Inference for Iterated GMM Under Misspecification and Clustering

Bruce E. Hansen*
University of Wisconsin

Seojeong Lee†
University of New South Wales

April 2018

Abstract

This paper develops a new distribution theory and inference methods for over-identified Generalized Method of Moments (GMM) estimation focusing on the iterated GMM estimator, allowing for moment misspecification, and for clustered dependence with heterogeneous and growing cluster sizes. This paper is the first to provide a rigorous theory for the iterated GMM estimator. We provide conditions for its existence by demonstrating that the iteration sequence is a contraction mapping. Our asymptotic theory allows the moments to be possibly misspecified, which is a general feature of approximate over-identified models. This form of moment misspecification causes bias in conventional standard error estimation. Our results show how to correct for this standard error bias. Our paper is also the first to provide a rigorous distribution theory for the GMM estimator under cluster dependence. Our distribution theory is asymptotic, and allows for heterogeneous and growing cluster sizes. Our results cover standard smooth moment condition models, including dynamic panels, which is a common application for GMM with cluster dependence. Our simulation results show that conventional heteroskedasticity-robust standard errors are highly biased under moment misspecification, severely understating estimation uncertainty, and resulting in severely over-sized hypothesis tests. In contrast, our misspecification-robust standard errors are approximately unbiased and properly sized under both correct specification and misspecification. We illustrate the method by extending the empirical work reported in Acemoglu, Johnson, Robinson, and Yared (2008, *American Economic Review*) and Cervellati, Jung, Sunde, and Vischer (2014, *American Economic Review*). Our results reveal an enormous effect of iterating the GMM estimator, demonstrating the arbitrariness of using one-step and two-step estimators. Our results also show a large effect of using misspecification robust standard errors instead of the Arellano-Bond standard errors. Our results support Acemoglu, Johnson, Robinson, and Yared's conclusion of an insignificant effect of income on democracy, but reveal that the heterogeneous effects documented by Cervellati, Jung, Sunde, and Vischer are less statistically significant than previously claimed.

*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 Reserach Award (DECRA) funding scheme (project number DE170100787).

1 Introduction

White (1980, 1982) advocated for robust inference, meaning that variance estimation should be constructed to 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 the pseudo-true parameter values without the requirement of 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 estimation by maximum likelihood could be quite sensitive to model misspecification. GMM focused estimation on the specific moment conditions specified by the application. Hansen's proposed covariance matrix estimators were also quite similar to those of White (1980) in that they did not exploit information beyond the moment conditions used for estimation.

However, when the model is over-identified Hansen's GMM covariance matrix estimator turns out to be quite sensitive to the assumption of correct moment specification. If we take the realistic view that an over-identified model is a constructive approximation rather than a literal truth, we should be cautious about requiring that our inference procedures rely on the literal assumption of correct specification.

This concern for robustness is echoed in the monograph by Hansen and Sargent (2008), where they argue that decisions should be robust to model misspecification.

This paper focuses on the problem of correct asymptotic inference in over-identified econometric models without requiring that all moment conditions hold exactly in the population. In this context it turns out that correct GMM inference requires a significant adjustment in covariance matrix calculation, as the asymptotic distribution turns out to depend on estimation error in the moment derivatives, on weight matrix estimation, and the degree of curvature of the model moments. Fortunately it is straightforward to characterize the correct covariance matrix structure, though some of the calculations are more tedious than the conventional case.

A second issue raised in this paper is a rigorous theory for the iterated GMM estimator. 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. In our empirical application we demonstrate that estimators can be highly sensitive to the initial weight matrix and the number of iterations. We provide a rigorous theory by providing simple conditions under which the iteration sequence is a contraction and thus the iterated GMM estimator exists. This results benefits from the theory of Dominitz and Sherman (2005). To our knowledge our paper is the first to provide this demonstration.

A third issue raised in this paper is inference allowing for clustered sampling dependence. In the past two decades there has been an explosion of empirical econometric interest in clustered sampling, but relatively little formal theory. This paper uses a new asymptotic theory developed in a companion paper Hansen and Lee (2017), which allows for quite general forms of clustered dependence, allowing for heterogeneous and growing cluster sizes. Our theory requires the number

of clusters to diverge to infinity (so-called "large G" asymptotics) so to obtain asymptotically normal limiting representations.

