off of the moment conditions, which have expectation zero. The objective is to select a $\hat{\theta}$ that agrees with the information contained in the moment conditions. GMM uses fixed weights to estimate the expected value of the estimating equations. If there are more moments than parameters, then not all m moment conditions can be jointly satisfied because the sample weights are fixed to n^{-1}. GMM sets k dimensions of the moment vector to zero and measures the distance from zero of the remaining $m-k$ dimensions with a quadratic penalty function. Optimality is defined by the parameter value associated with the lowest penalty. The GMM estimate of $\hat{\theta}$ solves the following optimization problem.

GMM estimator

$$\hat{\theta} = \underset{\theta \in \Theta}{\arg\min} \quad \underset{\text{GMM}}{Q(\theta)} = nG_n(\theta)'W_nG_n(\theta) \tag{13}$$

W_n is a symmetric positive definite $m \times m$ *weighting* matrix. The first-order conditions for GMM are

$$M_n(\hat{\theta})'W_nG_n(\hat{\theta}) = 0, \tag{14}$$

which is a $k \times 1$ vector identifying the k parameters. GMM is estimated efficiently by setting $W_n \xrightarrow{p} \Sigma^{-1}$, which is typically accomplished by using equation (13) twice. The first step uses a weighting matrix $\tilde{W}$, the identity matrix for example, to generate the first round consistent estimate $\tilde{\theta}$. The second step sets $W_n = V_n^{-1}(\tilde{\theta})$, which is a consistent estimate of Σ^{-1}. Under standard regularity conditions, the estimator $\hat{\theta}_{GMM}$ is consistent and asymptotically normally distributed

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, \left(M'WM\right)^{-1} M'W\Sigma WM \left(M'WM\right)^{-1}\right). \tag{15}$$

EL takes a different approach. EL calculates the sample average moment vector allowing weights to deviate from GMM's fixed weights and constrains the sample average moment vector to zero. Because sample weights are not fixed, it is possible to select a vector of non-negative probability weights under which the expected value of the estimating equations is equal to zero so long as zero is contained in the convex hull of the points $g(x_1, \theta), \ldots, g(x_n, \theta)$. EL is defined over the set of weights and parameter values where this condition is satisfied (i.e. $\mathcal{G}_n(\theta) = 0$ and $\omega > 0$).[4] When the system is overidentified, EL's moment vector constraint may not be satisfied under fixed weights. EL measures the distance from n^{-1} with a penalty function defined by the KLIC and with optimality defined by the parameter and weighting vectors associated with the lowest penalty.

[4] EL may select a θ such that the estimated parameter values are distant from the population parameter values because the condition $\omega > 0$ must also be satisified. This point will be discussed further below.

EL Estimator

$$\hat{\theta} = \underset{\substack{\theta \in \Theta \\ \omega \in \Psi}}{\arg\min} \quad \underset{\text{EL}}{Q}(\omega) = -2\sum_i \ln n\omega_i \quad \text{subject to} \quad \mathcal{G}_n(\theta) = 0 \tag{16}$$

EL minimizes the empirical discrepancy as measured by the KLIC. In other words, EL maximizes the likelihood of the multinomial distribution for the data subject to binding orthogonality conditions.[5] Unlike GMM, EL requires no iteration to achieve efficiency. The probability weights satisfying the first-order conditions solve

$$\omega_i = \frac{1}{n}\left(\frac{1}{1 + g_i(\theta)'\mathcal{V}_n^{-1}(\theta)G_n(\theta)}\right). \tag{17}$$

The weights cannot be found explicitly because $\mathcal{V}_n(\theta)$ is a function of the probability weights. Denote the implicit solution to (17) by $\omega_i(\theta)$. These optimal weights are conditional on θ and *may* be negative. In order for ω to be in Ψ, the estimation procedure restricts the parameter space to the region in which the optimal weights are positive. If this region is the null space, then the EL estimator is undefined for the combined model and data set.

