Inference for Iterated GMM Under Misspecification

Bruce E. Hansen*
University of Wisconsin

Seojeong Lee†
University of New South Wales

November 2020‡

Abstract

This paper develops inference methods for the iterated over-identified Generalized Method of Moments (GMM) estimator. We provide conditions for the existence of the iterated estimator and an asymptotic distribution theory which allows for mild misspecification. Moment misspecification causes bias in conventional GMM variance estimators which can lead to severely over-sized hypothesis tests. We show how to consistently estimate the correct asymptotic variance matrix. Our simulation results show that our methods are properly sized under both correct specification and mild to moderate misspecification. We illustrate the method with an application to the model of Acemoglu, Johnson, Robinson, and Yared (2008).

*Hansen thanks the National Science Foundation and the Phipps Chair for research support.

†Lee acknowledges that this research was supported under the Australian Research Council Discovery Early Career Research Award (DECRA) funding scheme (project number DE170100787).

‡We thank the Co-Editor Ulrich Müller and three referees for helpful comments on previous versions.

1 Introduction

White (1980ab, 1982) advocated for robust inference, meaning that variance estimation should be valid under broader assumptions than the model interpreted narrowly. His seminal papers showed how to construct robust covariance estimators for linear regression and for likelihood estimation which provide asymptotically valid inference for pseudo-true parameters without correct model specification. White's vision for robust covariance estimation dominates much of econometric practice. The metaphor of robust estimation also motivated the generalized method of moments (GMM) estimator of Lars Hansen (1982), as it was understood that maximum likelihood estimation can be sensitive to model misspecification. This concern for robustness is echoed in the monograph by Hansen and Sargent (2008), where they argue that decisions should be robust to mild model misspecification.

This paper provides a rigorous distribution theory for iterated GMM in over-identified econometric models under mild misspecification. We focus on the iterated estimator as it removes the arbitrary dependence of the one-step and two-step GMM estimators on the initial weight matrix. By "misspecification" we mean that some moment conditions may fail in the population. This is appropriate when the model is viewed as an approximation rather than as literally true. By "mild misspecification" we mean that the degree of misspecification is bounded. This is similar to the concept of bounded robustness proposed in Hansen and Sargent (2008). It is different from "local misspecification" which treats the degree of misspecification as diminishing in sample size. A reasonable interpretation is that "mild misspecification" is intermediate between "local misspecification" and "global misspecification".

Our contributions include the first formal demonstration of existence of the iterated estimator by establishing that the iteration sequence is a contraction, an asymptotic distribution theory for the iterated estimator allowing for misspecification, consistent covariance matrix estimation allowing for misspecification, and simulation evidence documenting the large distributional distortions in conventional test statistics due to misspecification and how our proposed standard errors dramatically reduce these distortions.

Moment misspecification alters the GMM asymptotic covariance matrix, as was first pointed out by Hall and Inoue (2003). Under misspecification the asymptotic covariance matrix contains terms which depend on estimation error in the moment derivatives, weight matrix estimation, the degree of curvature of the model moments, and the sensitivity of the weight matrix to the parameter values. Ignoring these terms leads to substantial size distortions. Hall and Inoue (2003) developed a distributional theory for the one-step and two-step GMM estimator under misspecification; this paper extends this theory to the iterated GMM estimator. An interesting by-product of our analysis is that the asymptotic distribution of the iterated estimator is simpler than the two-step estimator.

Our misspecification-robust covariance matrix estimator is closely related to the finite sample correction of Windmeijer (2000, 2005) which is routinely used in practice. The Windmeijer formula corrects for the bias in the standard error of the linear two-step and iterated GMM estimators by considering the extra variation arising from the efficient weight matrix being evaluated at an

estimate rather than the true value. Hwang, Kang, and Lee (2020) show that the misspecification-robust standard errors provide the same order of finite sample correction with the Windmeijer standard error under correct specification. Since the Windmeijer formula is not robust to misspecification, these calculations show that our misspecification-robust standard errors are more accurate than both conventional and Windmeijer standard errors.

The assumptions in this paper are closely related to over-identified instrumental variable (IV) regression with heterogeneous treatment effects (Imbens and Angrist (1994), Angrist and Imbens (1995), Kolesár (2013)). As shown in Lee (2018) and Evdokimov and Kolesár (2018), conventional inference methods are inappropriate in this context and alternative standard error formulas are necessary. The theory and methods presented in this paper include the heterogeneous treatment effect IV model as a special case and apply more broadly to linear and non-linear GMM estimation. We also build on the moment misspecification literature of White (1982), Maasoumi and Phillips (1982), Gallant and White (1988), Hall and Inoue (2003), Aguirre-Torres and Toribio (2004), Schennach (2007), Dovonon (2016), and Lee (2014, 2016). A relevant connection is Ai and Chen (2007), who derive the asymptotic distribution of the sieve minimum distance estimator for misspecified conditional moment restrictions models.

We focus on models which are potentially misspecified. These models have no true parameter, so the "true" parameter must be defined as a pseudo-true value – the value which minimizes the population version of the sample criterion. This follows a long tradition in econometrics. Goldberger (1991) proposed interpreting the least squares regression coefficient vector as the best linear predictor. White (1980a, 1982, 1984, 1994) recommended interpreting parameter values as minimizers of population criterion. Angrist, Chernozhukov and Fernández-Val (2006) derived the pseudo-true parameter value for the quantile regression estimator. In the asset pricing literature, Kan and Robotti (2008) and Gospodinov, Kan, and Robotti (2014) provide misspecification-robust inference methods for the GMM pseudo-true value minimizing the distance metric of Hansen and Jagannathan (1997). In the heterogeneous treatment effects literature, Angrist and Imbens (1995) show that the two-stage least squares (2SLS) estimand is the average causal response (ACR), which is a weighted average of the local average treatment effects (LATEs). As shown by Lee (2018), the moment condition model underlying 2SLS is misspecified if the instruments identify the instrument-specific LATEs. Thus, the ACR is a well-accepted GMM pseudo-true value.

More generally, our approach is related to recent work on characterizing various policy relevant treatment effects as weighted averages of the marginal treatment effect, e.g. Heckman and Vytlacil (2005), Heckman, Urzua, and Vytlacil (2006), Brinch, Mogstad, and Wiswall (2017), Mogstad, Santos, and Torgovitsky (2018), I. Andrews (2019), Evdokimov and Kolesár (2018), and Słoczyński (2018). Since these examples cannot be handled by a local misspecification framework (e.g. Cheng, Liao, and Shi, 2019) our approach is complementary to local misspecification.

The iterated GMM estimator is related to – but substantially different from – the continuously updated estimator (CU-GMM) of Hansen, Heaton and Yaron (1996). The CU-GMM is a one-step estimator (no iteration is required). Its asymptotic distribution under misspecification could

be derived by methods similar to those employed in this paper but we do not do so to keep the presentation focused.

There are a number of limitations to our analysis. First, our results assume that the moment conditions are smooth. Allowing for non-differentiable moment conditions would be desirable but would require a different technical approach. Second, it is difficult to give economic interpretation to pseudo-true parameter values. Consequently this limits interest in valid inference procedures for pseudo-true values. Third, the smoothness and moment assumptions needed for misspecification-robust inference are stronger than those needed for conventional inference methods. This is analogous to the fact that White's (1980a) heteroskedasticity-robust covariance matrix estimator requires stronger assumptions than the classical variance estimator. Fourth, we do not allow for weak identification, and this extension would be desirable but considerably more challenging. Fifth, we do not consider how to select among point estimators in the context of potential misspecification. This is a particularly difficult yet important topic. The methods of Cheng, Liao and Shi (2019) may be useful in this regard.

A common application of over-identified GMM is in dynamic panel models. Dynamic panel regression is highly susceptible to misspecification, as it is not credible that the dynamic structure (number of lags) is known *a priori*. Consequently dynamic panel models should generically be viewed as constructive approximations. We illustrate our methods by replicating and extending the results of Acemoglu, Johnson, Robinson, and Yared (2008) and Cervellati, Jung, Sunde, and Visher (2014). We show that the GMM estimates are highly sensitive to the number of GMM iterations. We also show that (depending on the specification) the standard errors can change enormously if the misspecification-robust variance estimator is used instead of the standard Arellano-Bond estimator. These results are consistent with mild misspecification and demonstrate the importance of using robust methods for empirical research.

A Matlab code which replicates the empirical work reported in the paper is available on the authors' webpages.

Regarding notation, let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote its smallest and largest eigenvalue of a positive semi-definite matrix A. For a vector a let $\|a\| = (a'a)^{1/2}$ denote the Euclidean norm. For a matrix A let $\|A\| = \sqrt{\lambda_{\max}(A'A)}$ denote the spectral norm.

2 Generalized Method of Moments Estimation

Consider a standard over-identified moment condition model. The moment equation is

$$E[m(X_i, \theta_0)] = 0 \tag{1}$$

where $m(x, \theta)$ is $l \times 1$ and $\theta \in \Theta \subset \mathbb{R}^k$ with $l > k$. Given a sample $\{X_1, ..., X_n\}$ let

$$\overline{m}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} m(X_i, \theta)$$

be the sample estimator of (1) given θ. Let

$$\overline{W}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} v(X_i,\theta)v(X_i,\theta)'$$

be a weight matrix estimator where $v(x,\theta)$ is $l \times 1$. Two leading examples are 2SLS which sets $v(x,\theta) = z$ for an instrument vector Z_i, and the efficient weight matrix which sets $v(x,\theta) = m(x,\theta)$.

The GMM criterion function is

$$\overline{J}_n(\theta,\phi) = \overline{m}_n(\theta)'\overline{W}_n(\phi)^{-1}\overline{m}_n(\theta). \tag{2}$$

The parameter ϕ is the initial value used to form the weight matrix. The GMM estimator is the value which minimizes the criterion function. Define the mapping

$$\overline{g}_n(\phi) = \arg\min_{\theta\in\Theta} \overline{J}_n(\theta,\phi). \tag{3}$$

This is the minimizer of the criterion function over θ, given the initial value ϕ used to form the weight matrix. Let $\widehat{\theta}_0$ be an initial value. The one-step estimator is $\widehat{\theta}_1 = \overline{g}_n(\widehat{\theta}_0)$, the two-step estimator is $\widehat{\theta}_2 = \overline{g}_n(\widehat{\theta}_1)$, and the s-step estimator is

$$\widehat{\theta}_s = \overline{g}_n(\widehat{\theta}_{s-1}). \tag{4}$$

The *iterated GMM estimator* is the limit of this sequence

$$\widehat{\theta} = \lim_{s\to\infty} \widehat{\theta}_s. \tag{5}$$

We discuss in Section 4 sufficient conditions such that this limit exists.

The limit (5) is a fixed point of the equation

$$\overline{g}_n(\widehat{\theta}) = \widehat{\theta}. \tag{6}$$

An interesting feature (previously unnoticed) is that the iterated estimator $\widehat{\theta}$ is identical if the centered version of the efficient weight matrix is used. That is, set $v(x,\theta) = m(x,\theta)$ and

$$\begin{aligned}
\overline{W}_n^*(\theta) &= \frac{1}{n}\sum_{i=1}^{n} m(X_i,\theta)m(X_i,\theta)' - \overline{m}_n(\theta)\overline{m}_n(\theta)' \\
\overline{J}_n^*(\theta,\phi) &= \overline{m}_n(\theta)'\overline{W}_n^*(\phi)^{-1}\overline{m}_n(\theta) \\
\overline{g}_n^*(\phi) &= \arg\min_{\theta\in\Theta} \overline{J}_n^*(\theta,\phi) \\
\widehat{\theta}_s &= \overline{g}_n^*(\widehat{\theta}_{s-1}).
\end{aligned}$$

Let $\widehat{\theta}^*$ be the limit of the sequence $\widehat{\theta}_s$.

Theorem 1. *Set $v(x,\theta) = m(x,\theta)$ and assume that $\overline{W}_n(\theta)$ and $\overline{W}_n^*(\theta)$ are non-singular. Then $\widehat{\theta}^* = \widehat{\theta}$.*

This result has antecedents. Newey and Smith (2004, footnote 2) asserted that weight matrix recentering does not affect the CU-GMM estimator, but did not provide a proof. Based on the first-order conditions for the s-step GMM estimator, Hall (2005, p. 129) asserted that recentering may not affect the probability limit of the iterated GMM estimator, but also did not provide a proof. Theorem 1 shows that this equivalence is finite sample exact for the iterated estimator. By similar reasoning, this result extends as well to pseudo-true parameter values.

