Journal of Econometrics 152 (2009) 28–36

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Choosing instrumental variables in conditional moment restriction models

Stephen G. Donald[a], Guido W. Imbens[b], Whitney K. Newey[c,*]

[a] Department of Economics, University of Texas, United States
[b] Department of Economics, Harvard University, United States
[c] Department of Economics, MIT, United States

ARTICLE INFO

Article history:
Received 1 January 2002
Accepted 1 October 2008
Available online 5 March 2009

JEL classification:
C13
C30

Keywords:
Conditional moment restrictions
Generalized method of moments
Generalized empirical likelihood
Mean squared error

ABSTRACT

Properties of GMM estimators are sensitive to the choice of instrument. Using many instruments leads to high asymptotic asymptotic efficiency but can cause high bias and/or variance in small samples. In this paper we develop and implement asymptotic mean square error (MSE) based criteria for instrument selection in estimation of conditional moment restriction models. The models we consider include various nonlinear simultaneous equations models with unknown heteroskedasticity. We develop moment selection criteria for the familiar two-step optimal GMM estimator (GMM), a bias corrected version, and generalized empirical likelihood estimators (GEL), that include the continuous updating estimator (CUE) as a special case. We also find that the CUE has lower higher-order variance than the bias-corrected GMM estimator, and that the higher-order efficiency of other GEL estimators depends on conditional kurtosis of the moments.

1. Introduction

It is important to choose carefully the instrumental variables for estimating conditional moment restriction models. Adding instruments increases asymptotic efficiency but also increases small sample bias and/or variance. We account for this trade-off by using a higher-order asymptotic mean-square error (MSE) of the estimator to choose the instrument set. We derive the higher-order MSE for GMM, a bias corrected version of GMM (BGMM), and generalized empirical likelihood (GEL). For simplicity we impose a conditional symmetry assumption, that third conditional moments of disturbances are zero, and use a large number of instrument approximations. We also consider the effect of allowing identification to shrink with the sample size n at a rate slower than $1/\sqrt{n}$. The resulting MSE expressions are quite simple and straightforward to apply in practice to choose the instrument set. The MSE criteria given here also provide higher order efficiency comparisons. We find that continuously updated GMM estimator (CUE) is higher-order efficient relative to BGMM. We also find that the higher order efficiency of the GEL estimators depends on conditional kurtosis, with all GEL estimators having the same higher-order variance when disturbances are Gaussian. With Gaussian disturbances and homoskedasticity, Rothenberg (1996) showed that empirical likelihood (EL) is higher order efficient relative to BGMM. Our efficiency comparisons generalize those of Rothenberg (1996) to other GEL estimators and heteroskedastic, non Gaussian disturbances. These efficiency results are different than the higher-order efficiency result for EL from Newey and Smith (2004), where all estimators are biased corrected, the number of moments is fixed, and symmetry is not imposed.

Donald and Newey (2001) gives analogous results for linear instrumental variable estimators with homoskedasticity. This paper focuses on GMM estimators with heteroskedasticity, leading to heteroskedasticity robust MSE that include terms from estimation of the weight matrix. Our MSE criteria is like that of Nagar (1959), being the MSE of leading terms in a stochastic expansion of the estimator. This approach is well-known to give the same answer as the MSE of leading terms in an Edgeworth expansion, under suitable regularity conditions (e.g. Rothenberg (1984)). The many-instrument simplification seems appropriate for many applications where there is a large number of potential instrumental variables. We also assume symmetry, in the sense that conditional third moments of the disturbances are zero. This symmetry assumption greatly simplifies calculations. Also, relaxing it may not change the results much, e.g. because the bias from asymmetry tends to be smaller than other bias sources for large numbers of moment conditions, see Newey and Smith (2004).

Choosing moments to minimize MSE may help reduce misleading inferences that can occur with many moments. For GMM, the MSE explicitly accounts for an important bias term (e.g. see Hansen et al. (1996), Newey and Smith (2004)), so choosing moments to

[*] Corresponding address: Department of Economics, Massachusetts Institute of Technology, Room E52-262D, 50 Memorial Drive, Cambridge, MA 02142, United States. Tel.: +1 617 253 6420; fax: +1 617 253 1330.
E-mail address: wnewey@mit.edu (W.K. Newey).

0304-4076/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2008.10.013

minimize MSE avoids cases where asymptotic inferences are poor due to the bias being large relative to the standard deviation. For GEL, the MSE explicitly accounts for higher order variance terms, so that choosing instruments to minimize MSE helps avoid underestimated variances. However, the criteria we consider do not generally provide the most accurate inference, as recently pointed out by Sun et al. (2007) in another context.

The problem addressed in this paper is different than that considered by Andrews (1999), where selection of the largest set of valid moments was considered. Here the problem is how to choose among moments known to be valid. Choosing among valid moments is important when there are many thought to be equally valid. Examples include various natural experiment studies, where multiple instruments are often available, as well as intertemporal optimization models, where all lags may serve as instruments.

In Section 2 we describe the estimators we consider and present the criteria we develop for choosing the moments. We also compare the criteria for different estimators, which corresponds to the MSE comparison for the estimators, finding that the CUE has smaller MSE than bias corrected GMM. In Section 3 we give the regularity conditions used to develop the approximate MSE, give the formal results, and consider higher order efficiency comparisons. A small scale Monte Carlo experiment is conducted in Section 4. Concluding remarks are offered in Section 5.

2. The model and estimators

We consider a model of conditional moment restrictions like Chamberlain (1987). To describe the model let z denote a single observation from an i.i.d. sequence $(z_1, z_2, \ldots)$, β a $p \times 1$ parameter vector, and $\rho(z, \beta)$ a scalar that can often be thought of as a residual.[1] The model specifies a subvector of x, acting as conditioning variables, such that for a value β_0 of the parameters

$$E[\rho(z, \beta_0)|x] = 0,$$

where $E[\cdot]$ the expectation taken with respect to the distribution of z_i.

To form GMM estimators we construct unconditional moment restrictions using a vector of K functions of x given by $q^K(x) = (q_{1K}(x), \ldots, q_{KK}(x))'$. Let $g(z, \beta) = \rho(z, \beta)q^K(x)$. Then the unconditional moment restrictions

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

are satisfied. Let $g_i(\beta) \equiv g(z_i, \beta)$, $\bar{g}_n(\beta) \equiv n^{-1}\sum_{i=1}^n g_i(\beta)$, and $\hat{\Upsilon}(\beta) \equiv n^{-1}\sum_{i=1}^n g_i(\beta)g_i(\beta)'$. A two-step GMM estimator is one that satisfies, for some preliminary consistent estimator $\tilde{\beta}$ for β_0,

$$\hat{\beta}^H = \arg\min_{\beta\in\mathcal{B}} \bar{g}_n(\beta)'\hat{\Upsilon}(\tilde{\beta})^{-1}\bar{g}_n(\beta), \qquad (2.1)$$

where $\mathcal{B}$ denotes the parameter space. For our purposes $\tilde{\beta}$ could be some other GMM estimator, obtained as the solution to an analogous minimization problem with $\hat{\Upsilon}(\tilde{\beta})^{-1}$ replaced by a different weighting matrix, such as $\tilde{W}_0 = [\sum_{i=1}^n q^K(x_i)q^K(x_i)'/n]^{-1}$.