Newey and Smith (2004) show that the EL estimator $\hat{\theta}$ solves the following first-order conditions.

$$\mathcal{M}_n(\hat{\theta})\mathcal{V}_n^{-1}(\theta)G_n(\hat{\theta}) = 0 \tag{18}$$

The $k \times 1$ vector in Equation (18) represents the identifying space for the k parameters and depends on the probability weights. Hence equations (17) and (18) must be solved jointly for the n weights and the k parameters. Note the similar structure between the first-order conditions provided by equations (7), (14), and (18). In each set of first-order conditions, a linear combination of the sample average moment vector is set to zero. The linear combination is comprised of estimates of the Jacobian and the second moment matrix. For EL, calculation of the expected Jacobian uses the probability weights minimizing the objective function with the weighting matrix calculated simultaneously. GMM, on the other hand, uses fixed weights to calculate the average score with the weighting matrix supplied *a priori*. PMM combines the two approaches.

Theorem 5 relates the optimally weighted to the equally weighted sample average moment vector for any θ in PMM estimation. The two estimates are explicitly related and δ has an important role in the relationship.

Theorem 5

Suppose $\delta \in (0,1)$, W_n is a full rank positive definite symmetric $m \times m$ matrix, and $\omega(\theta)$ is the vector of weights implied by the first order conditions of equation (2) and defined by (6). Then

$$\mathcal{G}_n(\theta) = (1-\delta)W_n^{-1}\mathcal{S}_n(\theta)G_n(\theta) \tag{19}$$

Theorem 5 may be used to provide a second view of the two sets of PMM first-order conditions defining $\hat{\theta}$ and

[5] When the system is just identified, the moment conditions can be set to zero under fixed weights, which results in the same solution provided by GMM.

the optimal weights. Substituting equation (19) into equation (7) gives an equivalent set of first-order conditions.

$$\mathcal{M}_n(\hat{\theta})'\mathcal{S}_n(\hat{\theta})G_n(\hat{\theta}) = 0 \tag{20}$$

Equation (20) rewrites PMM's first-order conditions as a linear combination of the equal-weight sample average moment vector set to zero. The first-order conditions depend on the weights through $\mathcal{M}_n(\hat{\theta})$ and $\mathcal{S}_n(\hat{\theta})$ and the optimal parameter vector and optimal weights are estimated simultaneously. Note, the PMM first-order conditions still rely on a two-step approach because $\mathcal{S}_n(\theta)$ depends on W_n. Theorem 5 may also be used to rewrite the weights in a form resembling the EL weights in equation (17).

$$\omega_i(\theta) = \frac{1}{n}\left(\frac{1}{1+\delta g_i(\theta)'\mathcal{S}_n(\theta)G_n(\theta) - \delta(1-\delta)G_n(\theta)'\mathcal{S}_n(\theta)W_n^{-1}\mathcal{S}_n(\theta)G_n(\theta)}\right) \tag{21}$$

Theorem 6 ***(Equivalency)***

Suppose the conditions of Theorem 5 are met. Let δ approach zero. Then PMM's first-order conditions limit to those of GMM. Alternatively, let δ approach one. Then PMM's first-order conditions limit to those of EL.

According to Theorem 6, GMM and EL are special cases of PMM. Through δ, PMM provides a continuum of estimators with GMM and EL at the extremes. When δ approaches one, PMM's weights limit to those of EL. When δ approaches zero, PMM's weights limit to the fixed weights of GMM.