3 Existence of Pseudo-True Parameter

In this section we define and demonstrate existence of the *pseudo-true* parameter θ_0 when the over-identified moment equation (1) is not necessarily satisfied. Following White (1982) we define θ_0 as the solution to the population analog of the estimator.

Define the population analogs of the sample moment and weight matrix

$$m(\theta) = \frac{1}{n}\sum_{i=1}^{n} E\left[m(X_i,\theta))\right] \tag{7}$$

$$W(\theta) = \frac{1}{n}\sum_{i=1}^{n} E\left[v(X_i,\theta)v(X_i,\theta)'\right]. \tag{8}$$

Under heterogeneous distributions the expectations (7) and (8) may vary with n. To not overburden the notation we do not index these and similar expressions by n. Define the population analogs of (2) and (3):

$$\begin{aligned} J(\theta,\phi) &= m(\theta)'W(\phi)^{-1}m(\theta) \\ g(\phi) &= \underset{\theta\in\Theta}{\arg\min}\, J(\theta,\phi). \end{aligned} \tag{9}$$

Definition (9) specifies $g(\phi)$ as the best fitting value of θ given the weight matrix $W(\phi)$ and an initial value ϕ. Under correct specification so that (1) holds and $W(\phi) > 0$, then the solution $g(\phi) = \theta_0$ is unique. Under moment misspecification, however, the solution (9) may vary with ϕ.

As an analog of the iterated GMM estimator (6) we define θ_0 to be the fixed point which solves

$$g(\theta_0) = \theta_0. \tag{10}$$

Conceptually, one could imagine obtaining θ_0 by iterating $g(\phi)$ until convergence.

The existence of the fixed point (10) has not been discussed in the previous literature. We now provide a formal justification. Define $Q(\theta) = \frac{\partial}{\partial\theta'}m(\theta)$ and $S(\theta) = \frac{\partial}{\partial\theta'}\operatorname{vec} W(\theta)$. $S(\theta)$ is a measure of the *sensitivity* of the weight matrix to the parameter value.

Assumption 1.

1. Θ *is compact*

2. *For all* $\phi \in \Theta$, $g(\phi)$ *is unique and in the interior of* Θ

3. $m(\theta)$ *is twice continuously differentiable in* $\theta \in \Theta$

4. $W(\theta)$ *is continuously differentiable in* $\theta \in \Theta$

5. $\inf_{\phi \in \Theta} \lambda_{\min}\left(W(\phi)\right) \geq C > 0$

6. $\sup_{\phi \in \Theta} J(g(\phi), \phi) < \frac{C_3^2}{4C_1C_2}$ *for*

$$C_1 = \sup_{\phi \in \Theta} \left\| Q(g(\phi))' W(\phi)^{-1} Q(g(\phi)) \right\| \tag{11}$$

$$C_2 = \sup_{\phi \in \Theta} S(g(\phi))' \left(W(\phi)^{-1} \otimes W(\phi)^{-1} \right) S(g(\phi)) \tag{12}$$

$$C_3 = \inf_{\phi \in \Theta} \left\| \frac{\partial^2}{\partial \theta \partial \theta'} J(\theta, \phi)|_{\theta = g(\phi)} \right\|. \tag{13}$$

Assumption 1.1 imposes compactness for technical convenience. Assumption 1.2 states that for any population weight matrix there is a unique population minimizer. Assumptions 1.3 and 1.4 are smoothness conditions on the population moments and weight function. These conditions are stronger than needed for the correctly specified case. Assumption 1.5 excludes singular population weight matrices.

Assumption 1.6 is unusual. It controls the degree of misspecification. It allows for misspecification since it allows $J(\theta_0, \phi) > 0$. It only allows, however, for mild misspecification since the magnitude of $J(g(\phi), \phi)$ is bounded. Assumption 1.6 is automatically satisfied under correct specification (since in that context $m(\theta_0) = 0$) but otherwise allows for mild deviations from the assumed model in the sense that for $\phi_1 \neq \phi_2$, $g(\phi_1) \neq g(\phi_2)$ but the difference is bounded. Mild misspecification allows global misspecification ($\|m(\theta)\| > 0$ for all $\theta \in \Theta$) but differs from local misspecification ($m(\theta_0) = \eta/n^s$ for some $\eta \neq 0$ and $s > 0$). Our definition of mild misspecification is similar to Hansen and Sargent (2008): the set of models around the benchmark model that lie within a fixed sized entropy ball. Our stated upper bound $C_3^2/4C_1C_2$ is technical, and depends on the norm of population moments. The upper bound $C_3^2/4C_1C_2$ is determined by the constants C_1, C_2, and C_3 defined in (11), (12) and (13). C_1 and C_2 are finite under Assumptions 1.1-5. $C_3 > 0$ typically holds under Assumption 1.2. The allowable range for the left-side of Assumption 1.6 is larger when C_3 is larger (which occurs when $g(\phi)$ is well identified) and/or C_2 is smaller (less sensitivity of the weight matrix). Assumption 1.6 is not a sharp bound. Below we provide examples where we explicitly calculate the contraction mapping. In two of these examples (1 & 2) we show that the contraction mapping holds for all parameter values so Assumption 1.6 is unnecessary. Furthermore, Assumption 1.6 is only used to establish the existence of the fixed point under misspecification, and could be replaced by any other sufficient condition for existence of the fixed point.

Assumption 1 is sufficient to establish the existence of the pseudo-true value θ_0.

Theorem 2. *Under Assumption 1,*

$$\sup_{\phi\in\Theta}\left\|\frac{\partial}{\partial\phi'}g(\phi)\right\| < 1. \tag{14}$$

The map $g(\phi)$ is a contraction and the fixed point θ_0 exists and is unique.

We now provide three examples where we can explicitly calculate the contraction mapping.

Example 1: Location model
Consider a simple location model with i.i.d. observations $X_i = (Y_i, Z_i)$ where Y and Z have the same unknown mean. The moment function is $m(X_i, \theta) = (Y_i - \theta, Z_i - \theta)'$. Assume that the true data generating process is a bivariate normal with $E[Y_i] = 0$, $E[Z_i] = \alpha$, $Var[Y_i] = Var[Z_i] = 1$, and $Cov(Y_i, Z_i) = 0.5$. We calculate that the GMM mapping with the weight matrix $W(\phi) = E[m(X_i, \phi)m(X_i, \phi)']$ is

$$g(\phi) = \frac{\alpha}{2(1+\alpha^2)} + \frac{\alpha^2}{1+\alpha^2}\phi.$$

Since $(\partial/\partial\phi)g(\phi) = \alpha^2/(1+\alpha^2) < 1$ for all real α, $g(\phi)$ is a contraction irrespective of the degree of misspecification. The minimized population criterion is $J(g(\phi), \phi) = \alpha^2/(1+\alpha^2)$ which does not depend on ϕ. It deviates from zero when $\alpha \neq 0$ but is bounded in α. The fixed point is $\theta = \alpha/2$ which can be obtained by solving the population first-order condition. Alternatively, the fixed point can be obtained without iteration by using the centered efficient weight matrix $W^* = W(\phi) - E[m(X_i, \phi)]\,E[m(X_i, \phi)]'$ because W^* does not depend on ϕ.[1] With W^* the minimized criterion is $J^*(g^*(\phi), \phi) = \alpha^2$ which is unbounded in α. In this example Assumption 1.6 is unnecessary.

Example 2: Linear IV
Consider a simple linear instrumental variable regression. The model is $Y_i = X_i\theta_0 + \varepsilon_i$, $E[Z_i\varepsilon_i] = 0$ where X_i and θ_0 are scalar and $Z_i = (Z_{1i}, Z_{2i})'$ is a vector of two instruments. The moment function is $m(W_i, \theta) = (Z_{1i}(Y_i - X_i\theta), Z_{2i}(Y_i - X_i\theta))'$ where $W_i = (Y_i, X_i, Z_{1i}, Z_{2i})$. Assuming that the data-generating process is $Y_i = X_i + \alpha(Z_{1i} - Z_{2i}) + e_i$, $X_i = (Z_{1i} + Z_{2i}) + u_i$, $Z_i \sim_{i.i.d} N(0, I_2)$, and $(e_i, u_i)' \sim_{i.i.d} N(0, [1, 0.5; 0.5, 1])$, the GMM mapping with the weight matrix $W(\phi) = E[m(W_i, \phi)m(W_i, \phi)']$ is

$$g(\phi) = \frac{5 - 7\phi + 3\phi^2 + \alpha^2(2 + 4\phi)}{5 - 7\phi + 3\phi^2 + 6\alpha^2}.$$

Since $(\partial/\partial\phi)g(\phi) < 2/3$ for all real α the mapping is a contraction irrespective of the degree of misspecification. The minimized population criterion is bounded because $J(g(\phi), \phi) = (3\alpha^2 + \frac{3}{2}\left(\phi - \frac{7}{6}\right)^2 + \frac{11}{24})^{-1}\alpha^2 < \frac{1}{3}$. The fixed point is $\theta = \theta_0$ which is invariant to the degree of misspecification. In this example Assumption 1.6 is unnecessary.

In general IV models a sufficient condition for the GMM estimator to be a contraction can be derived. Consider the model $Y_i = X_i\theta_0 + Z_i'\alpha + \varepsilon_i$, $X_i = Z_i'\pi + u_i$ where Y_i and X_i are scalars and Z_i

[1] This is true for any moment function in the form of $m(X_i, \theta) = \pi(X_i) + h(\theta)$, e.g., efficient minimum distance estimators.

is an $l \times 1$ vector. Assume that (Y_i, X_i, Z_i) are i.i.d. A nonzero α means a violation of the exclusion restriction. Further assume that $E[Z_i u_i] = 0$, $E[Z_i \varepsilon_i] = 0$, and $\pi \neq 0$ and let $\Sigma = E[Z_i Z_i']$. The GMM mapping with a weight matrix $W(\phi)$ is

$$g(\phi) = \theta_0 + (\pi' \Sigma W(\phi)^{-1} \Sigma \pi)^{-1} \pi' \Sigma W(\phi)^{-1} \Sigma \alpha.$$

Since the exact expression for $W(\phi)^{-1}$ in terms of model parameters is hard to obtain so is the exact condition for $|(\partial/\partial\phi) g(\phi)| < 1$. However, by manipulating the matrices it can be shown that $|(\partial/\partial\phi) g(\phi)| = 0$ if $\pi\alpha' = \alpha\pi'$. This condition holds if the instruments are all valid ($\alpha = 0$) or $\pi = \alpha$. The latter case may arise in practice where the strong instruments are likely to be invalid while the weaker instruments are likely to be valid. For example, lagged dependent variables are often used as instruments in the dynamic panel models. If there is a serial correlation in the errors, the instrument becomes weaker but less invalid as the lag increases.

Example 3: Nonlinear model of Schennach (2007)
Consider a simple nonlinear model where the mean is estimated imposing a known variance. The moment function is given by

$$m(X_i, \theta) = \begin{pmatrix} X_i - \theta \\ (X_i - \theta)^2 - 1 \end{pmatrix}. \tag{15}$$

The parameter of interest is θ and a unit variance is imposed. The moment condition is misspecified if the actual variance differs from one. The data-generating process is $X_i \sim_{i.i.d} N(\theta_0, \sigma^2)$ where $\sigma^2 > 0$. We set $\theta_0 = 0$. The degree of misspecification is defined as $\alpha = \sigma^2 - 1$. The model is correctly specified if $\alpha = 0$. The centered weight matrix is

$$\begin{aligned} W(\phi) &= E\left[m(X_i, \phi) m(X_i, \phi)'\right] - E\left[m(X_i, \phi)\right] E\left[m(X_i, \phi)\right]' \\ &= \begin{bmatrix} 1 + \alpha & -2(1+\alpha)\phi \\ -2(1+\alpha)\phi & 2(1+\alpha)(1+\alpha+2\phi^2) \end{bmatrix}. \end{aligned}$$

It is non-singular for all $\phi \in \Theta$ and for all $\alpha > -1$ because $\det(W(\phi)) = 2(1+\alpha)^3 \neq 0$. The GMM criterion function is

$$J(\theta, \phi) = \frac{\theta^4 - 4\phi\theta^3 + 2(2(1+\alpha) + 2\phi^2 - 1)\theta^2 - 4\alpha\phi\theta + \alpha^2}{2(1+\alpha)^2}.$$

If $\alpha = 0$ (correctly specified), the unique minimizer is $\theta_0 = 0$ regardless of ϕ. Under misspecification the minimizer generally depends on ϕ and α and a closed form expression for $g(\phi)$ is not available. Using numerical methods we verified that $g(\phi)$ is unique. By the implicit mapping theorem it follows that (formal justification is given in the proof of Theorem 2)

$$\frac{\partial}{\partial \phi'} g(\phi) = \left(\frac{\partial}{\partial \theta} \frac{\partial}{\partial \theta'} J(g(\phi), \phi) \right)^{-1} \frac{\partial}{\partial \phi} \frac{\partial}{\partial \theta'} J(g(\phi), \phi).$$

Being a contraction requires that $\sup_{\phi\in\Theta}\|(\partial/\partial\phi')g(\phi)\| < 1$. The fixed point is θ_0 which can be obtained by solving the population first-order condition evaluated at $\phi = \theta$. Since it can be numerically verified that $\sup_{\phi\in\Theta}\|(\partial/\partial\phi')g(\phi)\| = \|(\partial/\partial\phi')g(\phi)\|_{\phi=\theta_0}$ the degree of misspecification needs to be restricted as $\alpha > -1/3$. In other words, if $\sigma^2 \leq 2/3$ then the GMM mapping does not converge. Also it can be numerically verified that $\sup_{\phi\in\Theta} J(g(\phi),\phi) = J(\theta_0,\theta_0)$. Since $J(\theta_0,\theta_0) = \alpha^2/(2(1+\alpha)^2)^2 \geq 1/8$ for $\alpha \leq -1/3$, the upper bound in Assumption 1.6 can be set at $1/8$. In this example Assumption 1.6 is binding.