The MSE of the estimators will depend not only on the number of instruments but also on their form. In particular, instrumental variables that better predict the optimal instruments will help to lower the asymptotic variance of the estimator for a given K. Thus, for each K it is good to choose $q^K(x)$ that are the best predictors. Often it will be evident in an application how to choose the instruments in this way. For instance, lower order approximating functions (e.g. linear and quadratic) often provide the most information, and so should be used first. Also, the main terms may often be more important than interactions.

The instruments need not form a nested sequence. Letting $q_{kK}(x)$ depend on K allows different groups of instrumental variables to be used for different values of K. Indeed, K fills a double role here, as the index of the instrument set as well as the number of instruments. We could separate these roles by having a separate index for the instrument set. Instead here we allow for K to not be selected from all the integers, and let K fulfill both roles. This restricts the sets of instruments to each have a different number of instruments, but is often true in practice. Also, this double role for K restricts the number of instrument sets we can select among, as seems important for the asymptotic theory.

As demonstrated by Newey and Smith (2004), the correlation of the residual with the derivative of the moment function leads to an asymptotic bias that increases linearly with K. They suggested an approach that removes this bias (as well as other sources of bias that we will ignore for the moment). This estimator can be obtained by subtracting an estimate of the bias from the GMM estimator and gives rise to what we refer to as the bias adjusted GMM estimator (BGMM). To describe it, let $q_i = q^K(x_i)$, $\rho_i(\beta) = \rho(z_i, \beta)$, and

$$\hat{\rho}_i = \rho_i(\hat{\beta}^H), \qquad \hat{y}_i = [\partial\rho_i(\hat{\beta}^H)/\partial\beta], \qquad \hat{y} = [\hat{y}_1, \ldots, \hat{y}_n]',$$

$$\hat{\Gamma} = \sum_{i=1}^n q_i\hat{y}_i'/n,$$

$$\hat{\Sigma} = \hat{\Upsilon}(\hat{\beta}^H)^{-1} - \hat{\Upsilon}(\hat{\beta}^H)^{-1}\hat{\Gamma}(\hat{\Gamma}'\hat{\Upsilon}(\hat{\beta}^H)^{-1}\hat{\Gamma})^{-1}\hat{\Gamma}'\hat{\Upsilon}(\hat{\beta}^H)^{-1}.$$

The BGMM estimator is

$$\hat{\beta}^B = \hat{\beta}^H + (\hat{\Gamma}'\hat{\Upsilon}(\hat{\beta}^H)^{-1}\hat{\Gamma})^{-1}\sum_{i=1}^n \hat{y}_i\hat{\rho}_i q_i'\hat{\Sigma}q_i.$$

Also as shown in Newey and Smith (2004), GEL estimators have less bias than GMM when K is large. We follow the description of these estimators given in that paper. Let $s(v)$ be a concave function with domain that is an open interval $\mathcal{V}$ containing 0, $s_j(v) = \partial^j s(v)/\partial v^j$, and $s_j = s_j(0)$. We impose the normalizations $s_1 = s_2 = -1$. Define the GEL estimator as

$$\hat{\beta}^{\mathrm{GEL}} = \arg\min_{\beta\in B}\max_{\lambda\in\hat{\Lambda}_n(\beta)}\sum_{i=1}^n s(\lambda' g_i(\beta))$$

where, $\hat{\Lambda}_n(\beta) = \{\lambda : \lambda' g_i(\beta) \in \mathcal{V}, i = 1, \ldots, n\}$. This estimator includes as a special cases: empirical likelihood (EL, Qin and Lawless (1994), Owen (1988)), where $s(v) = \ln(1 - v)$, exponential tilting (ET, Imbens et al. (1998), Kitamura and Stutzer (1997)), where $s(v) = -\exp(v)$, and the continuous updating estimator (CUE Hansen et al., 1996), where $s(v) = -(1 + v)^2/2$. As we will see the MSE comparisons between these estimators depend on s_3, the third derivative of the s function, where

$$\mathrm{CUE}: s_3 = 0, \qquad \mathrm{ET}: s_3 = -1, \qquad \mathrm{EL}: s_3 = -2.$$

2.1. Instrument selection criteria

The instrument selection is based on minimizing the approximate mean squared error (MSE) of a linear combination $\hat{t}'\hat{\beta}$ of a GMM estimator or GEL estimator $\hat{\beta}$, where $\hat{t}$ is some vector of (estimated) linear combination coefficients. To describe the criteria, some additional notation is required. Let $\tilde{\beta}$ be some preliminary estimator, $\tilde{\rho}_i = \rho_i(\tilde{\beta})$, $\tilde{y}_i = \partial\rho_i(\tilde{\beta})/\partial\beta$, and

$$\hat{\Upsilon} = \sum_{i=1}^n \tilde{\rho}_i^2 q_i q_i'/n, \qquad \hat{\Gamma} = \sum_{i=1}^n q_i\tilde{y}_i'/n,$$

$$\hat{\Omega} = \hat{\Gamma}'\hat{\Upsilon}^{-1}\hat{\Gamma}, \qquad \hat{\tau} = \hat{\Omega}^{-1}\hat{t},$$

[1] The extension to the vector of residuals case is straightforward.

$$\tilde{d}_i = \hat{\Gamma}' \left(\sum_{j=1}^{n} q_j q_j'/n \right)^{-1} q_i, \qquad \tilde{\eta}_i = \tilde{y}_i - \tilde{d}_i,$$

$$\hat{\xi}_{ij} = q_i' \hat{\Upsilon}^{-1} q_j / n, \qquad \hat{D}_i^* = \hat{\Gamma}' \hat{\Upsilon}^{-1} q_i,$$

$$\hat{\Lambda}(K) = \sum_{i=1}^{n} \hat{\xi}_{ii} \left(\hat{\tau}' \tilde{\rho}_{\beta i} \right)^2, \qquad \hat{\Pi}(K) = \sum_{i=1}^{n} \hat{\xi}_{ii} \tilde{\rho}_i (\hat{\tau}' \tilde{\eta}_i),$$

$$\hat{\Phi}(K) = \sum_{i=1}^{n} \hat{\xi}_{ii} \left\{ \hat{\tau}' (\hat{D}_i^* \tilde{\rho}_i^2 - \tilde{\rho}_{\beta i}) \right\}^2 - \hat{\tau}' \hat{\Gamma}' \hat{\Upsilon}^{-1} \hat{\Gamma} \hat{\tau}.$$

The criteria for the GMM estimator, without a bias correction, is

$$S_{\mathrm{GMM}}(K) = \hat{\Pi}(K)^2/n + \hat{\Phi}(K).$$

Also, let

$$\hat{\Pi}_B(K) = \sum_{i,j=1}^{n} \tilde{\rho}_i \tilde{\rho}_j (\hat{\tau}' \tilde{\eta}_i)(\hat{\tau}' \tilde{\eta}_j) \hat{\xi}_{ij}^2 = \mathrm{tr}(\tilde{Q} \hat{\Upsilon}^{-1} \tilde{Q} \hat{\Upsilon}^{-1}),$$

$$\hat{\Xi}(K) = \sum_{i=1}^{n} \{5(\hat{\tau}' \hat{d}_i)^2 - \tilde{\rho}_i^4 (\hat{\tau}' \hat{D}_i^*)^2\} \hat{\xi}_{ii},$$

$$\hat{\Xi}_{\mathrm{GEL}}(K) = \sum_{i=1}^{n} \{3(\hat{\tau}' \hat{d}_i)^2 - \tilde{\rho}_i^4 (\hat{\tau}' \hat{D}_i^*)^2\} \hat{\xi}_{ii},$$

where $\tilde{Q} = \sum_{i=1}^{n} \tilde{\rho}_i (\hat{\tau}' \tilde{\eta}_i) q_i q_i'$. The criteria for the BGMM and GEL estimators are

$$S_{\mathrm{BGMM}}(K) = \left[\hat{\Lambda}(K) + \hat{\Pi}_B(K) + \hat{\Xi}(K) \right] /n + \hat{\Phi}(K),$$

$$S_{\mathrm{GEL}}(K) = \left[\hat{\Lambda}(K) - \hat{\Pi}_B(K) + \hat{\Xi}(K) + s_3 \hat{\Xi}_{\mathrm{GEL}}(K) \right] /n + \hat{\Phi}(K).$$

For each of the estimators, our proposed instrument selection procedure is to choose K to minimize $S(K)$. As we will show this will correspond to choosing K to minimize the higher-order MSE of the estimator.