This paper builds on the important contribution of Hall and Inoue (2003) who similarly explored the asymptotic distribution of the GMM estimator under moment misspecification. A limitation of their analysis was that they were unable to incorporate the iterated GMM estimator, and thus had the unfortunate finding that the limiting distribution depended on the specific weight matrix. Instead, our focus on the iterated GMM estimator simplifies the analysis by removing the dependence on the specified iteration.

One of the most common applications of over-identified GMM with clustered dependence is the dynamic panel model. The standard estimators, due to Arellano and Bond (1991), Arellano and Bover (1995) and Blundell and Bond (1998) all fall in this class, and are covered by our assumptions. 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 the models should generically be viewed as constructive approximations, and standard errors calculated using our misspecification-robust approach.

The assumptions in this paper are closely related to the context of over-identified IV regression with heterogeneous treatment effects (Imbens and Angrist (1994), Angrist and Imbens (1995), Kolesár (2013)). As shown in Lee (2017) and Evdokimov and Kolesár (2017), 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.

Our results assume that the moment conditions are smooth. Allowing for non-differentiable moment conditions would be desirable but would require a different approach.

The iterated GMM estimator is related to – but substantially different from – the continuously updated estimator of Hansen, Heaton and Yaron (1996). While it would be desirable to extend our results to cover the CU-GMM estimator, we do not do so in this paper to keep the presentation focused. Such an extension would be technically much more complicated.

The organization of the paper is as follows. Section 2 presents the iterated GMM estimator. Section 3 provides formal conditions for identification and existence. Section 4 discusses weight matrices. Section 5 presents the asymptotic distribution of the GMM estimator. Section 6 discusses covariance matrix estimation. Section 7 shows that the iterated GMM estimator defined with the efficient weight matrix is invariant to the weight matrix being constructed with re-centering. Section 8 discusses the GMM test of over-identifying restrictions. Section 9 describes the results for the linear model. Section 10 presents simulation evidence of the finite sample distributions. Section 11 is an application examining the dynamic panel regressions in Acemoglu, Johnson, Robinson, and Yared (2008) and Cervellati, Jung, Sunde, and Vischer (2014). Formal proofs are presented in the Appendix.

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

2 Generalized Method of Moments Estimation

Consider a standard over-identified moment condition model which specifies that

$$E\left(m(X_i,\theta)\right)=0 \tag{1}$$

where $m(\cdot,\cdot)$ is $l\times 1$ and $\theta\in\Theta$ is $k\times 1$ 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 estimate of (1).

The parameter θ is estimated by iterated GMM. Since the model is over-identified, the moment condition is augmented by an $l\times l$ positive definite user-specified weight matrix $\overline{W}_n(\theta)$ which possibly depends on the parameter vector θ. Given an initial value $\widehat{\theta}_0$ we create a sequence $\widehat{\theta}_s$ by iterative minimization

$$\widehat{\theta}_s=\underset{\theta\in\Theta}{\arg\min}\,\overline{m}_n(\theta)'\overline{W}_n(\widehat{\theta}_{s-1})^{-1}\overline{m}_n(\theta).$$

$\widehat{\theta}_s$ is known as the *s-step GMM estimator*. If the sequence is iterated until convergence we obtain the *iterated GMM estimator*:

$$\widehat{\theta}=\lim_{s\to\infty}\widehat{\theta}_s. \tag{2}$$

If the weight matrix $\overline{W}_n$ does not depend on θ then $\widehat{\theta}_s=\widehat{\theta}$ but they differ otherwise. We discuss in Section 3 sufficient conditions such that the limit in (2) exists.