4 Existence of Iterated GMM Estimator

The iterated GMM estimator (5) is defined as the limit of the s-step estimator or equivalently as the fixed point (6). The existence of this limit has not been discussed previous. In this section we provide sufficient conditions.

Assumption 2.

1. *The observations X_i are independent.*
2. *For all $\theta \in \Theta$ and $j = 1,...,k$ the random variables $\|m(X,\theta)\|$, $\left\|\frac{\partial}{\partial\theta'}m(X,\theta)\right\|$, $\left\|\frac{\partial^2}{\partial\theta_j\partial\theta'}m(X,\theta)\right\|$, $\|v(X,\theta)\|^2$, and $\left\|\frac{\partial}{\partial\theta'}v(X,\theta)\right\|^2$ are uniformly integrable.*
3. *For $j_1, j_2 = 1,...,k$, $\sup_i E\left[\sup_{\theta\in\Theta}\left\|\frac{\partial^3}{\partial\theta_{j_1}\partial\theta_{j_2}\partial\theta'}m(X_i,\theta)\right\|\right] < \infty$.*
4. *For $j = 1,...,k$, $\sup_i E\left[\sup_{\theta\in\Theta}\left\|\frac{\partial^2}{\partial\theta_j\partial\theta'}v(X_i,\theta)\right\|^2\right] < \infty$.*

Assumption 2 is used to verify that a set of sample moments converge uniformly in probability to their expectations.

Theorem 3. *Under Assumptions 1 and 2, as $n \to \infty$*

1. $\sup_{\phi\in\Theta}\|\overline{g}_n(\phi) - g(\phi)\| \xrightarrow[p]{} 0$.
2. *With probability tending to one, the map $\overline{g}_n(\phi)$ is a contraction and the fixed point $\widehat{\theta}$ exists and is unique.*
3. $\left\|\widehat{\theta} - \theta_0\right\| \xrightarrow[p]{} 0$.

The proof is based on Dominitz and Sherman (2005, Theorem 2 and Lemma 3). They show that if the population mapping $g(\phi)$ is a contraction and $\overline{g}_n(\phi)$ and its derivative are uniformly consistent then $\overline{g}_n(\phi)$ is a contraction, the fixed point exists (both with probability tending to one), and the fixed point $\widehat{\theta}$ is consistent. We verify these conditions by applying the uniform law of large numbers to $\overline{g}_n(\phi)$ and its derivative.

Remark 1. The fixed point θ_0 and the estimator $\widehat{\theta}$ are linked to the class of minimum discrepancy (MD) estimators (Corcoran, 1998) defined as

$$\widetilde{\theta} = \underset{\theta\in\Theta,\pi_1,\ldots,\pi_n}{\arg\min} \sum_{i=1}^{n} h(\pi_i), \quad s.t. \quad \sum_{i=1}^{n} \pi_i m(X_i,\theta) = 0, \quad \sum_{i=1}^{n} \pi_i = 1,$$

where $h(\pi)$ is a convex function of a scalar π. In particular, Imbens, Spady, and Johnson (1998) and Newey and Smith (2004) show that the first-order conditions of the log Euclidean likelihood (LEL) and the CU-GMM both corresponding to $h(\pi) = n^{-1}(n^2\pi^2 - 1)$ take the form

$$0 = \left(\sum_{i=1}^{n} \widehat{\pi}_i Q(X_i,\widehat{\theta})\right)' \overline{W}_n(\widehat{\theta})^{-1}\overline{m}_n(\widehat{\theta}) \tag{16}$$

where $Q(x,\theta) = (\partial/\partial\theta')m(x,\theta)$, $\overline{W}_n(\theta) = n^{-1}\sum_{i=1}^{n} m(X_i,\theta)m(X_i,\theta)'$, and

$$\widehat{\pi}_i = 1 - \overline{m}_n(\widehat{\theta})'\overline{W}_n(\widehat{\theta})^{-1}m(X_i,\widehat{\theta}).$$

Thus, the difference between (16) and the first-order condition of the iterated GMM (17) is how $E\left[Q(X_i,\theta_0)\right]$ is estimated. The iterated GMM uses equal weights and the LEL uses the optimal[2] weights. If $Q(x,\theta)$ does not depend on x (e.g. $m(x,\theta) = \pi(x) + h(\theta)$), then

$$\sum_{i=1}^{n} \widehat{\pi}_i Q(X_i,\widehat{\theta}) = \left(\sum_{i=1}^{n} \widehat{\pi}_i\right) Q(\widehat{\theta})$$

and (16) can be written as

$$0 = Q(\widehat{\theta})'\widehat{W}_n(\widehat{\theta})^{-1}\overline{m}_n(\widehat{\theta}),$$

which is the iterated GMM first-order condition. The same is true for the population first-order conditions. Under misspecification, the pseudo-true values of the MD estimators are defined as the minimizer of the population version of the discrepancy (Schennach, 2007). For example, the empirical likelihood estimators ($h(\pi) = -\ln\pi$) of Owen (1988), Qin and Lawless (1994), and Imbens (1997) minimizes the discrepancy called the Kullback-Leibler information criterion. The exponential tilting estimator ($h(\pi) = \pi\ln\pi$) of Kitamura and Stutzer (1997) and Imbens, Spady, and Johnson (1998) minimize the discrepancy called the entropy. These are all well-defined discrepancy measures and in general it is difficult to argue that one is superior to others.

5 Why Iterate?

A reasonable question is why an empirical researcher should prefer the iterated GMM estimator $\widehat{\theta}$ relative to the 2-step estimator $\widehat{\theta}_2$. Under correct specification both have the same first-order asymptotic distribution. In the lack of a higher-order theory what is the rationale for preferring

[2]The optimality of the LEL weights does not hold under misspecification.

one estimator over the other? We give three reasons.

First, the iterated GMM estimator removes the arbitrariness induced by the choice of initial estimator $\widehat{\theta}_0$. Two researchers may select distinct yet reasonable initial estimators and will thus obtain two distinct two-step GMM estimators $\widehat{\theta}_2$. Unless there is a compelling reason to select a specific initial estimator $\widehat{\theta}_0$ there is no compelling reason to select one of the two two-step estimators over the other. By iterating until convergence the estimators will coincide and the arbitrariness will be eliminated.

Second, as we show in Theorem 4 below, the asymptotic distribution of the iterated GMM estimator takes a simpler form than that of the 2-step estimator found by Hall and Inoue (2003). This means that using the iterated estimator makes it more convenient to implement misspecification-robust inference.

Third and possibly most importantly, the iterated GMM estimator appears to have reduced variance. This argument is a bit heuristic, but is based on the contraction property. Since the iteration mapping is a contraction, each iteration is approximately variance reducing.

To see this, combine equation (4) with Theorem 3.1 and equation (14) in Theorem 2. We find that the s-step estimator approximately equals

$$\begin{aligned}\widehat{\theta}_s - \theta_0 &= \overline{g}_n(\widehat{\theta}_{s-1}) - \theta_0 \\ &\simeq g(\widehat{\theta}_{s-1}) - \theta_0 \\ &\simeq \frac{\partial}{\partial \theta'} g(\theta_0)\left(\widehat{\theta}_{s-1} - \theta_0\right).\end{aligned}$$

For example, in the scalar case

$$\operatorname{var}\left(\widehat{\theta}_s\right) \simeq c^2 \operatorname{var}\left(\widehat{\theta}_{s-1}\right)$$

where c is from Theorem 2. Since $c < 1$, iteration reduced the variance.

These arguments are heuristic and are not meant to be either finite sample nor asymptotically rigorous statements. Rather, the heuristic is that since the iteration operation is a contraction which is an approximately affine function with a slope less than unity, it is approximately variance reducing.

The one potential downside to iteration is computation cost. The computation of the iterated GMM estimator is roughly linear in the number of iterations. In linear models, including linear GMM, computation cost is negligible so this should not be a practical concern. In nonlinear models computational cost can be a consideration. However this could be avoided by linearizing the moment condition. For example, Lewbel and Pendakur (2009) use a form of iterated linear GMM estimator to estimate a highly nonlinear model. The moment conditions are linearized by evaluating the nonlinear components at the previous step estimate. They report that the iterated estimator is numerically very close to the fully nonlinear estimator.

6 Asymptotic Distribution

In this section we provide the asymptotic distribution of the iterated GMM estimator while allowing for possible moment misspecification. The iterated GMM estimator $\widehat{\theta}$ satisfies the first-order condition

$$0 = \frac{1}{2}\frac{\partial}{\partial\theta}\overline{J}_n(\theta,\widehat{\theta})\Big|_{\theta=\widehat{\theta}} = \overline{Q}_n(\widehat{\theta})'\overline{W}_n(\widehat{\theta})^{-1}\overline{m}_n(\widehat{\theta}). \tag{17}$$

The standard approach is to make a first-order Taylor expansion of $\overline{m}_n(\widehat{\theta})$ about $\overline{m}_n(\theta_0)$ and then apply a central limit theory to $\overline{m}_n(\theta_0)$. Under correct specification this is appropriate since $E\left[m(X_i,\theta_0)\right] = 0$. Under misspecification this latter condition does not hold so this expansion does not lead to a constructive solution.

To obtain the correct asymptotic distribution, we can instead expand the entire first-order condition rather than just the sample moment $\overline{m}_n(\widehat{\theta})$. Define

$$\overline{F}_n(\theta) = \overline{Q}_n(\theta)'\overline{W}_n(\theta)^{-1}\overline{m}_n(\theta)$$

which satisfies $0 = \overline{F}_n(\widehat{\theta})$. Expand $\overline{F}_n(\widehat{\theta})$ about θ_0 and rearranging, we find that

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \simeq -\overline{H}_n(\theta_0)^{-1}\sqrt{n}\,\overline{F}_n(\theta_0) \tag{18}$$

where

$$\begin{aligned}\frac{\partial}{\partial\theta'}\overline{F}_n(\theta) &= \frac{1}{2}\frac{\partial^2}{\partial\theta\partial\theta'}\overline{J}_n(\theta,\phi) + \frac{1}{2}\,\frac{\partial^2}{\partial\theta\partial\phi'}\overline{J}_n(\theta,\phi)\Big|_{\phi=\theta} \\ &= \overline{Q}_n(\theta)'\overline{W}_n(\theta)^{-1}\overline{Q}_n(\theta) + \left(\overline{m}_n(\theta)'\overline{W}_n(\theta)^{-1}\otimes I_k\right)\overline{R}_n(\theta) \\ &\quad - \left(\overline{m}_n(\theta)'\overline{W}_n(\theta)^{-1}\otimes \overline{Q}_n(\theta)'\overline{W}_n(\theta)^{-1}\right)\overline{S}_n(\theta) \\ &\equiv \overline{H}_n(\theta)\end{aligned} \tag{19}$$

and $\overline{R}_n(\theta) = \frac{\partial}{\partial\theta'}\operatorname{vec}(\overline{Q}_n(\theta)')$. (This and other calculations are justified in the appendix.)