Each of the terms in the criteria has an interpretation. For GMM, $\hat{\Pi}(K)^2/n$ is an estimate of a squared bias term from Newey and Smith (2004). Because $\hat{\xi}_{ii}$ is of order K this squared bias term has order K^2/n. The $\hat{\Phi}(K)$ term in the GMM criteria is an asymptotic variance term. Its size is related to the asymptotic efficiency of a GMM estimator with instruments $q^K(x)$. As K grows these terms will tend to shrink, reflecting the reduction in asymptotic variance that accompanies using more instruments. The form of $\hat{\Phi}(K)$ is analogous to a Mallows criterion, in that it is a variance estimator plus a term that removes bias in the variance estimator.

The terms that appear in $S(K)$ for BGMM and GEL are all variance terms. No bias terms are present because, as discussed in Newey and Smith (2004), under symmetry GEL removes the GMM bias that grows with K. As with GMM, the $\hat{\Phi}(K)$ term accounts for the reduction in asymptotic variance that occurs from adding instruments. The other terms are higher-order variance terms, that will be of order K/n, because $\hat{\xi}_{ii}$ is of order K. The sum of these terms will generally increase with K, although this need not happen if $\hat{\Xi}(K)$ is too large relative to the other terms. Here $\hat{\Xi}(K)$ is an estimator of

$$\Xi(K) = \sum_{i=1}^{n} \xi_{ii} \left(\tau' d_i \right)^2 \{5 - E(\rho_i^4 | x_i)/\sigma_i^4\},$$

where $\rho_i = \rho(z_i, \beta_0)$ and $\sigma_i^2 = E[\rho_i^2 | x_i]$. As a result if the kurtosis of ρ_i is too high the higher-order variance of the BGMM and GEL estimators would actually decrease as K increases. This phenomenon is similar to that noted by Koenker et al. (1994) for the exogenous linear case. In this case the criteria could fail to be useful as a means of choosing the number of moment conditions, because they would monotonically decrease with K.

The terms $\hat{\Xi}(K)$ and $\hat{\Xi}_{\mathrm{GEL}}(K)$ arise from heteroskedasticity consistent estimation of the optimal weight matrix for GMM. When they are positive, as they will be if the kurtosis is not too large, they represent a penalty for using a weight matrix that is optimal under heteroskedasticity. It has been noted in simulations (including those below) that using a heteroskedasticity consistent weight matrix when it is not needed tends to degrade the performance of GMM estimators. The presence of these extra terms provides a theoretical counterpart to this feature of GMM.

It is also interesting to note that $\hat{\Xi}(K)$ and $\hat{\Xi}_{\mathrm{GEL}}(K)$ will be small relative to the other terms when identification shrinks, meaning that d_i goes to zero as the sample size grows. Thus, under shrinking identification the estimation of the weight matrix does not need to be accounted for in the MSE. This fact simplifies considerably the instrument choice criteria and was also noted in Newey and Windmeijer (2009).

It is interesting to compare the size of the criteria for different estimators, which comparison parallels that of the MSE. As previously noted, the squared bias term for GMM, which is $\hat{\Pi}(K)^2$, has the same order as K^2/n. In contrast the higher-order variance terms in the BGMM and GEL estimators generally have order K/n, because that is the order of ξ_{ii}. Consequently, for large K the MSE criteria for GMM will be larger than the MSE criteria for BGMM and GEL, meaning the BGMM and GEL estimators are preferred over GMM. This comparison parallels that in Newey and Smith (2004) and in Imbens and Spady (2005).

One interesting result is that for the CUE, where $s_3 = 0$, the MSE criteria is smaller than it is for BGMM, because $\hat{\Pi}_B(K)$ is positive. Thus we find that the CUE dominates the BGMM estimator, in terms of higher-order MSE, i.e. the CUE is higher-order efficient relative to BGMM. This result is analogous to the higher-order efficiency of the limited information maximum likelihood estimator relative to the bias corrected two-stage least squares estimator that was found by Rothenberg (1983).

The comparison of the higher-order MSE for the CUE and the other GEL estimators depends on the kurtosis of the residual. For conditionally normal ρ_i we have $E[\rho_i^4 | x_i] = 3\sigma_i^4$ and consequently $\hat{\Xi}_{\mathrm{GEL}}(K)$ will converge to zero for each K, and all the GEL estimators have the same higher-order MSE. When there is excess kurtosis, with $E[\rho_i^4 | x_i] > 3\sigma_i^4$, ET will have larger MSE than the CUE, and EL will have larger MSE than ET, with these rankings being reversed when $E[\rho_i^4 | x_i] < 3\sigma_i^4$. These comparisons parallel those of Newey and Smith (2004) for a heteroskedastic linear model with exogeneity.

The case with no endogeneity has some independent interest. In this setting the GMM estimator can often be interpreted as using "extra" moment conditions to improve efficiency in the presence of heteroskedasticity of unknown functional form. Here the MSE criteria will give a method for choosing the number of moments used for this purpose. Dropping the bias terms, which are not present in exogenous cases, leads to criteria of the form

$$S_{\mathrm{GMM}}(K) = \hat{\Xi}(K)/n + \hat{\Phi}(K)$$

$$S_{\mathrm{GEL}}(K) = \left[\hat{\Xi}(K) + s_3 \hat{\Xi}_{\mathrm{GEL}}(K) \right] /n + \hat{\Phi}(K).$$

Here GMM and the CUE have the same higher-order variance, as was found by Newey and Smith (2004). Also, as in the general case, these criteria can fail to be useful if there is so much kurtosis that the higher order variance terms shrink as K grows.

3. Assumptions and MSE results

As in Donald and Newey (2001), the MSE approximations are based on a decomposition of the form,

$$n t' (\hat{\beta} - \beta_0)(\hat{\beta} - \beta_0)' t = \hat{Q}(K) + \hat{R}(K), \tag{3.2}$$

$$E(\hat{Q}(K) | X) = t' \Omega^{*-1} t + S(K) + T(K),$$

$$[\hat{R}(K) + T(K)]/S(K) = o_p(1), \quad K \to \infty, \ n \to \infty$$

where $X = [x_1, \ldots, x_n]'$, $t = plim(\hat{t})$, $\Omega^* = \sum_{i=1}^{n} \sigma_i^{-2} d_i d_i'/n$, and $d_i = E[\partial \rho_i(\beta_0)/\partial \beta | x_i]$. Here $S(K)$ is part of conditional MSE of $\hat{Q}$ that depends on K and $\hat{R}(K)$ and $T(K)$ are remainder terms that goes to zero faster than $S(K)$. Thus, $S(K)$ is the MSE of the dominant terms for the estimator. All calculations are done assuming that K increases with n. The largest terms increasing and decreasing with K are retained. Compared to Donald and Newey (2001) we have the additional complication that none of our estimators has a closed form solution. Thus, we use the first-order condition that defines the estimator to develop approximations to the difference $\sqrt{n}t'(\hat{\beta} - \beta_0)$ where remainders are controlled using the smoothness of the relevant functions and the fact that under our assumptions the estimators are all root-n consistent.