Alternatively, we can view (2) as a fixed point. Define the mapping

$$\overline{g}_n(\phi)=\underset{\theta\in\Theta}{\arg\min}\,\overline{m}_n(\theta)'\overline{W}_n(\phi)^{-1}\overline{m}_n(\theta). \tag{3}$$

Given this notation, the iteration sequence can be written as

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

and the iterated GMM estimator (2) is the fixed point of the equation

$$\overline{g}_n(\widehat{\theta})=\widehat{\theta}. \tag{4}$$

3 Identification and Existence

Our goal is inference on θ allowing for robustness to possible moment misspecification. By this we mean that there may not exist a value θ solving (1). Following White (1982) it is appropriate in this context to define the *pseudo-true* parameter value θ_n as the vector which solves the population analog of the estimation problem. In an over-identified model this means the pseudo-true value

will depend on the weight matrix, as discussed in Hall and Inoue (2003).

Define the population analogs of the sample moment and weight matrix

$$m_n(\theta) = E\left(\overline{m}_n(\theta)\right) \tag{5}$$

$$W_n(\theta) = E\left(\overline{W}_n(\theta)\right). \tag{6}$$

Notice that we write the expectations (5) and (6) as functions of n. This allows heterogeneous distributions, and additionally under cluster sampling with non-homogeneous cluster sizes the weight matrix (6) is likely to vary with n. Under i.i.d. sampling the n subscripts can be omitted.

We then define the population analog of (3):

$$g_n(\phi) = \arg\min_{\theta\in\Theta} m_n(\theta)' W_n(\phi)^{-1} m_n(\theta). \tag{7}$$

Definition (7) specifies $g_n(\phi)$ as the best fitting value of θ given the weight matrix $W_n(\phi)$ and an initial value ϕ. Under correct specification so that (1) holds for some θ_n, and if $W_n(\phi) > 0$, it follows that the solution $g_n(\phi) = \theta_n$ is unique. Under moment misspecification, however, the solution (7) may vary with ϕ.

As an analog of the iterated GMM estimator we define the population *pseudo-true* value θ_n to be the fixed point of the population mapping $g_n(\phi)$. This solves

$$g_n(\theta_n) = \theta_n \tag{8}$$

Conceptually, one could imagine obtaining θ_n by iterating $g_n(\phi)$ from a starting point until convergence. We write the pseudo-true value θ_n as a function of the sample size n since the population weight matrix (6) may vary with the sample under cluster sampling.

The existence of the fixed points of (4) and (8) have not been discussed in the previous literature. We now provide formal justifications.

Define the population criterion $J_n(\theta,\phi) = m_n(\theta)' W_n(\phi)^{-1} m_n(\theta)$ and $D_n(\theta,\phi) = \frac{\partial^2}{\partial\theta\partial\theta'} J_n(\theta,\phi)$. For a vector a let $\|a\| = (a'a)^{1/2}$ denote the Euclidean norm. For a positive semi-definite matrix A let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote its smallest and largest eigenvalue, respectively. For a general matrix A let $\|A\| = \sqrt{\lambda_{\max}\left(A'A\right)}$ denote the spectral norm.

Assumption 1. *For some* $0 < C < \infty$

1. Θ *is compact*
2. $\inf_{\phi\in\Theta} \lambda_{\min}\left(W_n(\phi)\right) \geq C^{-1}$
3. $\inf_{\phi\in\Theta} \lambda_{\min}\left(D_n(g_n(\phi),\phi)\right) \geq C^{-1}$
4. $\sup_{\phi\in\Theta} \|m_n(g_n(\phi))\| \leq \delta$ *where* $\delta < \left(2kC^5\right)^{-1}$
5. $m(x,\theta)$ *is twice continuously differentiable in* $\theta \in \Theta$

6. $W_n(\phi)$ is continuously differentiable in $\phi \in \Theta$

Assumption 1.1 imposes compactness for technical convenience. Assumption 1.2 excludes singular population weight matrices. Assumption 1.3 is a global identification condition.

Assumption 1.4 is unusual. It specifies that the degree of misspecification is small, in the sense that the norm of the population moment $m_n(\theta)$ is small for all pseudo-true values of θ. This assumption is automatically satisfied under correct specification (since in that context $m_n(\theta_n) = 0$), but otherwise allows for mild moment misspecification. This assumption is only used to establish the existence of the pseudo-true value under misspecification, so could be replaced by any other sufficient condition for its existence.