Next, we expand $\overline{F}_n(\theta_0)$ in terms of sample moments. Set $\mu = m(\theta_0)$, $Q = Q(\theta_0)$, $W = W(\theta_0)$, $R(\theta) = \frac{\partial}{\partial\theta'}\operatorname{vec}(Q(\theta)')$, $R = R(\theta_0)$, and $S = S(\theta_0)$. Set $\overline{m}_n = \overline{m}_n(\theta_0)$, $\overline{Q}_n = \overline{Q}_n(\theta_0)$, and $\overline{W}_n = \overline{W}_n(\theta_0)$. The vector μ is the magnitude of the population moment condition evaluated at the pseudo-true values, and thus measures the magnitude of the degree of misspecification. The matrix Q is the derivative of the moment $m(\theta)$ and measures the degree of identification. The matrix R is the second derivative of the moment $m(\theta)$ and measures its *curvature*. The matrix S is the derivative of the weight matrix $W(\theta)$ and measures the *sensitivity* of the weight matrix to the parameter value.

In the appendix we show that

$$\sqrt{n}\,\overline{F}_n(\theta_0) = \sqrt{n}\widetilde{F}_n + o_p(1) \tag{20}$$

where

$$\widetilde{F}_n = Q'W^{-1}\overline{m}_n + \overline{Q}_n W^{-1}\mu - Q'W^{-1}\overline{W}_n W^{-1}\mu.$$

We set

$$\psi_i = Q'W^{-1}m(X_i,\theta_0) + Q(X_i,\theta_0)'W^{-1}\mu - Q'W^{-1}v(X_i,\theta_0)v(X_i,\theta_0)'W^{-1}\mu \tag{21}$$

so that

$$\sqrt{n}\widetilde{F}_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i. \tag{22}$$

The CLT can be applied to (22) which has variance

$$\Omega = \frac{1}{n}\sum_{i=1}^{n} E\left[\psi_i\psi_i'\right]. \tag{23}$$

Equations (18), (20), and (22) imply

$$\sqrt{n}\left(\widehat{\theta}-\theta_0\right) \simeq -\overline{H}_n(\theta_0)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i\right) + o_p(1).$$

This leads to an asymptotic distribution theory for $\widehat{\theta}$.

We now provide regularity conditions and a formal statement. Define

$$\begin{aligned} H(\theta) &= \frac{1}{2}\left(\frac{\partial^2}{\partial\theta\partial\theta'}J(\theta,\phi) + \frac{\partial^2}{\partial\theta\partial\phi'}J(\theta,\phi)\right)\Big|_{\phi=\theta} \\ &= Q(\theta)'W(\theta)^{-1}Q(\theta) + \left(m(\theta)'W(\theta)^{-1}\otimes I_k\right)R(\theta) - \left(m(\theta)'W(\theta)^{-1}\otimes Q(\theta)'W(\theta)^{-1}\right)S(\theta) \end{aligned}$$

and

$$H = H(\theta_0) = Q'W^{-1}Q + \left(\mu'W^{-1}\otimes I_k\right)R - \left(\mu'W^{-1}\otimes Q'W^{-1}\right)S. \tag{24}$$

The matrix H plays an important role in the asymptotic distribution. Its leading component $Q'W^{-1}Q$ is the inverse of the asymptotic covariance matrix under correct specification when W is the efficient weight matrix. The second term in H is proportional to μ (the magnitude of misspecification) and R (the curvature in the moment condition). The third term in H is proportional to μ and S (the sensitivity of the weight matrix to the parameters). Thus H will be close to $Q'W^{-1}Q$ when the degree of misspecification is small, and/or the curvature of $m(\theta)$ and sensitivity of $W(\theta)$ are small. Otherwise H will differ from $Q'W^{-1}Q$.

Define $V = H^{-1}\Omega H^{-1\prime}$. Let $\mathcal{N}$ be some neighborhood of θ_0.

Assumption 3.

1. *The random variables* $\|m(X_i,\theta)\|^2$, $\left\|\frac{\partial}{\partial\theta'}m(X_i,\theta)\right\|^2$, *and* $\|v(X_i,\theta)\|^4$ *are uniformly integrable for all* $\theta\in\mathcal{N}$.

2. $\lambda_{\min}\left(H'H\right) \geq C > 0$.

3. $\lambda_{\min}(\Omega) \geq C > 0$.

Assumption 3.1 is necessary in order to apply the central limit theorem to (22). Assumptions 3.2 and 3.3 exclude singular covariance matrices.

Theorem 4. *Under Assumptions 1-3, as* $n \to \infty$

$$V^{-1/2}\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \underset{d}{\longrightarrow} N(0, I_k). \tag{25}$$

Theorem 4 provides a simple characterization of the asymptotic distribution of the iterated GMM estimator under mild moment misspecification.

The asymptotic variance in Theorem 4 differs from the classical formula

$$(Q'W^{-1}Q)^{-1}\left(Q'W^{-1}\Omega_{11}W^{-1}Q\right)\left(Q'W^{-1}Q\right)^{-1}$$

where Ω_{11} is (23) with $\mu = 0$, in two ways. First, the matrix H defined in (24) differs from $Q'W^{-1}Q$ as discussed above. The magnitude of this difference depends on the degree of misspecification, the curvature in $m(\theta)$ and the sensitivity of $W(\theta)$. Second, the asymptotic covariance matrix Ω defined in (23) of the vector ψ_i is an augmented version of the classic covariance matrix. Ω is augmented by the variation in $Q(X_i, \theta_0)$ and $v(X_i, \theta_0)v(X_i, \theta_0)'$. Larger variance in these variables implies larger differences. These differences disappear under correct specification.

The asymptotic distribution in Theorem 4 is similar to that obtained by Hall and Inoue (2003). They are equivalent when $W(\theta)$ does not depend on θ (which excludes iterated GMM with an efficient weight matrix). Theorem 4 is the first distribution theory which formally covers the iterated GMM estimator, both under correct specification and misspecification.

The assumptions allow for heterogeneous distributions, thus the asymptotic covariance matrix V may vary with sample size. We have not indexed the population moments by n to keep the notation uncluttered, but it is useful to observe that the assumptions do not impose homogeneity. This is why the asymptotic distribution in (25) is written in self-normalized notation.

7 Covariance Matrix Estimation

It is straightforward to calculate an estimator of the asymptotic covariance matrix V. Construct the estimators $\widehat{Q} = \overline{Q}_n(\widehat{\theta})$, $\widehat{R} = \overline{R}_n(\widehat{\theta})$, $\widehat{S} = \overline{S}_n(\widehat{\theta})$, $\widehat{W} = \overline{W}_n(\widehat{\theta})$, $\widehat{\mu} = \overline{m}_n(\widehat{\theta})$ and

$$\widehat{H} = \overline{H}_n(\widehat{\theta}) = \widehat{Q}'\widehat{W}^{-1}\widehat{Q} + (\widehat{\mu}'\widehat{W}^{-1} \otimes I_k)\widehat{R} - (\widehat{\mu}'\widehat{W}^{-1} \otimes \widehat{Q}'\widehat{W}^{-1})\widehat{S}. \tag{26}$$

An alternative numerical method is

$$\widehat{H} = \frac{1}{2}\left(\frac{\partial^2}{\partial\theta\partial\theta'}\overline{J}_n(\theta,\phi) + \frac{\partial^2}{\partial\theta\partial\phi'}\overline{J}_n(\theta,\phi)\right)\Big|_{\phi=\theta=\widehat{\theta}}.$$

We use (26) for our numerical and empirical calculations.

Construct the vectors

$$\widehat{\psi}_i = \widehat{Q}'\widehat{W}^{-1}m(X_i,\widehat{\theta}) + Q(X_i,\widehat{\theta})'\widehat{W}^{-1}\widehat{\mu} - \widehat{Q}'\widehat{W}^{-1}v(X_i,\widehat{\theta})v(X_i,\widehat{\theta})'\widehat{W}^{-1}\widehat{\mu},$$

the variance estimator

$$\widehat{\Omega} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\psi}_i\widehat{\psi}_i', \tag{27}$$

and

$$\widehat{V} = \widehat{H}^{-1}\widehat{\Omega}\widehat{H}^{-1\prime}. \tag{28}$$

The asymptotic standard errors for $\widehat{\theta}$ are the square roots of the diagonal elements of $n^{-1}\widehat{V}$.

We now establish that $\widehat{V}$ is consistent for V and that replacement in the asymptotic distribution of V by $\widehat{V}$ has no effect.

Assumption 4.

1. $\sup_i E\left[\sup_{\theta\in\mathcal{N}}\left\|\frac{\partial^2}{\partial\theta_j\partial\theta'}m(X_i,\theta)\right\|^2\right] < \infty.$
2. $\sup_i E\left[\sup_{\theta\in\mathcal{N}}\left\|\frac{\partial}{\partial\theta'}v(X_i,\theta)\right\|^4\right] < \infty.$

Assumption 4 is used to establish the uniform convergence of $\widehat{\Omega}$.

Theorem 5. *Under Assumptions 1-4,*

$$\left\|\widehat{V} - V\right\| \underset{p}{\longrightarrow} 0$$

and

$$\widehat{V}^{-1/2}\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \underset{d}{\longrightarrow} N\left(0, I_k\right) \tag{29}$$

as $n \to \infty$.

Equation (29) implies that test statistics constructed with $\widehat{V}$ have standard asymptotic distributions. In particular, t-statistics are asymptotically standard normal, and Wald statistics have asymptotic chi-square distributions. Conventional confidence intervals constructed with the GMM estimator and our proposed standard errors have asymptotically correct coverage for the pseudo-true values.

To emphasize, Theorem 5 shows that robust t-statistics and Wald statistics calculated with (28) have conventional asymptotic distributions. The result fails if the standard covariance matrix estimator is used.

8 Simulation

In this section we illustrate our methods in two simulation experiments, one for a linear model and one for a nonlinear model. For both models we investigate inference for iterated efficient GMM

estimation using conventional and our recommended robust standard errors. The Windmeijer (2000, 2005) corrected standard errors are also calculated for the linear model. We find large and important improvements in performance by using our recommended methods. All calculations use 5000 Monte Carlo replications.

Our first experiment concerns a simple linear instrumental variable regression with a single endogenous regressor. The model is

$$\begin{aligned} Y_i &= X_i\theta_0 + \varepsilon_i \\ E\left[Z_i\varepsilon_i\right] &= 0 \end{aligned} \tag{30}$$

where X_i and θ_0 are scalar and $Z_i = (Z_{1i}, Z_{2i}, Z_{3i}, Z_{4i})'$ is a vector of instrumental variables. We estimate θ_0 by iterated efficient GMM, and calculate standard errors using the conventional heteroskedasticity-robust formula,, the Windmeijer corrected formula, and our misspecification-robust formula.

Our data-generating process is

$$\begin{aligned} Y_i &= X_i + \alpha\left(Z_{1i} - Z_{2i} + Z_{3i} - Z_{4i}\right) + e_i, \\ X_i &= \pi\left(Z_{1i} + Z_{2i} + Z_{3i} + Z_{4i}\right) + u_i, \\ Z_i &\sim_{i.i.d} N\left(0, I_4\right), \\ \begin{pmatrix} e_i \\ u_i \end{pmatrix} &\sim_{i.i.d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix}\right) \end{aligned} \tag{31}$$

with Z_i independent of (e_i, u_i). We vary α from 0 to 1 in steps of 0.1, and set the first-stage coefficient π so that the first-stage $R^2 = 0.20$ or 0.02, corresponding to relatively strong and weak instrument settings. We set the number of observations as $n = 250$ and 2500.

The key parameter is α. At $\alpha = 0$, the model is correctly specified. For $\alpha \neq 0$ we find

$$E\left[Z_i(Y_i - X_i\theta_0)\right] = \begin{pmatrix} \alpha \\ -\alpha \\ \alpha \\ -\alpha \end{pmatrix} \neq 0 \tag{32}$$

so the moment condition (30) fails to hold. Note that (32) is the moment condition evaluated at the pseudo-true value. The instruments are invalid due to the violation of the exclusion restriction. The model is designed so that the pseudo-true value θ_0 is invariant to α.[3] (This choices eases reporting and interpretation. It is worth mentioning once again that one challenge with inference allowing for misspecification is that in general the pseudo-true parameter value is not invariant to misspecification.)

[3]Since the model is linear the pseudo-true value can be obtained by solving the population first-order condition, $Q'W(\theta_0)^{-1}m(\theta_0) = 0$.