To describe the results, let

$$\rho_i = \rho(z_i, \beta_0), \qquad \rho_{\beta i} = \partial \rho_i(\beta_0)/\partial \beta, \qquad \eta_i = \rho_{\beta i} - d_i,$$
$$q_i = q^K(x_i), \qquad \kappa_i = E[\rho_i^4 | x_i]/\sigma_i^4,$$
$$\Upsilon = \sum_{i=1}^{n} \sigma_i^2 q_i q_i'/n, \qquad \Gamma = \sum_i q_i d_i'/n, \qquad \tau = \Omega^{*-1} t,$$
$$\xi_{ij} = q_i' \Upsilon^{-1} q_j/n, \qquad E[\tau' \eta_i \rho_i | x_i] = \sigma_i^{\rho\eta},$$
$$\Pi = \sum_{i=1}^{n} \xi_{ii} \sigma_i^{\rho\eta}, \qquad \Pi_B = \sum_{i,j=1}^{n} \sigma_i^{\rho\eta} \sigma_j^{\rho\eta} \xi_{ij}^2,$$
$$\Lambda = \sum_{i=1}^{n} \xi_{ii} E[(\tau' \eta_i)^2 | x_i],$$
$$\Xi = \sum_{i=1}^{n} \xi_{ii} \left(\tau' d_i\right)^2 (5 - \kappa_i), \qquad \Xi_{\mathrm{GEL}} = \sum_{i=1}^{n} \xi_i \left(\tau' d_i\right)^2 (3 - \kappa_i),$$

where we suppress the K argument for notational convenience. The terms involving fourth moments of the residuals are due to estimation of the weight matrix Υ^{-1} for the optimal GMM estimator. This feature did not arise in the homoskedastic case considered in Donald and Newey (2001) where an optimal weight matrix depends only on the instruments.

We impose the following fundamental condition on the data, the approximating functions $q^K(x)$ and the distribution of x:

Assumption 1 (*Moments*). Assume that z_i are i.i.d., and

(i) β_0 is unique value of β in $\mathcal{B}$ (a compact subset of $\mathbb{R}^p$) satisfying $E[\rho(z_i, \beta)|x_i] = 0$;
(ii) $\sum_{i=1}^{n} \sigma_i^{-2} d_i d_i'/n$ is uniformly positive definite and finite (w.p.1.).
(iii) σ_i^2 is bounded and bounded away from zero.
(iv) $E(\eta_{ji}^{\iota_1} \rho_i^{\iota_2} | x_i) = 0$ for any non-negative integers ι_1 and ι_2 such that $\iota_1 + \iota_2 = 3$.
(v) $E(\|\eta_i\|^{\iota} + |\rho_i|^{\iota} | x_i)$ is bounded for $\iota = 6$ for GMM and BGMM and $\iota = 8$ for GEL.

For identification, this condition only requires that $E[\rho(z_i, \beta)|x_i] = 0$ has a unique solution at $\beta = \beta_0$. Estimators will be consistent under this condition because K is allowed to grow with n, as in Donald et al. (2003). Consistency could also be achieved by using the approach of Dominguez and Lobato (2004) or Lavergne and Patilea (2008). Part of this assumption is a restriction that the third moments are zero. This greatly simplifies the MSE calculations. The last condition is a restriction on the moments that are used to control the remainder terms in the MSE expansion. The condition is more restrictive for GEL which has a more complicated expansion involving more terms and higher moments. The next assumption concerns the properties of the derivatives of the moment functions. Specifically, in order to control the remainder terms we will require certain smoothness conditions so that Taylor series expansions can be used and so that we can bound the remainder terms in such expansions.

Assumption 2 (*Expansion*). Assume that $\rho(z, \beta)$ is at least five times continuously differentiable in a neighborhood $\mathcal{N}$ of β_0, with derivatives that are all dominated in absolute value by the random variable b_i with $E(b_i^2) < \infty$ for GMM and BGMM and $E(b_i^5) < \infty$ for GEL.

This assumption is used to control remainder terms and has as an implication that for instance,

$$\sup_{\beta \in \mathcal{N}} \| \left(\partial/\partial \beta'\right) \rho(z, \beta) \| < b_i.$$

It should be noted that in the linear case only the first derivative needs to be bounded since all other derivatives would be zero. It is also interesting to note that although we allow for nonlinearities in the MSE calculations, they do not have an impact on the dominant terms in the MSE. The condition is stronger for GEL reflecting the more complicated remainder term. Our next assumption concerns the "instruments" represented by the vector $q^K(x_i)$.

Assumption 3 (*Approximation*). (i) There is $\zeta(K)$ such that for each K there is a nonsingular constant matrix B such that $\tilde{q}^K(x) = Bp^K(x)$ for all x in the support X of x_i, $\sup_{x \in X} \|\tilde{q}^K(x)\| \le \zeta(K)$, and $E[\tilde{q}^K(x)\tilde{q}^K(x)']$ has smallest eigenvalue that is bounded away from zero, and $\sqrt{K} \le \zeta(K) \le CK$ for some finite constant C. (ii) For each K there exists a sequence of constant vectors π_K and π_K^* such that $E(\|d_i - q_i'\pi_K\|^2) \to 0$ and $\zeta(K)^2 E(\|d_i/\sigma_i^2 - q_i'\pi_K^*\|^2) \to 0$ as $K \to \infty$.

The first part of the assumption gives a bound on the norm of the basis functions, and is used extensively in the MSE derivations to bound remainder terms. The second part of the assumption implies that d_i and d_i/σ_i^2 can be approximated by linear combinations of q_i. Because σ_i^2 is bounded and bounded away from zero, it is easily seen that for the same coefficients π_K, $\|d_i/\sigma_i - \sigma_i q_i \pi_K^*\|^2 \le \sigma_i^2 \|d_i/\sigma_i^2 - q_i'\pi_K\|^2$ so that d_i/σ_i can be approximated by a linear combination of $\sigma_i q_i$. Indeed the variance part of the MSE measures the mean squared error in the fit of this regression. Since $\zeta(K) \to \infty$ the approximation condition for d_i/σ_i^2 is slightly stronger than for d_i. This is to control various remainder terms where d_i/σ_i needs to be approximated in uniform manner. Since in many cases one can show that the expectations in (ii) are bounded by $K^{-2\alpha}$ where α depends on the smoothness of the function d_i/σ_i^2, the condition can be met by assuming that d_i/σ_i^2 is a sufficiently smooth function of x_i.

We will assume that the preliminary estimator $\tilde{\beta}$ used to construct the weight matrix is a GMM estimator is itself a GMM estimator with weighting matrix that may not be optimal. We do not require either optimal weighting or that the number of moments increases, though to get consistency under Assumption 1 (i) K would need to grow. In other words we let $\tilde{\beta}$ solve,

$$\min_{\beta} \tilde{g}_n(\beta)' \tilde{W}_0 \tilde{g}_n(\beta), \qquad \tilde{g}_n(\beta) = (1/n) \sum_{i=1}^{n} \tilde{q}(x_i) \rho_i(\beta)$$

for some $\tilde{K}$ vector of functions $\tilde{q}(x_i)$ and some $\tilde{K} \times \tilde{K}$ matrix $\tilde{W}_0$ which potentially could be $I_{\tilde{K}}$ or it could be random as would be the case if more than one iteration were used to obtain the GMM estimator. We make the following assumption regarding this preliminary estimator.