2.4 Higher-order Bias and Tests of Overidentifying Restrictions

Let $\mathbf{P}^{\perp}_{M} \equiv \mathbb{I}_m - \Sigma^{-\frac{1}{2}}M\left(M'\Sigma^{-1}M\right)^{-1}M'\Sigma^{-\frac{1}{2}}$ be the projection matrix orthogonal to the space spanned by the asymptotic normalized Jacobian, where $\Sigma^{-\frac{1}{2}}$ is the Cholesky decomposition of Σ^{-1}. Define $\Omega \equiv M'\Sigma^{-1}M$, $\tilde{\Omega} \equiv M'\tilde{W}M$, $\Upsilon \equiv \Omega^{-1}M'\Sigma^{-1}$, and $\tilde{\Upsilon} \equiv \tilde{\Omega}^{-1}M'\tilde{W}$. The following four definitions will be used to compare the higher-order bias of GMM, EL, and PMM

$$\begin{aligned}
B_I &= n^{-1}\Upsilon\left(\mathbb{E}\left[m_i(\theta_0)\Upsilon g_i(\theta_0)\right] - a/2\right) \\
B_M &= -n^{-1}\Omega^{-1}\mathbb{E}\left[m_i(\theta_0)'^{-\frac{1}{2}}\mathbf{P}^{\perp}_{M}\Sigma^{-\frac{1}{2}}g_i(\theta_0)\right] \\
B_\Sigma &= n^{-1}\Upsilon\mathbb{E}\left[v_i(\theta_0)\Sigma^{-\frac{1}{2}}\mathbf{P}^{\perp}_{M}\Sigma^{-\frac{1}{2}}g_i(\theta_0)\right] \\
B_W &= -n^{-1}\Upsilon\mathbb{E}\left[\sum_{j=1}^{k}\frac{\partial V_n(\theta_0)}{\partial\theta_j}(\tilde{\Upsilon}-\Upsilon)'e_j\right],
\end{aligned} \tag{22}$$

where e_j is the jth unit vector, a is an $m \times 1$ vector with

$$a_j \equiv \mathrm{tr}\left(\Omega\mathbb{E}\left[\frac{\partial^2 g_{ij}(\theta_0)}{\partial\theta\partial\theta'}\right]\right) \qquad (j = 1, ..., m),$$

and $g_{ij}(\theta_0)$ denotes the jth element of $g_i(\theta_0)$.

Newey and Smith (2004) show that GMM's higher-order $\mathcal{O}\left(n^{-1}\right)$ bias is

$$\underset{\text{GMM}}{\text{Bias}}(\hat{\theta}) = B_I + B_M + B_\Sigma + B_W. \tag{23}$$

The first component, B_I, is the bias for an estimator with the optimal (non-random) transformation $M'\Sigma^{-1}$ applied to the moment conditions $G_n(\theta_0)$. The second component and third components, B_M and B_Σ, are the bias due to the estimation of the Jacobian and second moment matrices. The final term, B_W, is the bias due to the choice of the first-round weighting matrix.

Newey and Smith (2004) show that EL's higher-order $\mathcal{O}\left(n^{-1}\right)$ bias is

$$\underset{\text{EL}}{\text{Bias}}(\hat{\theta}) = B_I. \tag{24}$$

Comparing the higher-order bias of EL against GMM, three of the four components disappear. EL provides *efficient* estimates of the Jacobian and second moment matrices, in that its estimates do not contribute to the higher order bias. As a one-step estimator, EL has no bias component from the inclusion of a preliminary weighting matrix. Interpretation of the bias requires caution. Although EL certainly has fewer higher-order bias contributors, EL may not have the smaller higher-order bias in a given application. The four bias-terms do not necessarily have the same sign and the three missing bias terms may cancel out some of the B_I bias.

Theorem 7 provides the higher-order $\mathcal{O}\left(n^{-1}\right)$ bias for the family of PMM estimators.