Assumption 1.5 is a stronger smoothness condition than typical for GMM distribution theory, but is needed to allow for moment misspecification. Assumption 1.6 is a mild smoothness condition on the population weight matrix.

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

Theorem 1. *Under Assumption 1 the map $g_n(\phi)$ is a contraction. The fixed point θ_n exists and is unique.*

We next provide formal justification for existence of the iterated GMM estimator (2).

Define the sample derivatives $\overline{Q}_n(\theta) = \frac{\partial}{\partial \theta'}\overline{m}_n(\theta)$, $\overline{R}_n(\theta) = \frac{\partial}{\partial \theta'}\mathrm{vec}(\overline{Q}_n(\theta)')$, and $\overline{S}_n(\phi) = \frac{\partial}{\partial \phi'}\mathrm{vec}\overline{W}_n(\phi)$, and the population analogs $Q_n(\theta) = \frac{\partial}{\partial \theta'}m_n(\theta)$, $R_n(\theta) = \frac{\partial}{\partial \theta'}\mathrm{vec}(Q_n(\theta)')$, and $S_n(\phi) = \frac{\partial}{\partial \phi'}\mathrm{vec}W_n(\phi)$.

Assumption 2. *As $n \to \infty$*

$$\sup_{\theta\in\Theta} \|\overline{m}_n(\theta) - m_n(\theta)\| \xrightarrow{p} 0 \tag{9}$$

$$\sup_{\theta\in\Theta} \left\|\overline{Q}_n(\theta) - Q_n(\theta)\right\| \xrightarrow{p} 0 \tag{10}$$

$$\sup_{\theta\in\Theta} \left\|\overline{R}_n(\theta) - R_n(\theta)\right\| \xrightarrow{p} 0 \tag{11}$$

$$\sup_{\theta\in\Theta} \left\|\overline{W}_n(\theta) - W_n(\theta)\right\| \xrightarrow{p} 0 \tag{12}$$

$$\sup_{\theta\in\Theta} \left\|\overline{S}_n(\theta) - S_n(\theta)\right\| \xrightarrow{p} 0 \tag{13}$$

and the functions $m_n(\theta)$, $Q_n(\theta)$, $R_n(\theta)$, $W_n(\theta)$ and $S_n(\theta)$ are continuous in θ uniformly over $\theta \in \Theta$.

Assumption 2 states that the sample moments converge uniformly over θ to their expectations. Sufficient conditions for these results are available for specific sampling contexts. In Section 5 we provide primitive conditions in the context of cluster sampling.

Assumptions 1-2 are sufficient to establish the existence of the iterated GMM estimator $\widehat{\theta}$ and its consistency for θ_n.

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

1. $\sup_{\phi\in\Theta} \|\overline{g}_n(\phi) - g_n(\phi)\| \to_p 0$.

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. $\left\|\widehat{\theta} - \theta_n\right\| \to_p 0$.

Our proof of parts 2 and 3 of Theorem 2 builds on Dominitz and Sherman (2005, Theorem 2 and Lemma 3). They show that if the population mapping $g_n(\phi)$ is a contraction (which was established in Theorem 1), the sample mapping $\overline{g}_n(\phi)$ is uniformly consistent (established in part 1), and similarly its derivative, then $\overline{g}_n(\phi)$ is a contraction, the fixed point exists, and the fixed point $\widehat{\theta}$ is consistent. We use the uniform consistency of Assumption 2 to establish the uniform consistency of $\overline{g}_n(\phi)$ and its derivative.

4 Weight Matrices and Clustering

For the remainder of the paper we focus attention on two specific classes of weight matrices which encompass the typical choices used in empirical practice. In particular, we allow for either unclustered or clustered weight matrices.

We assume that the observations are grouped into G mutually independent known clusters, indexed $g = 1, ..., G$, where the g^{th} cluster has n_g observations. The number of observations per cluster may vary across clusters. Thus the total number of observations is $n = \sum_{g=1}^{G} n_g$. When convenient, we index the observations as X_{gj} for $g = 1, ..., G$ and $j = 1, ..., n_g$. Random sampling is the special case where $n_g = 1$.