Assumption 4 (*Preliminary Estimator*). Assume (i) $\tilde{\beta} \xrightarrow{p} \beta_0$ (ii) there exist some non-stochastic matrix W_0 such that $\left\|\tilde{W}_0 - W_0\right\| \xrightarrow{p} 0$ and we can write $\tilde{\beta} = \beta_0 + \frac{1}{n}\sum_{i=1}^{n} \tilde{\phi}_i \rho_i + o_p(n^{-1/2})$, $\tilde{\phi}_i = -(\tilde{\Gamma}' W_0 \tilde{\Gamma})^{-1} \tilde{\Gamma} W_0 \tilde{q}_i$ with $\tilde{\Gamma} = \sum_{i=1}^{n} \tilde{q}(x_i) d_i/n$ and $E\left(\left\|\rho_i^2 \tilde{\phi}_i \tilde{\phi}_i'\right\|\right) < \infty$.

Note that the assumption requires that we just use some root-n consistent and asymptotically normally distributed estimator. The asymptotic variance of the preliminary estimator will be,

and for different specifications of π we generate artificial random samples under the assumptions that

$$E\left(\begin{pmatrix}\rho_i\\ \eta_i\end{pmatrix}\begin{pmatrix}\rho_i & \eta_i\end{pmatrix}\right) = \Sigma = \begin{pmatrix}1 & c\\ c & 1\end{pmatrix}$$

and $X_i \sim N(0, I_{\bar{K}})$ where $\bar{K}$ is the maximal number of instruments considered. As shown in Hahn and Hausman (2002) this specification implies a theoretical first stage R-squared that is of the form,

$$R_f^2 = \frac{\pi'\pi}{\pi'\pi + 1}. \tag{4.4}$$

We consider one of the models that was considered in Donald and Newey (2001) where,

$$\pi_k^2 = c(\bar{K})\left(1 - \frac{k}{\bar{K}+1}\right)^4 \quad \text{for } k = 1, \ldots, \bar{K},$$

where the constant $c(\bar{K})$ is chosen so that $\pi'\pi = R_f^2/(1 - R_f^2)$. In this model all the instruments are relevant but they have coefficients that are declining. This represents a situation where one has prior information that suggests that certain instruments are more important than others and the instruments have been ranked accordingly. In this model all of the potential $\bar{K}$ moment conditions should be used for the estimators to be asymptotically efficient. Note also, that in our setup LIML and 2SLS are also asymptotically efficient estimators provided that we eventually use all of the instruments X_{ji}. Indeed in the experiments we compute not only GMM, BGMM, ET, EL and the CUE (the last three being members of the GEL class) but we also examine the performance of 2SLS and LIML along with the instrument selection methods proposed in Donald and Newey (2001). This allows us to gauge the small sample cost of not imposing heteroskedasticity. As in Donald and Newey (2001) we report for each of the seven different estimators, summary statistics for the version that uses all available instruments or moment conditions plus the summary statistics for the estimators based on a set of moment conditions or instruments that were chosen using the respective moment or instrument selection criterion.

For each model experiments were conducted with the specifications for sample sizes of $n = 200$ and $n = 800$. When the sample size is 200 we set $R_f^2 = 0.1, \bar{K} = 10$ and performed 500 replications, while in the larger sample size we set $R_f^2 = 0.1, \bar{K} = 20$ and we performed 200 replications (due to time constraints). Both of these choices reflect the fairly common situation where there may be a relatively small amount of correlation between the instruments and the endogenous variable (see Staiger and Stock (1997) and Stock and Wright (2000)) as well as the fact that with larger data sets empirical researchers are more willing to use more moment conditions to improve efficiency. For each of these cases we consider $c \in \{.1, .5, .9\}$. In addition we consider the impact of having excess kurtosis, which as noted above has differential effect on the higher order MSE across the different estimators. The distribution we consider is that of

$$\begin{pmatrix}\rho_i\\ \eta_i\end{pmatrix} = |e_i|\begin{pmatrix}\rho_i^*\\ \eta_i^*\end{pmatrix}, \quad \begin{pmatrix}\rho_i^*\\ \eta_i^*\end{pmatrix} \sim N(0, \Sigma), e_i \sim \text{logistic}(0,1)$$

where e_i is independent of ρ_i^* and η_i^* and is distributed as a logistic random variable with mean zero and variance equal to one. Given this particular setup we will have that (ρ_i, η_i) are jointly distributed with mean zero and a covariance matrix equal to Σ, and a coefficient of kurtosis of approximately $\kappa = 12.6$. With two different models, two different distributions for the errors, and three different choices for residual correlations there are a total of 12 specifications for each sample size.

The estimator that uses all moments or instruments is indicated by the suffix "-all" while the estimator that uses a number of moment conditions as chosen by the respective moment or instrument selection criterion is indicated by "-op". We consider a moment selection criteria where $\hat{D}_i^* \tilde{\rho}_i^2 - \tilde{\rho}_{\beta i}$ is replaced by $\tilde{\rho}_{\beta i} - \tilde{d}_i$ in $\hat{\Phi}(K)$, constituting a weak identification approximation. For instance, GMM-all and GMM-op are the two-step estimators that use all of the moment conditions and the moment conditions the minimize the estimated MSE criterion respectively. The preliminary estimates of the objects that appear in the criteria were in each case based on a number of moment conditions that was optimal with respect to cross validation in the first stage.

As in Donald and Newey (2001) we present robust measures of central tendency and dispersion. We computed the median bias (Med. Bias) for each estimator, the median of the absolute deviations (MAD) of the estimator from the true value of $\gamma = 0$ and examined dispersion through the difference between the 0.1 and 0.9 quantile (Dec. Rge) in the distribution of each estimator. We also examined statistical inference by computing the coverage rate for 95% confidence intervals as well as the rejection rate for an overidentification test (in cases where overidentifying restrictions are present) using the test statistic corresponding to the estimator and a significance level of 5%. In addition we report some summary statistics concerning the choices of K in the experiments, including the modal choice of K if one used the actual MSE to choose K. There was very little dispersion in this variable across replications and generally the optimal K with the true criterion was equal to the same value in most if not all replications. In cases where there was some dispersion it was usually either being some cases on either side of the mode. To indicate such cases we use $+$ and $-$, so that for instance $3+$ means that the mode was 3 but that there were some cases where 4 was optimal. The notation $3{+}{+}$ means that the mode was 3 but that a good proportion of the replications had 4 as being optimal.

Tables 1–6 and 10–15 contain the summary statistics for the estimators for $n = 200$ and $n = 800$ respectively, while Tables 7–9 and 16–18 contain the summary statistics for the chosen number of moments across the replications. In general the results are encouraging for all the estimators. As expected the GEL and LIML estimators are less dispersed when the optimal number of moments is used, while for GMM and 2SLS the use of the criterion reduces the bias that occurs when there is a high degree of covariance between the residuals. The improvements for the GEL estimators are more marked when the there is a low to moderate degree of covariance. It is noteworthy that in such situations there is also a dramatic improvement in the quality of inference as indicated by the coverage rates for the confidence interval. As far as testing the overidentifying restrictions only when there is a high degree of covariance is there any problem with testing these restrictions. This occurs with most of the estimators in the small sample with a high covariance and with GMM and TSLS in the large sample with a high covariance. It also seems that using the criteria does not really help in fixing any of the problems with the overidentifying test statistic.