Theorem 7 ***(Higher-order $\mathcal{O}\left(n^{-1}\right)$ Bias)***

Suppose $\hat{\theta}$ satisfies equation (7) where $W_n \xrightarrow{p} \Sigma^{-1}$, $\hat{\theta} \xrightarrow{p} \theta_0$, and $\tilde{W}$ is the preliminary first-round weighting matrix. Then

$$\underset{PMM}{Bias}(\hat{\theta}) = B_I + (1-\delta)B_M + \left(1-\delta^2\right)B_\Sigma + (1-\delta)B_W. \tag{25}$$

The PMM bias has a similar form to that of GMM and EL and is comprised of the four bias components discussed above. The bias for PMM is a continuous function of the δ parameter. As δ approaches zero, the coefficients on B_M, B_Σ, and B_W approach one and the PMM bias approaches that of GMM. Similarly, as δ approaches one, the coefficients to B_M, B_Σ, and B_W approach zero and the PMM bias approaches that of EL.

The objective function (1) value serves as a test statistic of overidentifying restrictions in PMM estimation when the parameters are estimated efficiently.

Theorem 8 ***(Test of Overidentifying Restrictions)***

Suppose $\hat{\theta}$ satisfies equation (7) and $\hat{\theta} \xrightarrow{p} \theta_0$. Further suppose that $W_n \xrightarrow{p} \Sigma^{-1}$. Then

$$\underset{PMM}{Q}(\hat{\theta}) \xrightarrow{d} \chi^2_{m-k}. \tag{26}$$

The objective function (1) has two components. The first component is the quadratic penalty of GMM that, when appropriately scaled, is similar in structure to GMM's test statistic of overidentifying restrictions. The second component is the KLIC penalty of EL that, when appropriately scaled, is similar in structure to EL's likelihood ratio test statistic of overidentifying restrictions.

3 PMM versus EL in Practice

The EL objective function is undefined for non-positive weights. Yet, the first-order conditions may imply negative weights are optimal, conditional on θ.[6] The following must be satisfied for EL to be defined at a given parameter vector.

$$1 + g_i(\theta)'\mathcal{V}_n^{-1}(\theta)G_n(\theta) \geq 0 \tag{27}$$

Unfortunately, the above restriction may be violated at the population parameter vector. The probability of the population parameter vector residing outside the support of the parameter space decreases with sample size because $G_n(\theta_0) = \mathcal{O}\left(n^{-\frac{1}{2}}\right)$ and increases with model misspecification because $E[G_n(\theta_0)] \neq 0$ in a misspecified model. If $\sup_{x\in\mathcal{X}} g(x, \theta_0)$ is not bounded, then even with a correctly specified model, restriction (27) may not be satisfied at the population parameter value as $n \to \infty$.

Violation of condition (27) leads to two possible outcomes. First, the condition may be violated for only a subset of the parameter space, restricting the parameter space. Second, the restriction could be violated for the entire parameter space, which occurs when the convex hull of the moments $g_1(\theta), \ldots, g_n(\theta)$ does not contain the origin for any choice of θ. In the second scenario, EL is undefined for the given model and sample. Non-positive weights also leave PMM's objective function undefined. For PMM to be defined for a given parameter vector, the following must hold

$$1 + \delta g_i(\theta)'\mathcal{S}_n(\theta)G_n(\theta) - \delta(1-\delta)G_n(\theta)'\mathcal{S}_n(\theta)W_n^{-1}\mathcal{S}_n(\theta)G_n(\theta) \geq 0. \tag{28}$$

Fortunately, PMM provides the econometrician with a solution to the negative weight problem. The coefficient δ may be selected to ensure positive weights in some region of the parameter space.

Using PMM estimation to increase the support of the parameter space may be desirable even when the parameter

[6]See Bertille, Bonnal, and Renault (2007) and Liu and Chen (2010) for more on negative weights in EL estimation.