The results are reported in Table 1. In the fourth column we report the ratio of the mean of our proposed misspecification-robust standard errors relative to the actual standard deviation of $\widehat{\theta}$ across the 5000 simulation replications. The standard error is unbiased if this ratio is 1, is biased downwards for values less than 1, and biased upwards for values greater than 1. We can see that under strong identification ($R^2 = 0.20$) our proposed standard errors are nearly unbiased in all cases examined. Under weak identification ($R^2 = 0.02$) our proposed standard errors are upward biased for $n = 250$, but nearly unbiased for $n = 2500$.

In the fifth column we report the ratio of the mean of the Windmeijer corrected standard errors relative to the standard deviation of $\widehat{\theta}$. The smaller size distortion and reduced bias in the Windmeijer standard errors under correct specification ($\alpha = 0$) are clearly seen, as is shown in Windmeijer (2000, 2005). In contrast, the bias of the Windmeijer standard error increases in the degree of misspecification. The standard errors are moderately downward biased under strong identification but severely (60%) biased under weak identification. This is consistent with the finding of Hwang, Kang, and Lee (2020) that the Windmeijer formula only partially corrects the misspecification bias.

In the sixth column we report the ratio of the mean of the conventional heteroskedasticity-robust standard errors relative to the standard deviation of $\widehat{\theta}$. We can see that the standard errors are unbiased for $\alpha = 0$ but highly biased for $\alpha \neq 0$. The standard errors are downward biased, meaning that the reported standard errors understate estimation uncertainty. The bias is severe even for the smallest departure from $\alpha = 0$, with approximately a 30% downwards bias for $\alpha = 0.2$ under strong identification, and a 50-60% downward bias under weak identification. The bias is increasing in α and does not improve with sample size. Indeed the worst case arises for $\alpha = 1$, $R^2 = 0.02$, and $n = 2500$, where the conventional standard error is about one-fifth the true standard deviation. These results demonstrate unambiguously that the conventional heteroskedasticity-robust standard errors are severely affected by moment misspecification.

In columns seven to nine we report the size of nominal asymptotic 5% t-tests for $H_0 : \theta_0 = 1$ against $H_1 : \theta_0 \neq 1$. Column seven reports the size of tests using our proposed misspecification-robust standard errors. We can see that there is meaningful size distortion from our misspecification-robust t-tests when the sample size is small (rejection rates range from 6% to 12% under strong identification, and from 7% to 8% under weak identification), but this disappears as the sample size increases. Column eight reports the size of the test using the Windmeijer corrected standard errors. Similar to column five there is moderate size distortion (rejection rates range from 6% to 13.7%) under strong identification and it improves as the sample size increases (range from 5% to 9%). Under weak identification, the test exhibits severe size distortion under misspecification (rejection rates range from 22% to 32%) and it worsens as the sample size increases (range from 35% to 47%). Column nine reports the size of the test using the conventional heteroskedasticity-robust standard errors. We can see that the test is highly over-sized with the size distortion increasing in the degree of misspecification, as the strength of the instruments weaken, and as the sample size increases. The rejection rates are severely distorted even for the mildest departures from correct specification

in the presence of weak instruments. Indeed, the size of the t-test is 46.7% for $\alpha = 0.2$ and $n = 2500$ and exceeds 70% for $\alpha > 0.6$.

In column ten we report rejection rates for the J test using the asymptotic 5% critical value. While the J test will properly detect misspecification when $n = 2500$, it may not when $n = 250$, especially in the presence of weak instruments.

In the final column we report the median number of iterations required to obtain convergence, which is defined as $\|\widehat{\theta}_s - \widehat{\theta}_{s-1}\| < 10^{-5}$. The results show that the number of required iterations is increasing in the degree of misspecification. This is consistent with Assumption 1.4 which is used to establish the convergence of the GMM iteration sequence. As misspecification increases the contraction property weakens and thus iterative convergence slows. It is noteworthy that in all our simulation runs the GMM iteration sequence did converge.

Our theory does not cover the case of weak instruments. To investigate the impact of weak instruments we computed a further simulation using $n = 2500$ and $R^2 = 0.002$, which corresponds to a concentration parameter similar to the $n = 250$ and $R^2 = 0.02$ case. As might be expected, we found the performance of the method to be similar to the $n = 250$ and $R^2 = 0.02$ case. This illustrates that our methods are not robust to very weak instruments.

It is worth pointing out the behavior of the statistics when there is no misspecification ($\alpha = 0$). In this setting both conventional and robust methods are appropriate, and in fact one might expect the conventional methods to work better since the covariance matrix is estimating fewer terms. However, the misspecification-robust t-statistic has less size distortion than the conventional t-statistic, in particular when the sample size is small and the instruments are weak. Surprisingly, this is not a coincidence. Hwang, Kang, and Lee (2020) show that the misspecification-robust standard errors provide finite sample corrections up to the same order with the Windmeijer correction under correct specification for linear GMM assuming strong identification. Thus, it is strongly preferred to use our new robust standard error for linear models.

Our second experiment involves a simple nonlinear model of Schennach (2007), which is presented in Example 3 in Section 3. We estimate θ_0 by iterated efficient GMM using the centered weight matrix and calculate the standard error using the conventional heteroskedasticity-robust formula and our misspecification-robust formula. Replications where convergence failed are excluded.

Since we have shown in Example 3 that the GMM mapping does not converge if $\alpha \leq -1/3$, we set $\alpha = -0.3$, -0.25, 0 (correct specification), 0.5, 1, and 2. The iterated efficient GMM pseudo-true value is $\theta_0 = 0$ which is invariant to α. It can be found by solving the first-order condition and setting $\phi = \theta$ because the first-order condition has a unique root. We set the number of observations as $n = 250$ and 2500.

The overall results are quite similar to those presented in Table 1, showing excellent performance of our misspecification-robust standard errors and t tests across α relative to the conventional standard errors and t tests. Notably, the performance of our proposed standard errors is not affected conditional on convergence. At $\alpha = -0.3$, which is close to the value $\alpha = -1/3$ where convergence fails, the convergence failure rate is about 36% of the repetitions, which are reported

R^2	n	α	$s_r(\hat{\theta})/s.d$	$s_w(\hat{\theta})/s.d$	$s(\hat{\theta})/s.d$	Size(t_r)	Size(t_w)	Size(t)	Reject(J)	Med. Iter
0.2	250	0	1.012	0.999	0.974	0.058	0.059	0.066	0.052	3
		0.2	1.001	0.919	0.712	0.066	0.079	0.162	0.998	7
		0.4	0.997	0.902	0.511	0.078	0.098	0.312	1.000	11
		0.6	0.980	0.895	0.433	0.105	0.119	0.408	1.000	14
		0.8	0.973	0.896	0.393	0.109	0.127	0.450	1.000	15
		1	0.955	0.882	0.372	0.121	0.137	0.488	1.000	16
	2500	0	0.990	0.989	0.987	0.051	0.051	0.052	0.050	2
		0.2	1.011	0.935	0.732	0.047	0.069	0.157	1.000	6
		0.4	0.992	0.891	0.501	0.060	0.087	0.319	1.000	11
		0.6	1.001	0.900	0.415	0.053	0.081	0.414	1.000	14
		0.8	0.971	0.876	0.363	0.062	0.091	0.477	1.000	16
		1	0.980	0.887	0.345	0.067	0.094	0.507	1.000	18
0.02	250	0	1.277	1.013	0.992	0.083	0.116	0.125	0.054	4
		0.2	1.096	0.594	0.516	0.068	0.223	0.284	0.861	6
		0.4	1.270	0.541	0.438	0.077	0.276	0.426	0.912	8
		0.6	1.004	0.504	0.402	0.081	0.312	0.490	0.910	8
		0.8	1.019	0.505	0.396	0.076	0.327	0.527	0.913	9
		1	1.003	0.501	0.391	0.074	0.318	0.533	0.915	9
	2500	0	1.048	1.018	1.016	0.048	0.053	0.054	0.051	3
		0.2	1.007	0.463	0.370	0.060	0.349	0.467	1.000	7
		0.4	1.002	0.397	0.247	0.058	0.441	0.661	1.000	10
		0.6	1.000	0.388	0.215	0.064	0.470	0.721	1.000	12
		0.8	0.992	0.386	0.204	0.067	0.474	0.745	1.000	14
		1	1.010	0.393	0.199	0.062	0.469	0.751	1.000	15

Table 1: Monte Carlo Results for Linear Model

n	α	$s_r(\hat{\theta})/s.d$	$s(\hat{\theta})/s.d$	Size(t_r)	Size(t)	Reject(J)	Med. Iter	Failed Conv.
250	-0.3	0.994	0.918	0.035	0.048	0.623	36	36.6%
	-0.25	0.987	0.937	0.051	0.060	0.805	13	13.4%
	0	1.002	0.997	0.050	0.051	0.063	3	0%
	0.5	0.997	0.920	0.051	0.073	0.988	7	0%
	1	1.004	0.859	0.053	0.095	1.000	8	0%
	2	0.999	0.787	0.054	0.128	1.000	10	0%
2500	-0.3	0.983	0.875	0.051	0.079	0.927	25	7.3%
	-0.25	1.007	0.935	0.050	0.062	1.000	11	0%
	0	0.997	0.996	0.050	0.050	0.050	2	0%
	0.5	1.010	0.935	0.048	0.063	1.000	6	0%
	1	0.990	0.845	0.052	0.100	1.000	8	0%
	2	1.006	0.782	0.048	0.125	1.000	9	0%

Table 2: Monte Carlo Results for Nonlinear Model

in the final column of the table.

9 Application: Income and Democracy

In an influential paper, Acemoglu, Johnson, Robinson, and Yared (2008, AJRY hereinafter) find no evidence of a causal effect of income on democracy. This contrasts to the conventional wisdom that income has a positive causal effect. AJRY estimate the dynamic panel regression

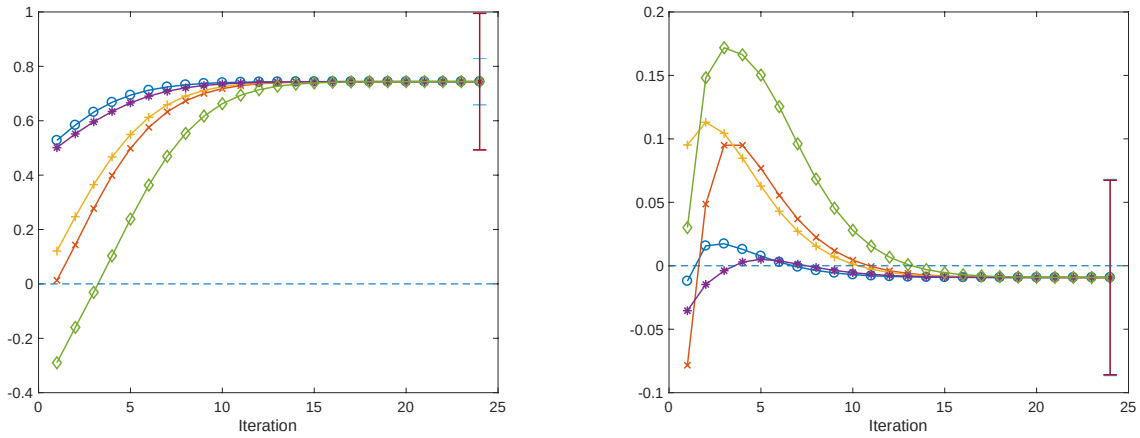
$$Y_{it} = \alpha Y_{i,t-1} + \gamma X_{i,t-1} + \mu_t + \delta_i + u_{it}, \tag{33}$$

where Y_{it} is a measure of democracy, X_{it} is log income per capita, μ_t is a time effect, and δ_i is a country fixed effect. The error term u_{it} has mean zero for all i and t. The parameter of interest is γ, the effect of income on democracy. Their data set includes 127 countries observed over 1960-2000 at both 5-year and 10-year frequencies. While AJRY consider several estimators we focus on the one-step GMM estimator of Arellano and Bond (1991). This is an overidentified GMM setting. It is unlikely that the AJRY model is correctly specified as the dynamics, functional form, and coefficient homogeneity assumption could all be incorrect. This is a natural application to investigate the impact of our proposed estimators.

Dynamic GMM applications with $T > 3$ time periods have clustered dependence structures. Therefore clustered variance estimation is used. The extension of our variance estimators to the clustered setting is reasonably straightforward and is described in our companion papers Hansen and Lee (2019, 2020)[4].

The AJRY estimates of (33) are reported in their Table 2. We repeat their estimates in Table 3 below. Columns I and III are the estimates reported in AJRY (one-step GMM). Columns II

[4]The theoretical treatment in the original version of this paper allowed for clustered dependence but was removed at the Co-Editor's request to allow a clean focus on the primary issues.