There is a number of things to note about the results for $\hat{K}$. First, the estimated criteria give values for $\hat{K}$ that are often near the values that minimize the true criterion, suggesting that the estimated criterion is a good approximation to the true criterion. It also noteworthy that, as one would expect, the criteria suggest use of a small number of moments for GMM and 2SLS when there is a high error covariance and for the GEL estimators when there is a low covariance. For BGMM the optimal number is quite stable as the covariance increases. In the larger sample the optimal number decreases as the covariance increases, but is slightly larger when the residuals have fat tails compared to the situation where they

Table 1
$n = 200$, Cov $= 0.1$, normal.

Est.	Med. Bias	Med. AD	Dec. Rge	Cov.	Over.
GMM-all	0.028	0.129	0.489	0.934	0.018
GMM-op	0.019	0.143	0.537	0.942	0.022
BGMM-all	0.013	0.163	0.616	0.864	0.012
BGMM-op	0.011	0.152	0.586	0.936	0.036
EL-all	−0.011	0.190	0.712	0.806	0.054
EL-op	0.011	0.158	0.597	0.934	0.048
ET-all	−0.004	0.195	0.716	0.790	0.048
ET-op	0.010	0.155	0.593	0.936	0.042
CUE-all	0.006	0.192	0.733	0.770	0.010
CUE-op	0.013	0.151	0.596	0.924	0.032
2SLS-all	0.027	0.126	0.447	0.958	0.026
2SLS-op	0.018	0.137	0.509	0.974	0.034
LIML-all	−0.009	0.183	0.649	0.974	0.030
LIML-op	0.009	0.141	0.564	0.980	0.026

Table 2
$n = 200$, Cov $= 0.1$, logistic.

Est.	Med. Bias	Med. AD	Dec. Rge	Cov.	Over.
GMM-all	0.018	0.113	0.422	0.932	0.034
GMM-op	0.013	0.125	0.478	0.926	0.044
BGMM-all	0.001	0.135	0.513	0.864	0.032
BGMM-op	0.021	0.137	0.529	0.916	0.040
EL-all	−0.018	0.173	0.646	0.782	0.174
EL-op	0.007	0.149	0.586	0.882	0.104
ET-all	−0.008	0.158	0.601	0.798	0.110
ET-op	0.014	0.148	0.564	0.878	0.088
CUE-all	−0.006	0.160	0.590	0.787	0.024
CUE-op	0.008	0.144	0.562	0.880	0.042
2SLS-all	0.034	0.118	0.443	0.948	0.040
2SLS-op	0.031	0.143	0.516	0.952	0.050
LIML-all	0.001	0.182	0.710	0.972	0.044
LIML-op	0.017	0.152	0.567	0.962	0.052

Table 3
$n = 200$, Cov $= 0.5$, normal.

Est.	Med. Bias	Med. AD	Dec. Rge	Cov.	Over.
GMM-all	0.149	0.165	0.436	0.782	0.038
GMM-op	0.065	0.153	0.530	0.858	0.036
BGMM-all	0.064	0.169	0.598	0.842	0.032
BGMM-op	0.047	0.154	0.532	0.91	0.036
EL-all	−0.002	0.182	0.761	0.854	0.072
EL-op	0.036	0.162	0.552	0.896	0.052
ET-all	0.003	0.180	0.711	0.860	0.066
ET-op	0.035	0.155	0.533	0.898	0.048
CUE-all	0.002	0.177	0.734	0.840	0.022
CUE-op	0.039	0.153	0.528	0.886	0.038
2SLS-all	0.143	0.161	0.426	0.836	0.066
2SLS-op	0.066	0.152	0.517	0.900	0.046
LIML-all	0.006	0.170	0.680	0.964	0.044
LIML-op	0.041	0.154	0.527	0.946	0.048

Table 4
$n = 200$, Cov $= 0.5$, logistic.

Est.	Med. Bias	Med. AD	Dec. Rge	Cov.	Over.
GMM-all	0.131	0.161	0.438	0.768	0.038
GMM-op	0.079	0.154	0.516	0.854	0.044
BGMM-all	0.062	0.160	0.540	0.816	0.032
BGMM-op	0.048	0.148	0.527	0.880	0.038
EL-all	0.016	0.187	0.701	0.796	0.160
EL-op	0.041	0.156	0.578	0.860	0.090
ET-all	0.012	0.178	0.635	0.796	0.108
ET-op	0.039	0.153	0.555	0.868	0.078
CUE-all	−0.004	0.170	0.638	0.776	0.014
CUE-op	0.041	0.154	0.530	0.866	0.036
2SLS-all	0.147	0.172	0.461	0.800	0.076
2SLS-op	0.081	0.160	0.550	0.874	0.058
LIML-all	−0.007	0.175	0.707	0.936	0.06
LIML-op	0.045	0.149	0.581	0.920	0.054

Table 5
$n = 200$, Cov $= 0.9$, normal.

Est.	Med. Bias	Med. AD	Dec. Rge	Cov.	Over.
GMM-all	0.274	0.275	0.368	0.460	0.180
GMM-op	0.124	0.189	0.565	0.798	0.078
BGMM-all	0.128	0.183	0.583	0.738	0.092
BGMM-op	0.091	0.171	0.600	0.814	0.072
EL-all	0.016	0.165	0.688	0.876	0.096
EL-op	0.056	0.168	0.599	0.846	0.126
ET-all	0.020	0.165	0.690	0.874	0.084
ET-op	0.059	0.166	0.603	0.842	0.126
CUE-all	0.024	0.165	0.681	0.880	0.034
CUE-op	0.063	0.169	0.589	0.838	0.078
2SLS-all	0.274	0.275	0.334	0.484	0.198
2SLS-op	0.115	0.186	0.559	0.820	0.062
LIML-all	0.006	0.161	0.648	0.944	0.056
LIML-op	0.041	0.156	0.623	0.900	0.108

Table 6
$n = 200$, Cov $= 0.9$, logistic.

Est.	Med. Bias	Med. AD	Dec. Rge	Cov.	Over.
GMM-all	0.213	0.213	0.349	0.568	0.134
GMM-op	0.092	0.136	0.505	0.874	0.076
BGMM-all	0.065	0.146	0.484	0.802	0.078
BGMM-op	0.073	0.134	0.472	0.854	0.084
EL-all	−0.006	0.146	0.620	0.886	0.158
EL-op	0.044	0.133	0.504	0.870	0.160
ET-all	−0.010	0.134	0.551	0.898	0.118
ET-op	0.039	0.126	0.470	0.892	0.144
CUE-all	−0.016	0.129	0.530	0.879	0.034
CUE-op	0.030	0.122	0.472	0.886	0.066
2SLS-all	0.242	0.244	0.347	0.580	0.190
2SLS-op	0.081	0.134	0.485	0.882	0.076
LIML-all	−0.008	0.131	0.595	0.952	0.056
LIML-op	0.032	0.127	0.557	0.934	0.108

Table 7
Statistics for $\hat{K}$, $n = 200$, cov $= 0.1$.