space support under EL is non-empty. Suppose

$$nG_n(\theta_1)'\mathcal{V}_n^{-1}(\theta_1)G_n(\theta_1) < nG_n(\theta_2)'\mathcal{V}_n^{-1}(\theta_2)G_n(\theta_2)$$

for parameter vectors θ_1 and θ_2. Ideally, the first parameter vector will be selected. However, for one or more observations the following could be true

$$g_i(\theta_1)'\mathcal{V}_n^{-1}(\theta_1)G_n(\theta_1) < -1$$

and

$$g_i(\theta_2)'\mathcal{V}_n^{-1}(\theta_2)G_n(\theta_2) \geq -1.$$

EL is undefined at θ_1, so EL selects θ_2 to avoid the negative weight(s). This scenario is more likely to occur when the sample size is small, the number of moments is large, the moments are unbounded, or the model is misspecified. By shrinking the weights through δ, PMM offers a less restricted parameter space than EL. We may select a δ to ensure positive weights for both θ_1 and θ_2, and PMM will select θ_1 over θ_2.

Why not let δ approach zero and avoid the scenario outlined in the previous paragraph entirely? Because, doing so prohibits us from attaining the desirable higher-order properties offered by EL. PMM allows the econometrician to continuously exchange the higher-order properties provided by EL for the less restricted parameter space associated with GMM by adjusting δ.

The increased support for the parameter space provided by PMM also offers a practical advantage for operationalizing the estimator. To begin the numerical optimization, the econometrician must supply a starting value. EL and PMM might be undefined at the supplied starting parameter vector due to a restricted parameter space. When this problem occurs, the econometrician must search for a starting value in which the estimators are defined; the search becomes increasingly difficult as the number of parameters increases. The larger support provided by PMM can make finding an appropriate starting value easier.

3.1 Monte Carlo Simulations

We use the Hall and Horowitz (1996) model as modified by Schennach (2007) to compare EL to PMM with $\delta = 0.5$.[7] The orthogonality conditions are

$$g_i(\theta) = r(x_i, \theta)\left[1 \quad x_{i2} \quad (x_{i3}-1) \quad \ldots \quad (x_{iK}-1)\right]', \tag{29}$$

[7] The GMM estimator posseses notoriously poor properties when the sample size is small or number of moments is large, so GMM has been left out of the comparison.

where $r(x_i, \theta) = exp(-0.72 - (x_{i1} + x_{i2})\theta + 3x_{i2}) - 1$. When $(x_{i1}, x_{i2}) \sim N(0, 0.16\mathbb{I}_2)$ and $x_{ik} \sim \chi_1^2$, for $k = 3, \ldots, K$ the moment conditions are satisfied at $\theta_0 = 3$. The $\mathcal{O}(n^{-1})$ term does not trivially vanish because the third moments of all elements of $g_i(\theta)$ are non-zero and $g_i(\theta)$ is a nonlinear function of θ.

The simulations use small samples ($n = 25$, 50, and 100) with a large number of moment conditions to investigate the properties of PMM when the restricted parameter space is likely to be problematic. We calculate the realized estimator mean, volatility (standard deviation of the estimates across simulations), and Root Mean Squared Error (RMSE) for PMM and EL. The size of the rejection region for the relevant test of the overidentifying restrictions and the probability of rejecting the population parameter value are also calculated. Appendix B details how to implement PMM and provides the complete simulation results; here, we only summarize the main findings.

PMM dramatically outperforms EL when the number of observations is low and the number of moments is high. For example, with 25 observations and 5 moments, the volatility of the EL estimates is approximately 0.94 while the volatility of PMM's estimates equals 0.85. At the 5 percent level, PMM rejects the population parameter value 36 percent of the time and EL rejects 44 percent of the time. PMM and EL reject the model at the 5 percent level 61 and 48 percent of the time, respectively. The average parameter estimate for PMM is about 2.9, while EL is not close at 3.5. The results are similar for the 50 observation and 10 moment model. The volatility of EL's estimates equals 0.72, which is approximately 18 percent higher than for the PMM estimates. PMM and EL reject the population parameter value at the 5 percent level approximately 26 and 43 percent of the time, respectively. The model is rejected at the 5 percent level 63 percent of the time for PMM and 73 percent of the time for EL. For the simulations with 100 observations, EL's estimates are less volatile than PMM's when the number of moments are low, but the volatility of EL's estimates increases more than for PMM as the number of moment conditions grows. EL's estimator volatility increases from 0.30 under 2 moments to 0.45 under 20 moments. PMM's estimator volatility increases from 0.32 under 2 moments to 0.37 with 20 moment conditions.[8]

When the sample size is large and the number of moments are low, EL sometimes slightly outperforms PMM (with $\delta = 0.5$). In most of these cases, however, EL and PMM have similar properties, with hypothesis tests and tests of model-specification being equally mis-sized.