(a) Democracy$_{t-1}$ (b) Income$_{t-1}$

Figure 1: Convergence path of the iterated GMM estimator in Column II of Table 3 with initial weight matrices $\overline{W}_n(\widetilde{\phi})$ where $\widetilde{\phi}$ is the one-step GMM of AJRY (circle); $\widetilde{\phi}$ is the one-step GMM using the identity matrix (x); $\widetilde{\phi} = (0, 0, \cdots)$ (+); $\widetilde{\phi} = (0.5, 0.5, \cdots)$ (∗); $\widetilde{\phi} = (-0.5, -0.5, \cdots)$ (diamond), Thicker (lighter) error bar at the converging point is the asymptotic 95% confidence interval based on the robust (conventional) standard error.

and IV are iterated GMM (which are not reported in AJRY). We report Arellano-Bond standard errors and our new misspecification-robust standard errors, both clustered by country. We report the number of instruments used, the number of total observations, the number of countries G, and the p-value of the over-identifying restrictions J test. The J statistics are constructed using the uncentered clustered efficient weight matrix.

We focus on two issues: (1) The difference between the one-step and iterated estimates. (2) The difference between the Arellano-Bond and misspecification-robust standard errors.

First, we find that the difference between the one-step and iterated GMM estimates can be large. For example, in the five-year data the one-step point estimate for γ is -0.129 while the iterated GMM estimate is -0.009. This large difference means that one-step estimation is sensitive to the initial estimator. Two econometricians with different initial weight matrices will find two meaningfully different estimates. Any choice except the iterated solution is arbitrary.

To emphasize the strong and arbitrary dependence of the GMM estimator on the initial weight matrix and the importance of iterating until convergence, we display in Figure 1 the point estimates for $\widehat{\alpha}$ (panel (a)) and $\widehat{\gamma}$ (panel (b)) as a function of the iteration. Five lines are plotted corresponding to distinct starting values. Also displayed are the asymptotic 95% confidence intervals for the iterated GMM estimates. What can be seen is that while the sequence of GMM estimates converge to a well-defined limit as the number of iterations increase, the convergence takes a fairly large number of iterations (over 20). While the change in the point estimates between iterations is small, the overall change by iterating to convergence is substantial. Quite intriguingly, the income coefficient iterates are non-monotonic. This demonstrates substantial arbitrariness of using any estimator other than iterated GMM.

Second, the difference between the two sets of standard errors can be large. For example, in the five-year data the misspecification-robust standard error for the iterated estimator of the lagged democracy coefficient is about three times the Arellano-Bond standard error. This shows that taking into account misspecification can make an enormous difference in the standard errors.

The p-values of the over-identification J tests provide mixed answers to the validity of model specifications but overall the tests suggest that the dynamic panel regression equation may be misspecified. This is consistent with our finding that standard errors are affected by the use of the misspecification-robust formula. However, note that the iterated GMM estimates have only mildly significant J statistics (the p-value is 0.09). In this context it is common for applied researchers to treat the statistic as "borderline significant" and continue with their analysis unadjusted. Our view is that regardless of the value of the J statistic it is better to report the misspecification-robust standard errors as these are agnostic to whether the model is correctly specified or mildly misspecified.

It is noteworthy that the p-values of the J test reported in AJRY are 0.26 and 0.07, which are different from our values 0.04 and 0.08 in Columns I and III. The reason is that their statistic is the two-step GMM criterion with the efficient weight matrix evaluated at the one-step estimate while ours is the two-step GMM criterion with both the moment and efficient weight matrix being evaluated at the two-step estimate. Comparing the p-values one finds that the J test can be quite sensitive to the user's choice of the estimate where the efficient weight matrix is evaluated. The result of AJRY is based on the popular Stata command `xtabond2` for dynamic panel models. The command calculates the J statistic and the p-value based on the two-step GMM with the one-step weight matrix even when the one-step GMM estimates are reported. This does not seem to be a reliable test but there is no clear guideline on this. We recommend using the J statistic evaluated at the iterated GMM, which is not subject to such arbitrariness.

What are the causes of potential misspecification? One possibility is that the dynamic structure in (33) is incorrect – that lagged values are omitted. If the dynamics are misspecified, then the moment conditions are not satisfied and the Arellano-Bond standard errors will be incorrect. Since the "true" dynamic structure of a panel regression is not known *a priori*, this is a strong reason to generically allow for misspecification.

Another reason for potential misspecification is coefficient heterogeneity. If the coefficients are heterogeneous across countries, then moment conditions will not be satisfied. For example, in model (33), if the coefficient γ_i (the effect of income on democracy) varies with country i, then the moment conditions will be invalid. To see this, if we set $\gamma = E[\gamma_i]$ as the mean coefficient, then the effective error in the differenced equation (33) is $\Delta u_{it} + (\gamma_i - \gamma)\Delta X_{i,t-1}$ which will be correlated with the instrument $Y_{i,t-2}$.

There is strong evidence for coefficient heterogeneity in equation (33). Cervellati, Jung, Sunde, and Vischer (2014, CJSV hereinafter) argue that the income effect is heterogeneous between former colonies and non-colonies, and furthermore within colonies based on the quality of political institutions. Bonhomme and Manresa (2015) find evidence of grouped patterns of unobserved het-

	Column (4) five-year data		Column (8) ten-year data	
	One-step	Iterated	One-step	Iterated
	I	II	III	IV
Democracy_{t-1}	0.489	0.744	0.227	0.288
Arellano-Bond s.e.	(0.085)	(0.043)	(0.123)	(0.111)
Misspecification-Robust s.e.	(0.095)	(0.128)	(0.125)	(0.146)
Income_{t-1}	-0.129	-0.009	-0.318	-0.280
Arellano-Bond s.e.	(0.076)	(0.040)	(0.180)	(0.170)
Misspecification-Robust s.e.	(0.088)	(0.039)	(0.183)	(0.202)
Cumulative Income Effect	-0.253	-0.036	-0.411	-0.393
Arellano-Bond s.e.	(0.148)	(0.152)	(0.243)	(0.243)
Misspecification-Robust s.e.	(0.163)	(0.149)	(0.246)	(0.290)
Hansen J Test	[0.04]	[0.42]	[0.08]	[0.09]
# of Iteration	0	23	0	9
# of Instruments	55		15	
Observations	838		338	
Countries	127		118	

Standard errors clustered by country

Table 3: Extension of Acemoglu, Johnson, Robinson and Yared (2008), Table 2

erogeneity in the same dataset. Lu and Su (2017) also find strong evidence of heterogeneity in the income effect across countries. This literature makes a clear case that the coefficients (primarily γ) vary across countries. In this case, model (33) should be viewed as an approximation rather than a tight statistical model. The coefficients should be viewed as projections and the moment conditions acknowledged to be potentially invalid.

To highlight this issue further we examine a key table from CJSV (their Table 4) where they present Arellano-Bond estimates of model (33) augmented to allow the income effect to vary across groups. Their model is

$$Y_{it} = \alpha Y_{i,t-1} + \gamma X_{i,t-1} + \phi X_{i,t-1} c_i + \mu_t + \delta_i + u_{it}$$

where c_i is a country-specific dummy variable for "historically strong institutions". (Acemoglu, Johnson, Robinson and Yared (2009) make a similar distinction, describing colonies with "historically weak institutions" as "extractive".) CJSV estimate this model for the sub-sample of former colonies using three distinct measures of institutional quality: (i) the level of constraints on the executive in 1900; (ii) whether the country became independent before 1900; and (iii) whether the colony was subject to the rule of a late colonial power. We repeat their estimates in Table 4 below for the five-year sample. CJSV reported one-step Arellano-Bond estimates and standard errors

	Constraints		Independence		No Late Colonial	
	One-step	Iterated	One-step	Iterated	One-step	Iterated
	I	II	III	IV	V	VI
Democracy_{t-1}	0.289	-0.423	0.343	0.724	0.355	0.666
Arellano-Bond s.e.	(0.123)	(0.039)	(0.110)	(0.044)	(0.101)	(0.040)
Misspecification-Robust s.e.	(0.142)	(0.380)	(0.127)	(0.152)	(0.115)	(0.125)
Income_{t-1}	-0.417	-0.337	-0.270	-0.011	-0.303	-0.052
Arellano-Bond s.e.	(0.194)	(0.116)	(0.113)	(0.050)	(0.110)	(0.047)
Misspecification-Robust s.e.	(0.221)	(0.289)	(0.134)	(0.047)	(0.122)	(0.041)
$\text{Income}_{t-1} \times c_i$	0.345	0.296	0.224	0.020	0.318	0.111
Arellano-Bond s.e.	(0.162)	(0.073)	(0.121)	(0.037)	(0.122)	(0.039)
Misspecification-Robust s.e.	(0.169)	(0.309)	(0.125)	(0.077)	(0.130)	(0.053)
Hansen J Test	[0.03]	[0.02]	[0.04]	[0.38]	[0.15]	[0.37]
# of Iteration	0	297	0	32	0	28
# of Instruments	56		56		56	
Observations	531		628		631	
Countries	79		99		100	

Standard errors clustered by country

Table 4: Extension of Cervellati, Jung, Sunde, and Vischer (2014), Table 4

which are reported in our columns I, III, and V. In addition, we report iterated GMM estimates (in columns II, IV, and VI) and misspecification-robust standard errors.

We find that some of the iterated GMM estimates are quite different from the one-step estimates. This means that the one-step estimates are dependent on the initial weight matrix. We also find that the misspecification-robust standard errors differ from the Arellano-Bond standard errors; in some cases they are three to four times as large. Examining the over-identification J tests we find that three of the six p-values[5] are significant at the 5% level indicating potential misspecification.

Turning to the question raised by CJSV – is there heterogeneity in the income effect across institutional structure? – our results (iterated GMM with misspecification-robust standard errors) are that in two of the three specifications the t-statistics for $\phi = 0$ are statistically far from significant. This is due to both smaller coefficient estimates and larger standard errors relative to the results reported in CJSV. In the third specification (no late colonial power) the t-ratio of 2.1 is marginally significant at the 5% level. Our conclusion is that there is no strong evidence of the heterogeneity allegedly found by CJSV.

While this finding (no statistical evidence of coefficient heterogeneity) may appear to contradict our claim of possible misspecification in the AJRY analysis, the key is the need for standard errors to be robust to *potential* misspecification. Only by using robust standard errors can we make

[5]The reported p-values of the J test in CJSV are different from our p-values for the same reason given in the AJRY analysis.

inferences which are not fragile to specification choices.

Finally, we believe the empirical analysis reported in this section reveals that the misspecification-robust standard errors are useful for empirical practice. As we have illustrated they are helpful not only to provide robust inference but to develop improved specifications.