Unclustered weight matrices take the form

$$\overline{W}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} v(X_i, \theta)v(X_i, \theta)', \tag{14}$$

for some $l \times 1$ vector $v(x, \theta)$. Unclustered weight matrices are standard under independent sampling but can also be used under clustered sampling. The two leading examples of unclustered weight matrices (14) are 2SLS, which sets $v(X_i, \theta) = Z_i$ for an instrument vector Z_i, and the standard efficient weight matrix which sets $v(x, \theta) = m(x, \theta)$.

Clustered weight matrices take the form

$$\begin{aligned}\overline{W}_n(\theta) &= \frac{1}{n}\sum_{g=1}^{G}\left(\sum_{j=1}^{n_g} v(X_{gj}, \theta)\right)\left(\sum_{j=1}^{n_g} v(X_{gj}, \theta)\right)' \\ &= \frac{1}{n}\sum_{g=1}^{G} \widetilde{v}_g(\theta)\widetilde{v}_g(\theta)',\end{aligned} \tag{15}$$

for some $l \times 1$ vector $v(x, \theta)$, and $\widetilde{v}_g(\theta) = \sum_{j=1}^{n_g} v(X_{gj}, \theta)$. Clustered weight matrices are often used under cluster sampling. The standard GMM estimator under clustering uses an efficient weight

matrix (15) with $v(x,\theta) = m(x,\theta)$. The Arellano and Bond (1991) two-step GMM estimator for dynamic panels uses (15) with $v((Y,X,Z),\theta) = Z'(Y - X\theta)$.

Even when the observations are clustered for the purpose of standard error calculation, it is not necessary for a user to select a clustered weight matrix. It is quite common, for example, for users to estimate a model by 2SLS (which uses a non-clustered weight matrix of the form (14)) and then use cluster-robust standard errors. Consequently, we allow for both (14) and (15) as feasible choices, separately from the choice of covariance matrix estimator.

The weight matrices (14) and (15) are uncentered. Centered versions are also commonly used. In Section 7 we show that our GMM estimators are invariant to recentering, and thus all our results apply to such weight matrices as well.

5 Asymptotic Distribution

The iterated GMM estimator $\widehat{\theta}$ minimizes the criterion $\overline{m}_n(\theta)'\overline{W}_n(\widehat{\theta})^{-1}\overline{m}_n(\theta)$ and thus satisfies the first-order condition

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

The standard approach to obtain the asymptotic distribution for $\widehat{\theta}$ makes a first-order Taylor expansion of $\overline{m}_n(\widehat{\theta})$ about $\overline{m}_n(\theta_n)$ and then solves to find

$$\sqrt{n}\left(\widehat{\theta} - \theta_n\right) \simeq -\left(\overline{Q}_n(\widehat{\theta})'\overline{W}_n(\widehat{\theta})^{-1}\overline{Q}_n(\theta_n)\right)^{-1}\overline{Q}_n(\widehat{\theta})'\overline{W}_n(\widehat{\theta})^{-1}\sqrt{n}\overline{m}_n(\theta_n).$$

Under correct specification $E\overline{m}_n(\theta_n) = 0$ so the central limit theorem applies. However, under misspecification $E\overline{m}_n(\theta_n) = \mu_n \neq 0$ and we cannot apply the central limit theorem without first re-centering $\overline{m}_n(\theta_n)$ about μ_n. This invalidates the above argument and 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 $\overline{m}_n(\widehat{\theta})$. There are three steps. The first expands the sample function $\overline{F}_n(\theta)$ about θ_n. To do so, its derivative equals

$$\begin{aligned}\frac{\partial}{\partial\theta'}\overline{F}_n(\theta) &= \overline{Q}_n(\theta)'\overline{W}_n(\theta)^{-1}\overline{Q}_n(\theta) + \left(\overline{m}_n(\theta)'\overline{W}_n(\theta)^{-1}\otimes I_k\right)\overline{R}_n(\theta) \\ &\quad - \left(\overline{m}_n(\theta)'\overline{W}_n(\theta)^{-1}\otimes \overline{Q}_n(\theta)'\overline{W}_n(\theta)^{-1}\right)\overline{S}_n(\theta) \\ &\equiv \overline{H}_n(\theta).\end{aligned} \tag{16}$$