Overall, the simulations highlight the improvements achieved by expanding the parameter space through shrinkage. The restricted parameter space due to negative weight avoidance can make EL have more volatile estimates and greater over-rejection of hypothesis tests. Shrinking the weights towards fixed weights through PMM can provide more desirable estimator properties.

Although the simulations consider the particular PMM estimator with $\delta = 0.5$, the results provide some guidance on how to select δ. If the sample size is small or there are many moments, then shrinkage towards GMM might be best. While model estimation using larger samples or less moment conditions may have more desirable properties

[8] Appendix B visually depicts the PMM and EL parameter estimates for each set of simulations. In the 25x5, 50x5, 100x15, and 100x20 trials, PMM's density functions have higher peaks and tighter distributions than those of EL. The differences between EL and PMM are striking. The distribution for EL is not even close to being centered on the true parameter value, demonstrating the superiority of PMM in these particular examples.

when PMM is selected to resemble EL. This is only an intuitive way to think about δ. We hope future research will uncover how to optimally select δ.

4 Conclusion

This paper presents a new family of estimators by merging the objective functions of GMM and EL. GMM allows the elements of the sample average moment vector to deviate from zero and requires the sample weights to be fixed. EL allows the sample weights to vary and forces the weighted moments to equal zero. The PMM family of estimators allows both weights and sample moments to deviate from n^{-1} and 0 and measures the respective deviations with EL's KLIC penalty function and GMM's quadratic penalty function. Through a free parameter, a continuum of estimators is obtained with GMM and EL at the extremes. When the sample size is small and the number of moments is large, the new estimator outperforms EL in Monte Carlo simulations.

References

Bertille, A., H. Bonnal, and E. Renault (2007): "On the Efficient Use of the Informational Content of Estimating Equations: Implied Probabilities and Euclidean Empirical Likelihood," *Journal of Econometrics*, 138, 461–487.

Corocoran, S. A. (1998): "Bartlett Adjustment of Empirical Discrepancy Statistics," *Biometrika*, 85, 967–972.

Hall, P., and J. L. Horowitz (1996): "Bootstrap Critical Values for Tests based on Generalized-Method-Of-Moments Estimators," *Econometrica*, 64, 891–916.

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 Economics Statistics*, 14, 262–280.

Imbens, G. W. (1997): "One-Step Estimators for Over-Identified Generalized Method of Moments Models," *Review of Economic Studies*, 64, 359–383.

Imbens, G. W., P. Johnson, and R. H. Spady (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," *Econometrica*, 65, 861–874.

Liu, Y., and J. Chen (2010): "Adjusted Empirical Likelihood with High-Order Precision," *The Annals of Statistics*, 38(3), 1341–1362.

Newey, W. K., and D. L. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," *In: Engle, R. F., McFadden, D. L. (Eds.), Handbook of Econometrics*, 4, 2111–2245.

Newey, W. K., and R. J. Smith (2004): "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica*, 72, 219–255.

Owen, A. (1988): "Empirical Likelihood Ratio Confidence Intervals for a Single Functional," *Biometrika*, 75, 237–249.

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

Ragusa, G. (2008): "Minimum Divergence, Generalized Empirical Likelihoods, and Higher Order Expansions," *Working Paper*.

Schennach, S. M. (2007): "Point Estimation with Exponentially Tilted Empirical Likelihood," *Annals of Statistics*, 35, 634–672.