		GMM	BGMM	EL	ET	CUE	TSLS	LIML
Normal	K	5	3	3	3	3	5	3+
	Mode	2	2	2	2	2	3	2
	1Q	3	2	2	2	2	3	2
	Med.	5	3	3	3	3	4	3
	3Q	8	4	5	5	4	6	4
Logistic	K	5	4+	2+	3	5+	5	3+
	Mode	10	3	2	2	3	3	3
	1Q	4	3	2	3	3	3	2
	Med.	6	4	4	4	4	4	3
	3Q	9	6	6	7	7	6	4

Table 8
Statistics for $\hat{K}$, $n = 200$, cov $= 0.5$.

		GMM	BGMM	EL	ET	CUE	TSLS	LIML
Normal	K	3	3−	4−	4−	4−	3	4
	Mode	2	2	2	2	2	2	3
	1Q	2	2	2	2	2	2	2
	Med.	3	3	3	3	3	3	3
	3Q	4	4	5	5	5	4	5
Logistic	K	3	4−	3−	4−	10−	3	4−
	Mode	3	3	3	3	3	2	2
	1Q	2	2	2	3	3	2	2
	Med.	4	4	4	4	5	3	3
	3Q	6	6	7	7	7	4	5

do not. Among the GEL estimators increasing the covariance and having fat tailed errors has the most dramatic impact on the CUE as one would expect given the criteria.

Concerning the effect of excess kurtosis, it does appear that the improvement from using the criteria is more noticeable for EL, which is most sensitive to having fat-tailed errors. There also was some evidence that going from normal to fat-tailed errors helped

Table 9
Statistics for $\hat{K}$, $n = 200$, cov = 0.9.

		GMM	BGMM	EL	ET	CUE	TSLS	LIML
Normal	K	2	2+	4+	4+	4+	2	5
	Mode	2	2	3	3	3	2	3
	1Q	2	2	3	3	3	2	3
	Med.	2	3	4	4	4	2	4
	3Q	3	3	7	7	6	3	7
Logistic	K	2	3+	3−	4	10	2	5
	Mode	2	2	10	10	10	2	3
	1Q	2	2	3	4	4	2	3
	Med.	2	3	6	6	6	2	4
	3Q	3	5	9	9	9	3	7

Table 10
$n = 800$, cov = 0.1, normal.

Est.	Med. Bias	Med. AD	Dec. Rge	Cov.	Over.
GMM-all	0.019	0.07	0.262	0.925	0.035
GMM-op	0.01	0.078	0.265	0.92	0.055
BGMM-all	0.006	0.081	0.302	0.92	0.025
BGMM-op	0.001	0.084	0.298	0.92	0.06
EL-all	0.000	0.085	0.305	0.895	0.065
EL-op	−0.001	0.080	0.296	0.900	0.060
ET-all	0.005	0.084	0.314	0.895	0.070
ET-op	0.001	0.080	0.290	0.905	0.065
CUE-all	0.006	0.082	0.307	0.895	0.025
CUE-op	0.004	0.082	0.298	0.915	0.055
2SLS-all	0.022	0.066	0.24	0.925	0.050
2SLS-op	0.005	0.074	0.275	0.920	0.060
LIML-all	−0.001	0.079	0.305	0.945	0.060
LIML-op	−0.006	0.083	0.296	0.930	0.055

Table 11
$n = 800$, cov = 0.1, logistic.

Est.	Med. Bias	Med. AD	Dec. Rge	Cov.	Over.
GMM-all	0.007	0.068	0.234	0.93	0.025
GMM-op	0.007	0.073	0.253	0.925	0.025
BGMM-all	−0.007	0.076	0.271	0.89	0.025
BGMM-op	−0.002	0.07	0.269	0.905	0.025
EL-all	−0.012	0.082	0.300	0.850	0.135
EL-op	−0.002	0.082	0.288	0.910	0.085
ET-all	−0.015	0.083	0.286	0.845	0.105
ET-op	−0.003	0.073	0.290	0.900	0.080
CUE-all	−0.005	0.08	0.281	0.856	0.025
CUE-op	−0.001	0.073	0.276	0.887	0.035
2SLS-all	0.005	0.067	0.243	0.965	0.060
2SLS-op	0.007	0.072	0.260	0.975	0.025
LIML-all	−0.012	0.078	0.314	0.975	0.060
LIML-op	−0.010	0.069	0.297	0.98	0.045

Table 12
$n = 800$, cov = 0.5, normal.

Est.	Med. Bias	Med. AD	Dec. Rge	Cov.	Over.
GMM-all	0.087	0.094	0.237	0.770	0.070
GMM-op	0.034	0.081	0.269	0.910	0.035
BGMM-all	0.022	0.085	0.297	0.860	0.065
BGMM-op	0.016	0.077	0.278	0.940	0.020
EL-all	0.004	0.089	0.322	0.890	0.075
EL-op	0.015	0.084	0.282	0.930	0.065
ET-all	0.005	0.089	0.314	0.880	0.075
ET-op	0.015	0.085	0.282	0.935	0.065
CUE-all	0.009	0.090	0.322	0.870	0.050
CUE-op	0.015	0.082	0.276	0.935	0.040
2SLS-all	0.089	0.090	0.231	0.805	0.065
2SLS-op	0.035	0.077	0.255	0.915	0.035
LIML-all	0.004	0.085	0.319	0.960	0.055
LIML-op	0.018	0.083	0.281	0.955	0.040

the CUE more than the other estimators, as suggested in the theory, although this led to a lower improvement from using the moment selection criterion.

Table 13
$n = 800$, cov = 0.5, logistic.

Est.	Med. Bias	Med. AD	Dec. Rge	Cov.	Over.
GMM-all	0.082	0.091	0.241	0.790	0.05
GMM-op	0.042	0.078	0.274	0.870	0.045
BGMM-all	0.022	0.079	0.291	0.870	0.025
BGMM-op	0.025	0.081	0.285	0.895	0.055
EL-all	−0.003	0.089	0.312	0.875	0.155
EL-op	0.016	0.083	0.287	0.885	0.100
ET-all	0.000	0.082	0.303	0.870	0.115
ET-op	0.018	0.080	0.275	0.885	0.080
CUE-all	0.001	0.086	0.302	0.838	0.025
CUE-op	0.017	0.077	0.273	0.880	0.045
2SLS-all	0.093	0.093	0.224	0.800	0.055
2SLS-op	0.041	0.076	0.278	0.890	0.060
LIML-all	−0.009	0.076	0.286	0.955	0.055
LIML-op	0.021	0.078	0.274	0.925	0.06

Table 14
$n = 800$, cov = 0.9, normal.

Est.	Med. Bias	Med. AD	Dec. Rge	Cov.	Over.
GMM-all	0.176	0.176	0.212	0.415	0.145
GMM-op	0.064	0.093	0.273	0.880	0.055
BGMM-all	0.060	0.100	0.297	0.815	0.070
BGMM-op	0.044	0.083	0.294	0.875	0.060
EL-all	0.010	0.078	0.287	0.915	0.085
EL-op	0.032	0.082	0.272	0.895	0.105
ET-all	0.016	0.078	0.279	0.920	0.115
ET-op	0.025	0.079	0.273	0.920	0.135
CUE-all	0.012	0.080	0.280	0.925	0.075
CUE-op	0.025	0.079	0.276	0.920	0.115
2SLS-all	0.166	0.166	0.217	0.455	0.140
2SLS-op	0.061	0.089	0.268	0.88	0.050
LIML-all	0.020	0.079	0.285	0.95	0.060
LIML-op	0.035	0.080	0.276	0.93	0.100

Table 15
$n = 800$, cov = 0.9, logistic.