Appendix

Proof of Theorem 1: By the Woodbury matrix identity,

$$\overline{W}_n(\phi)^{-1} = \left[\overline{W}_n^*(\phi) + \overline{m}_n(\phi)\overline{m}_n(\phi)'\right]^{-1} = \overline{W}_n^*(\phi)^{-1} - \frac{W_n^*(\phi)^{-1}\overline{m}_n(\phi)\overline{m}_n(\phi)'\overline{W}_n^*(\phi)^{-1}}{1 + \overline{m}_n(\phi)'\overline{W}_n^*(\phi)^{-1}\overline{m}_n(\phi)}.$$

Thus

$$\begin{aligned}\overline{J}_n(\theta,\phi) &= \overline{m}_n(\theta)'\overline{W}_n^*(\phi)^{-1}\overline{m}_n(\theta) - \frac{\left(\overline{m}_n(\theta)'W_n^*(\phi)^{-1}\overline{m}_n(\phi)\right)^2}{1 + \overline{m}_n(\phi)'\overline{W}_n^*(\phi)^{-1}\overline{m}_n(\phi)}. \\ &= \overline{J}_n^*(\theta,\phi)\left(1 - \rho_n(\theta,\phi)^2\frac{\overline{J}_n^*(\phi,\phi)}{1+\overline{J}_n^*(\phi,\phi)}\right)\end{aligned}$$

where

$$\rho_n(\theta,\phi)^2 = \frac{\left(\overline{m}_n(\theta)'W_n^*(\phi)^{-1}\overline{m}_n(\phi)\right)^2}{\overline{J}_n^*(\phi,\phi)\overline{J}_n^*(\theta,\phi)}$$

is the squared weighted correlation between $\overline{m}_n(\theta)$ and $\overline{m}_n(\phi)$. Note that $\overline{J}_n^*(\theta,\widehat{\theta}^*)$ is minimized over θ at $\widehat{\theta}^*$ and $\rho_n(\theta,\widehat{\theta}^*)^2$ is maximized at $\widehat{\theta}^*$. It follows that $\overline{J}_n(\theta,\widehat{\theta}^*)$ is minimized over θ at $\widehat{\theta}^*$. This means that $\widehat{\theta}^*$ solves the fixed point of $\overline{g}_n(\widehat{\theta}^*) = \widehat{\theta}^*$ so $\widehat{\theta} = \widehat{\theta}^*$ as claimed. ■

Proof of Theorem 2: $g(\phi)$ is an interior minimizer of $J(\theta,\phi)$ so solves the first-order condition

$$0 = \frac{\partial}{\partial\theta}J(\theta,\phi) = 2Q(\theta)'W(\phi)^{-1}m(\theta). \tag{34}$$

Since (34) is continuously differentiable under Assumptions 1.3 and 1.4, and $W(\phi)$ is uniformly invertible under Assumption 1.5, it follows by the implicit function theorem that $g(\phi)$ exists, is continuously differentiable, and its derivative equals

$$\frac{\partial}{\partial\phi'}g(\phi) = -D(\phi)^{-1}B(\phi) \tag{35}$$

where

$$\begin{aligned}D(\phi) &= \frac{\partial^2}{\partial\theta\partial\theta'}J(\theta,\phi)|_{\theta=g(\phi)} \\ B(\phi) &= \frac{\partial^2}{\partial\theta\partial\phi'}J(\theta,\phi)|_{\theta=g(\phi)}.\end{aligned}$$

We calculate that

$$B(\phi) = 2\left(m(g(\phi))' \otimes Q(g(\phi))'\right) \frac{\partial}{\partial \phi'} \operatorname{vec}\left(W(\phi)^{-1}\right). \tag{36}$$

Using (11), (12) and Assumption 1.6 we find that

$$\begin{aligned}
\|B(\phi)\| &= 2\left\|\left(m(g(\phi))' \otimes Q(g(\phi))'\right) \frac{\partial}{\partial \phi'} \operatorname{vec}\left(W(\phi)^{-1}\right)\right\| \\
&= 2\left\|\left(m(g(\phi))' \otimes Q(g(\phi))'\right)\left(W(\phi)^{-1} \otimes W(\phi)^{-1}\right) S(\phi)\right\| \\
&\leq 2\left(J(g(\phi),\phi)\left\|Q(g(\phi))'W(\phi)^{-1}Q(g(\phi))\right\|\left\|S(\phi)'\left(W(\phi)^{-1} \otimes W(\phi)^{-1}\right) S(\phi)\right\|\right)^{1/2} \\
&< C_3.
\end{aligned}$$

Combined with (13) we find

$$\left\|\frac{\partial}{\partial \phi'} g(\phi)\right\| \leq \left\|D(\phi)^{-1}\right\| \|B(\phi)\| < 1$$

This is (14) and implies the map $g(\phi)$ is a contraction. By the Banach fixed point theorem this implies that the fixed point θ_0 exists and is unique. ■

Proof of Theorem 3.1: Define the sample derivatives $\overline{Q}_n(\theta) = \frac{\partial}{\partial \theta'}\overline{m}_n(\theta)$, $\overline{R}_n(\theta) = \frac{\partial}{\partial \theta'}\operatorname{vec}(\overline{Q}_n(\theta)')$, and $\overline{S}_n(\phi) = \frac{\partial}{\partial \phi'}\operatorname{vec}\overline{W}_n(\phi)$. Our first task is to show that the sample averages $\overline{m}_n(\theta)$, $\overline{Q}_n(\theta)$, $\overline{R}_n(\theta)$, $\overline{W}_n(\theta)$, and $\overline{S}_n(\theta)$ converge in probability uniformly to their population expectations. Each is a sample average of the form $\frac{1}{n}\sum_{i=1}^n q\left(X_i,\theta\right)$ where the $q\left(X_i,\theta\right)$ are independent under Assumption 2.1, uniformly integrable under Assumption 2.2, and have a bounded first derivative under Assumptions 2.3 and 2.4. Thus each satisfies the uniform weak law of large numbers over $\theta \in \Theta$ as required. For the ULLN see D. Andrews (1992, Theorem 3).

The uniform convergence of $\overline{m}_n(\theta)$ and $\overline{W}_n(\phi)$, plus the uniform invertibility of $W(\phi)$ from Assumption 1.5 imply

$$\sup_{\phi,\theta}\left\|\overline{J}_n(\theta,\phi) - J(\theta,\phi)\right\| \xrightarrow[p]{} 0. \tag{37}$$

Since $g(\phi)$ minimizes $J(\theta,\phi)$ and $\overline{g}_n(\phi)$ minimizes $\overline{J}_n(\theta,\phi)$

$$\begin{aligned}
0 &\leq J(\overline{g}_n(\phi),\phi) - J(g(\phi),\phi) \\
&= J(\overline{g}_n(\phi),\phi) - \overline{J}_n(\overline{g}_n(\phi),\phi) - J(g(\phi),\phi) + \overline{J}_n(\overline{g}_n(\phi),\phi) \\
&\leq J(\overline{g}_n(\phi),\phi) - \overline{J}_n(\overline{g}_n(\phi),\phi) - J(g(\phi),\phi) + \overline{J}_n(g(\phi),\phi) \\
&\leq 2\sup_{\phi,\theta}\left\|\overline{J}_n(\theta,\phi) - J(\theta,\phi)\right\| \xrightarrow[p]{} 0,
\end{aligned}$$

where the final convergence by (37). This implies

$$\sup_{\phi}\left|J(\overline{g}_n(\phi),\phi) - J(g(\phi),\phi)\right| \xrightarrow[p]{} 0.$$

Fix $\epsilon > 0$. Under Assumption 1.2, $g(\phi)$ uniquely minimizes $J(\theta, \phi)$. The latter is continuous as a by-product of the proof of (37), so we can find a $\eta > 0$ such that for all ϕ, $\|g(\phi) - \theta\| > \epsilon$ implies $|J(g(\phi), \phi) - J(\theta, \phi)| > \eta$. Thus

$$\sup_{\phi} |J(g(\phi), \phi) - J(\overline{g}_n(\phi), \phi)| \leq \eta$$

implies $\sup_{\phi} \|g_n(\phi) - \overline{g}_n(\phi)\| \leq \epsilon$. Hence

$$P\left(\sup_{\phi} \|g(\phi) - \overline{g}_n(\phi)\| \leq \epsilon\right) \geq P\left(\sup_{\phi} |J(g(\phi), \phi) - J(\overline{g}_n(\phi), \phi)| \leq \eta\right) \to 1$$

as asserted. ■

Proof of Theorem 3.2: The fixed point $\widehat{\theta}$ exists and is unique if $\overline{g}_n(\phi)$ is a contraction mapping, in the sense that there is a $0 \leq c < 1$ such that

$$\|\overline{g}_n(\phi_1) - \overline{g}_n(\phi_2)\| \leq c \|\phi_1 - \phi_2\| \tag{38}$$

for all $\phi_1, \phi_2 \in \Theta$. Dominitz and Sherman (2005) Lemma 3 show that sufficient conditions for (38) to hold with probability tending to one as $n \to \infty$ are that (i) $g(\phi)$ is a contraction mapping (established in Theorem 2); (ii) $\sup_{\phi} \|\overline{g}_n(\phi) - g(\phi)\| \to_p 0$ (established in part 1); and (iii) $\sup_{\phi} \left\| \frac{\partial}{\partial \phi'} \overline{g}_n(\phi) - \frac{\partial}{\partial \phi'} g(\phi) \right\| \to_p 0$. Hence it is sufficient to verify this final condition.

Recall that $\frac{\partial}{\partial \phi'} g(\phi)$ can be expressed as (35) where $B(\theta, \phi)$ equals (36). We calculate that

$$D(\theta, \phi) = 2 \left\{ Q(\theta)' W(\phi)^{-1} Q(\theta) + \left(m(\theta)' W(\phi)^{-1} \otimes I \right) R(\theta) \right\}$$

and

$$\frac{\partial}{\partial \phi'} \overline{g}_n(\phi) = -\overline{D}_n(\overline{g}_n(\phi), \phi)^{-1} \overline{B}_n(\overline{g}_n(\phi), \phi)$$

where

$$\overline{B}_n(\theta, \phi) = -2 \left[\overline{m}_n(\theta)' \otimes \overline{Q}_n(\theta)' \right] \left[\overline{W}_n(\phi)^{-1} \otimes \overline{W}_n(\phi)^{-1} \right] \overline{S}_n(\phi)$$

and

$$\overline{D}_n(\theta, \phi) = 2 \left\{ \overline{Q}_n(\theta)' \overline{W}_n(\phi)^{-1} \overline{Q}_n(\theta) + \left(\overline{m}_n(\theta)' \overline{W}_n(\phi)^{-1} \otimes I \right) \overline{R}_n(\theta) \right\}.$$

Earlier we demonstrated that $\overline{m}_n(\theta)$, $\overline{Q}_n(\theta)$, $\overline{R}_n(\theta)$, $\overline{W}_n(\theta)$, and $\overline{S}_n(\theta)$ satisfy the ULLN. The continuous mapping theorem implies that $\overline{B}_n(\theta, \phi) - B(\theta, \phi)$ and $\overline{D}_n(\theta, \phi) - D(\theta, \phi)$ converge uniformly, and thus $\frac{\partial}{\partial \phi'} \overline{g}_n(\phi) - \frac{\partial}{\partial \phi'} g(\phi)$ as well. This completes the proof. ■

Proof of Theorem 3.3: Dominitz and Sherman (2005, Theorem 2) show that if $s(n) \to \infty$ then $\left\| \widehat{\theta}_{s(n)} - \theta_0 \right\| \underset{p}{\longrightarrow} 0$ since $g(\phi)$ is a contraction mapping (Theorem 2) and $\sup_{\phi} \|\overline{g}_n(\phi) - g(\phi)\| \underset{p}{\longrightarrow} 0$

(Theorem 3.1). Combined with Theorem 3.2 we find

$$\left\|\widehat{\theta} - \theta_0\right\| \leq \left\|\widehat{\theta}_{s(n)} - \theta_0\right\| + \left\|\widehat{\theta} - \widehat{\theta}_{s(n)}\right\| \underset{p}{\longrightarrow} 0.$$

■

Proof of Theorem 4: We can write $F = Q'W^{-1}m$ using the alternative representations

$$\begin{aligned} F &= \left(m'W^{-1} \otimes I_k\right) \operatorname{vec} Q' \\ &= \left(m' \otimes Q'\right) \operatorname{vec} W^{-1} \end{aligned}$$

and recall the identity

$$\frac{\partial}{\partial \theta'} \operatorname{vec} W^{-1} = -\left(W^{-1} \otimes W^{-1}\right) \frac{\partial}{\partial \theta'} \operatorname{vec} W.$$

An application of the chain rule yields (19) in the text. Similarly, defining

$$F(\theta) = Q(\theta)'W(\theta)^{-1}m(\theta)$$

we calculate that its derivative equals

$$\begin{aligned} \frac{\partial}{\partial \theta'} F(\theta) &= Q(\theta)'W(\theta)^{-1}Q(\theta) + \left(m(\theta)'W(\theta)^{-1} \otimes I_k\right) R(\theta) \\ &\quad - \left(m(\theta)'W(\theta)^{-1} \otimes Q(\theta)'W(\theta)^{-1}\right) S(\theta) \\ &\equiv H(\theta). \end{aligned}$$

Notice that the first-order condition for the estimator satisfies $\overline{F}_n(\widehat{\theta}) = 0$ and that for the pseudo-true value satisfies $F(\theta_0) = 0$. They satisfy the expansion

$$0 = \overline{F}_n(\widehat{\theta}) = \overline{F}_n(\theta_0) + H_n^* \left(\widehat{\theta} - \theta_0\right)$$

where the j^{th} row of H_n^* is the j^{th} row of $\overline{H}_n(\theta_{nj})$ and θ_{nj} is on the line segment joining $\widehat{\theta}$ and θ_0. Using the square root matrix $V^{-1/2} = \Omega^{-1/2}H$ we find

$$\sqrt{n}V^{-1/2}\left(\widehat{\theta} - \theta_0\right) = -\Omega^{-1/2}HH_n^{*-1}\sqrt{n}\,\overline{F}_n(\theta_0).$$

The Theorem follows from the two limit results:

$$\left\|HH_n^{*-1} - I_k\right\| \underset{p}{\longrightarrow} 0 \tag{39}$$

$$\sqrt{n}\,\overline{F}_n(\theta_0) \underset{d}{\longrightarrow} N\left(0, I_k\right). \tag{40}$$

Take (39). In the proof of Theorem 3.1 we showed that $\overline{m}_n(\theta)$, $\overline{Q}_n(\theta)$, $\overline{R}_n(\theta)$, $\overline{W}_n(\theta)$, and $\overline{S}_n(\theta)$ satisfy the ULLN over $\theta \in \Theta$. Since $\overline{H}_n(\theta)$ is a continuous function of these moments it converges

uniformly as well. Together with $\left\|\widehat{\theta}-\theta_0\right\| \underset{p}{\longrightarrow} 0$ and $\left\|H^{-1}\right\| = \sqrt{1/\lambda_{\min}\left(H'H\right)} \leq C^{-1/2} < \infty$ we obtain

$$\left\|H^{-1/2}H_n^*H^{-1/2\prime} - I_k\right\| = \left\|H_n^* - H\right\| \left\|H^{-1}\right\| \underset{p}{\longrightarrow} 0.$$

This implies that the eigenvalues of $H^{-1/2}H_n^*H^{-1/2\prime}$ converge in probability to 1, which implies the same for $H^{1/2\prime}H_n^{*-1}H^{1/2}$ and the singular values of HH_n^{*-1}. This implies (39).

Take (40). Assumption 3.1 and independence of the observations implies that $\sqrt{n}\left(\overline{m}_n - \mu\right)$, $\sqrt{n}\left(\overline{Q}_n - Q\right)$ and $\sqrt{n}\left(\overline{W}_n - W\right)$ are $O_p(1)$. Assumption 1.3 implies $W^{-1} > 0$, so $\sqrt{n}\left(\overline{W}_n^{-1} - W^{-1}\right) = O_p(1)$ as well. Then by standard expansions and $0 = Q'W^{-1}\mu$ we find

$$\begin{aligned}
\sqrt{n}\,\overline{F}_n(\theta_0) =& \sqrt{n}\,\overline{Q}_n'\overline{W}_n^{-1}\overline{m}_n \\
=& \sqrt{n}\left(Q'W^{-1}\overline{m}_n + \overline{Q}_n'W^{-1}\mu - Q'W^{-1}\overline{W}_nW^{-1}\mu\right) + o_p(1) \\
=& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i + o_p(1)
\end{aligned}$$

where ψ_i is defined in (21). Since ψ_i are independent across i, mean zero (since $Q'W^{-1}\mu = 0$), uniformly square integrable (Assumption 3.1), and $\lambda_{\min}(\Omega) \geq C > 0$ (Assumption 3.3), an application of the multivariate CLT for heterogeneous random vectors establishes (40). ■

Proof of Theorem 5: Given (39) it is sufficient to show that

$$\left\|\widehat{\Omega} - \Omega\right\| \underset{p}{\longrightarrow} 0. \tag{41}$$

We can write $\psi_i = D'f_i(\theta_0)$ and $\widehat{\psi}_i = \widehat{D}'f_i(\widehat{\theta})$ where

$$D = \begin{bmatrix} W^{-1}Q \\ W^{-1}\mu \otimes I_k \\ -W^{-1}\mu \otimes W^{-1}Q \end{bmatrix}$$

and

$$f_i(\theta) = \begin{bmatrix} m(X_i,\theta) \\ \text{vec}(Q(X_i,\theta)') \\ v(X_i,\theta) \otimes v(X_i,\theta) \end{bmatrix},$$

and $\widehat{D}$ is defined as the sample version of D. We can write $\widehat{\Omega} = \widehat{D}'\widetilde{G}(\widehat{\theta})\widehat{D}$ where

$$\widetilde{G}(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)f_i(\theta)'.$$

It is straightforward to establish that $\widehat{D} - D \underset{p}{\longrightarrow} 0$. Equation (41) and the theorem will follow if $\widetilde{G}(\theta)$ satisfies the ULLN in a neighborhood of θ_0. $f_i(\theta)f_i(\theta)'$ is uniformly integrable under

Assumption 3.1. It satisfies the derivative bound

$$\sup_i E\left|\frac{\partial}{\partial\theta}\left\|f_i(\theta)f_i(\theta)'\right\|\right| \leq 2\sqrt{\sup_i E\left(\sup_{\theta\in\mathcal{N}}\|f_i(\theta)\|^2\right)\sup_i E\left(\sup_{\theta\in\mathcal{N}}\left\|\frac{\partial}{\partial\theta'}f_i(\theta)\right\|^2\right)} < \infty$$

under Assumptions 3.1 and 4. Thus $\widetilde{G}(\theta)$ satisfies the ULLN as required and the proof is complete. ■

References

1. Acemoglu, Daron, Simon Johnson, James A. Robinson, and Pierre Yared (2008). Income and democracy. *American Economic Review*, 98(3), 808-842.

2. Acemoglu, Daron, Simon Johnson, James A. Robinson, and Pierre Yared (2009). Reevaluating the modernization hypothesis. *Journal of Monetary Economics*, 56(8), 1043-1058.

3. Aguirre-Torres, Víctor, and Manuel Domínguez Toribio (2004). Efficient method of moments in misspecified iid models. *Econometric Theory*, 20(3), 513-534.

4. Ai, Chunrong and Xiaohong Chen (2007). Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables. *Journal of Econometrics*, 141(1), 5-43.

5. Andrews, Donald W. K. (1992). Generic uniform convergence. *Econometric Theory*, 8(2), 241-257.

6. Andrews, Isaiah (2019). On the structure of IV estimands. *Journal of Econometrics*, 211(1), 294-307.

7. Angrist, Joshua D., Victor Chernozhukov, and Iván Fernández-Val (2006). Quantile regression under misspecification, with an application to the U.S. wage structure. *Econometrica*, 74, 539-563.

8. Angrist, Joshua D. and Guido W. Imbens (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90:430, 431-442

9. Arellano, Manuel, and Stephen Bond (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies*, 58(2), 277-297.

10. Bonhomme, Stephane, and Elena Manresa (2015). Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3), 1147-1184.

11. Brinch, Christian N., Magne Mogstad, and Matthew Wiswall (2017). Beyond LATE with a discrete instrument. *Journal of Political Economy*, 125(4), 985-1039.

12. Cervellati, Matteo, Florian Jung, Uwe Sunde, and Thomas Vischer (2014). Income and democracy: Comment. *American Economic Review*, 104(2), 707-719.

13. Cheng, Xu, Zhipeng Liao, and Ruoyao Shi (2019). On uniform asymptotic risk of averaging GMM estimators. *Quantitative Economics*, 10(3), 931-979.

14. Corcoran, Stephen A (1998). Bartlett adjustment of empirical discrepancy statistics. *Biometrika*, 85, 967-972.

15. Dominitz, Jeff, and Robert P. Sherman (2005). Some convergence theory for iterative estimation procedures with an application to semiparametric estimation. *Econometric Theory*, 21 (04), 838-863.

16. Dovonon, Prosper (2016). Large Sample Properties of the Three-Step Euclidean Likelihood Estimators under Model Misspecification. *Econometric Reviews*, 35(4), 465-514.

17. Evdokimov, Kirill, and Michal Kolesár (2018). Inference in Instrumental Variable Regression Analysis with Heterogeneous Treatment Effects. Working paper.

18. Gallant, A. Ronald, and Halbert White (1988). *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. Blackwell.

19. Goldberger, Arthur S. (1991). *A Course in Econometrics*. Harvard University Press.

20. Gospodinov, Nikolay, Raymond Kan, and Cesare Robotti (2014). Misspecification-robust inference in linear asset-pricing models with irrelevant risk factors. *The Review of Financial Studies*, 27(7), 2139-2170.

21. Hall, Alastair R. (2005). *Generalized Method of Moments*. Oxford University Press.

22. Hall, Alastair R., and Atsushi Inoue (2003). The large sample behaviour of the generalized method of moments estimator in misspecified models. *Journal of Econometrics*, 114(2), 361-394.

23. Hansen, Bruce E. and Seojeong Lee (2019). Asymptotic theory for clustered samples. *Journal of Econometrics*, 210(2), 268-290.

24. Hansen, Bruce E. and Seojeong Lee (2020). Inference for GMM under misspecification and clustering.

25. Hansen, Lars Peter (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50, 1029-1054.

26. Hansen, Lars Peter, John Heaton, and Amir Yaron (1996). Finite-Sample properties of some alternative GMM estimators. *Journal of Business & Economic Statistics*, 14, 262-280.

27. Hansen, Lars Peter, and Ravi Jagannathan (1997). Assessing specification errors in stochastic discount factor models. *The Journal of Finance*, 52(2), 557-590.

28. Hansen, Lars Peter and Thomas J. Sargent (2008). *Robustness*. Princeton University Press.

29. Heckman, James J., Sergio Urzua, and Edward Vytlacil (2006). Understanding instrumental variables in models with essential heterogeneity. *The Review of Economics and Statistics*, 88(3), 389-432.

30. Heckman, James J., and Edward Vytlacil (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, 73(3), 669-738.

31. Hwang, Jungbin, Byunghoon Kang, and Seojeong Lee (2020). A doubly corrected robust variance estimator for linear GMM. *Journal of Econometrics*, forthcoming.

32. Imbens, Guido W (1997). One-step estimators for over-identified generalized method of moments models. *The Review of Economic Studies*, 64(3), 359-383.

33. Imbens, Guido W. and Joshua D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62, 467-475.

34. Imbens, Guido W., Richard H. Spady, and Phillip Johnson (1998). Information theoretic approaches to inference in moment condition models. *Econometrica*, 66(2), 333-358.

35. Kan, Raymond, and Cesare Robotti (2008). Model comparison using the Hansen-Jagannathan distance. *The Review of Financial Studies*, 22(9), 3449-3490.

36. Kitamura, Yuichi, and Michael Stutzer (1997). An information-theoretic alternative to generalized method of moments estimation. *Econometrica*, 65(4), 861-874.

37. Kolesár, Michal (2013). Estimation in an instrumental variables model with treatment effect heterogeneity. Working paper.

38. Lee, Seojeong (2014). Asymptotic refinements of a misspecification-robust bootstrap for generalized method of moments estimators. *Journal of Econometrics*, 178(3), 398-413.

39. Lee, Seojeong (2016). Asymptotic refinements of a misspecification-robust bootstrap for GEL estimators. *Journal of Econometrics*, 192(1), 86-104.

40. Lee, Seojeong (2018). A consistent variance estimator for 2SLS when instruments identify different LATEs. *Journal of Business & Economic Statistics*, 36(3), 400-410.

41. Lewbel, Arthur, and Krishna Pendakur (2009). Tricks with Hicks: The EASI demand system. *American Economic Review*, 99(3), 827-63.

42. Lu, Xun and Liangjun Su (2017). Determining the number of groups in latent panel structures with an application to income and democracy. *Quantitative Economics*, 8, 729-760.

43. Maasoumi, Esfandiar, and Peter C. B. Phillips (1982). On the behavior of inconsistent instrumental variable estimators. *Journal of Econometrics*, 19(2), 183-201.

44. Mogstad, Magne, Andres Santos, and Alexander Torgovitsky (2018). Using instrumental variables for inference about policy relevant treatment parameters. *Econometrica*, 86(5), 1589-1619.

45. Newey, Whitney K. and Richard J. Smith (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72(1), 219-255.

46. Owen, Art B (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2), 237-249.

47. Qin, Jin, and Jerry Lawless (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22(1), 300-325.

48. Schennach, Susanne M. (2007). Point estimation with exponentially tilted empirical likelihood. *The Annals of Statistics*, 35(2), 634-672.

49. Słoczyński, Tymon (2018). A general weighted average representation of the ordinary and two-stage least squares estimands. Woking paper.

50. White, Halbert (1980a). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4), 817-838.

51. White, Halbert (1980b). Using least squares to approximate unknown regression functions. *International Economic Review*, 149-170.

52. White, Halbert (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1), 1-25.

53. White, Halbert (1984). *Asymptotic Theory for Econometricians*. Academic Press.

54. White, Halbert (1994). *Estimation, Inference and Specification Analysis*. Cambridge University Press.

55. Windmeijer, Frank (2000). A finite sample correction for the variance of linear two-step GMM estimators (No. W00/19). Working paper. Institute for Fiscal Studies.

56. Windmeijer, Frank (2005). A finite sample correction for the variance of linear efficient two-step GMM estimators. *Journal of Econometrics* 126(1), 25-51.