Est.	Med. Bias	Med. AD	Dec. Rge	Cov.	Over.
GMM-all	0.143	0.145	0.179	0.530	0.130
GMM-op	0.046	0.089	0.274	0.885	0.045
BGMM-all	0.033	0.070	0.230	0.885	0.075
BGMM-op	0.039	0.074	0.263	0.875	0.075
EL-all	0.003	0.066	0.289	0.910	0.180
EL-op	0.024	0.072	0.256	0.920	0.165
ET-all	−0.002	0.086	0.281	0.910	0.115
ET-op	0.015	0.079	0.281	0.910	0.125
CUE-all	−0.001	0.083	0.264	0.919	0.035
CUE-op	0.013	0.079	0.277	0.911	0.040
2SLS-all	0.161	0.161	0.199	0.510	0.135
2SLS-op	0.058	0.089	0.304	0.870	0.085
LIML-all	0.004	0.077	0.304	0.975	0.075
LIML-op	0.016	0.076	0.263	0.955	0.115

Table 16
Statistics for $\hat{K}$, $n = 800$, cov = 0.1.

		GMM	BGMM	EL	ET	CUE	TSLS	LIML
Normal	K	10	7	7+	7+	7+	10	8−
	Mode	8	6	6	6	6	8	6
	1Q	8	6	6	6	6	7	6
	Med.	12	7	7	7	7	9	7
	3Q	17	9	9	9	9	13	9
Logistic	K	10	9+	6	7	11−	10	8
	Mode	20	10	7	8	8	8	7
	1Q	10	8	6	7	8	7	6
	Med.	15	11	8	10	12	9	7
	3Q	19	16	12	15	17	12	8

5. Conclusion

In this paper we have developed approximate MSE criteria for moment selection for a variety of estimators in conditional moment contexts. We found that the CUE has smaller MSE than

Table 17

Statistics for $\hat{K}$, $n = 800$, $\text{cov} = 0.5$.

		GMM	BGMM	EL	ET	CUE	TSLS	LIML
Normal	*K*	6−	7−	8	8	8	6−	9−
	Mode	5	6	7	7	7	5	7
	1Q	5	6	6	6	6	5	6
	Med.	6	7	8	8	8	6	8
	3Q	7	8	10	10	9	7	10
Logistic	*K*	6−	8	7−	8	20	6−	9−
	Mode	6	6	6	6	20	6	6
	1Q	5	7	6	7	9	5	6
	Med.	6	10	9	11	13	6	8
	3Q	8	15	14	15	18	7	10

Table 18

Statistics for $\hat{K}$, $n = 800$, $\text{cov} = 0.9$.

		GMM	BGMM	EL	ET	CUE	TSLS	LIML
Normal	*K*	4	6+	9	9	9	4	11
	Mode	4	5	8	9	9	4	8
	1Q	4	5	8	8	7	4	8
	Med.	4	6	10	10	10	4	10
	3Q	5	7	15	15	14	5	14
Logistic	*K*	4	7	7	9	20	4	11
	Mode	4	7	20	20	20	4	8
	1Q	4	6	8	10	13	3	7
	Med.	4	8	12	15	17	4	9
	3Q	5	12	18	19	19	4	14

the bias-corrected GMM estimator. In addition we proposed data based methods for estimating the approximate MSE, so that in practice the number of moments can be selected by minimizing these criteria. The criteria seemed to perform adequately in a small scale simulation exercise.

The present paper has considered a restrictive environment in which the data are considered a random sample. It would be useful to extend the results in two directions. The first would be to the dynamic panel data case. In that situation there will typically be different sets of instruments available for each residual coming from sequential moment restrictions. It would also be useful to extend the results to a purely time series context where one would need to deal with serial correlation. Kuersteiner (2002) has derived interesting results in this direction.

Acknowledgements

This work reported in this paper was supported in part by the National Science Foundation. Miguel Delgado and a referee provided useful comments.

References

Andrews, D.W.K., 1999. Consistent moment selection procedures for generalized method of moments estimation. Econometrica 67, 543–564.

Chamberlain, G., 1987. Asymptotic efficiency in estimation with conditional moment restrictions. Journal of Econometrics 34, 305–334.

Dominguez, M.A., Lobato, I.N., 2004. Consistent estimation of models defined by conditional moment restrictions. Econometrica 72, 1601–1615.

Donald, S.G., Newey, W.K., 2001. Choosing the number of instruments. Econometrica 69, 1161–1191.

Donald, S.G., Imbens, G., Newey, W.K., 2003. Empirical likelihood estimation and consistent tests with conditional moment restrictions. Journal of Econometrics 117, 55–93.

Hahn, J., Hausman, J.A., 2002. A new specification test for the validity of instrumental variables. Econometrica 70, 163–189.

Hansen, L.P., Heaton, J., Yaron, A., 1996. Finite-sample properties of some alternative GMM estimators. Journal of Business and Economic Statistics 14, 262–280.

Imbens, G.W., Spady, R.H., Johnson, P., 1998. Information theoretic approaches to inference in moment condition models. Econometrica 66, 333–357.

Imbens, G.W., Spady, R.H., 2005. The performance of empirical likelihood and its generalizations. In: Andrews, D., Stock, J. (Eds.), Identification and Inference for Econometric Models, Essays in Honor of Thomas Rothenberg. Cambridge University Press, Cambridge.

Kitamura, Y., Stutzer, M., 1997. An information-theoretic alternative to generalized method of moments estimation. Econometrica 65, 861–874.

Koenker, R., Machado, J.A.F., Skeels, C., Welsh, A.H., 1994. Momentary lapses: Moment expansions and the robustness of minimum distance estimation. Econometric Theory 10, 172–190.

Kuersteiner, G.M., 2002. Selecting the number of instruments for GMM estimators of linear time series models. Mimeo UC Davis.

Lavergne, P., Patilea, V., 2008. Bandwidth-robust inference with conditional moment restrictions. Preprint.

Nagar, A.L., 1959. The bias and moment matrix of the general *k*-class estimators of the parameters in simultaneous equations. Econometrica 27, 575–595.

Newey, W.K., Smith, R.J., 2004. Higher-order properties of GMM and generalized empirical likelihood estimators. Econometrica 72, 219–255.

Newey, W.K., Windmeijer, F., 2009. GMM estimation with many weak moment conditions. Econometrica 77, 687–719.

Owen, A., 1988. Empirical likelihood ratio confidence regions for a single functional. Biometrika 75, 237–249.

Qin, J., Lawless, J., 1994. Empirical likelihood and general estimating equations. Annals of Statistics 22, 300–325.

Rothenberg, T.J., 1983. Asymptotic properties of some estimators in structural models. In: Karlin, S., Amemiya, T., Goodman, L.A. (Eds.), Studies in Econometrics, Time Series and Multivariate Statistics. Academic Press, New York.

Rothenberg, T.J., 1984. Approximating the distributions of econometric estimators and test statistics. In: Griliches, Z., Intriligator, M.D. (Eds.), Handbook of Econometrics, vol. 2. North-Holland, New York.

Rothenberg, T.J., 1996. Empirical likelihood parameter estimation under moment restrictions. In: Seminar Notes. Harvard/M.I.T., Bristol.

Staiger, D., Stock, J.H., 1997. Instrumental variables regression with weak instruments. Econometrica 65, 557–586.

Stock, J.H., Wright, J.H., 2000. GMM with weak identification. Econometrica 68, 1055–1096.

Sun, Y., Phillips, P.C.B., Jin, S., 2007. Optimal bandwidth selection in heteroscedasicity-autocorrelation robust testing. Econometrica 76, 175–194.