(This and other calculations are justified in the appendix.) Expanding $\overline{F}_n(\widehat{\theta})$ about θ_n, we find that

$$0 = \overline{F}_n(\widehat{\theta}) \simeq \overline{F}_n(\theta_n) + \overline{H}_n(\theta_n)\left(\widehat{\theta} - \theta_n\right). \tag{17}$$

Thus

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

Second, we expand $\overline{F}_n(\theta_n)$ in terms of sample moments. Set $\mu_n = m_n(\theta_n)$, $Q_n = Q_n(\theta_n)$, $W_n = W_n(\theta_n)$, $R_n = R_n(\theta_n)$, and $S_n = S_n(\theta_n)$. Set $\overline{m}_n = \overline{m}_n(\theta_n)$, $\overline{Q}_n = \overline{Q}_n(\theta_n)$, and $\overline{W}_n = \overline{W}_n(\theta_n)$. In the Appendix we show that

$$\sqrt{n}\,\overline{F}_n(\theta_n) = \sqrt{n}\widetilde{F}_n\,(1 + o_p(1)) \tag{19}$$

where

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

We write $\widetilde{F}_n$ as a sum across the cluster sums. Define

$$\begin{aligned}
\widetilde{m}_g &= \sum_{j=1}^{n_g} m(X_{gj}, \theta_n) \\
\widetilde{Q}_g &= \sum_{j=1}^{n_g} Q(X_{gj}, \theta_n) \\
\widetilde{W}_g &= \begin{cases} \sum_{j=1}^{n_g} v(X_{gj}, \theta_n) v(X_{gj}, \theta_n)' & \text{under (14)} \\ \widetilde{v}_g(\theta_n)\widetilde{v}_g(\theta_n)' & \text{under (15).} \end{cases}
\end{aligned}$$

Then set

$$\widetilde{\psi}_g = Q_n' W_n^{-1}\widetilde{m}_g + \widetilde{Q}_g W_n^{-1}\mu_n - Q_n' W_n^{-1}\widetilde{W}_g W_n^{-1}\mu_n$$

so that

$$\sqrt{n}\widetilde{F}_n = \frac{1}{\sqrt{n}}\sum_{g=1}^{G}\widetilde{\psi}_g. \tag{20}$$

The CLT can be applied to (20), which has variance

$$\Omega_n = \frac{1}{n}\sum_{g=1}^{G} E\left(\widetilde{\psi}_g\widetilde{\psi}_g'\right). \tag{21}$$

Equations (18), (19), and (20) imply

$$\sqrt{n}\left(\widehat{\theta} - \theta_n\right) \simeq -\overline{H}_n(\theta_n)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{g=1}^{G}\widetilde{\psi}_g\right)(1 + o_p(1)).$$

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

We now provide regularity conditions and a formal statement. Define

$$H_n = H_n(\theta_n) = Q_n' W_n^{-1} Q_n + \left(\mu_n' W_n^{-1} \otimes I_k\right) R_n - \left(\mu_n' W_n^{-1} \otimes Q_n' W_n^{-1}\right) S_n \tag{22}$$

and $U(x, \theta) = \frac{\partial}{\partial \theta'} v(x, \theta)$.

Assumption 3. *For some* $0 < C < \infty$

1. $\lambda_{\min}(H_n) \geq C^{-1}$

2. $\lambda_{\min}(\Omega_n) \geq C^{-1}$

3. *The weight matrix* $\overline{W}_n(\theta)$ *takes either the unclustered (14) or clustered form (15)*

4. *For each* $\theta \in \Theta$, *and* $f(x) = \|m(x,\theta)\|^2$, $\|Q(x,\theta)\|^2$, $\|R(x,\theta)\|$, $\|v(x,\theta)\|^4$, *and* $\|U(x,\theta)\|^2$

$$\lim_{B\to\infty} \sup_i E\left(\|f(X_i)\| 1\left(\|f(X_i))\| > B\right)\right) = 0$$

5. *For each* $\theta_1, \theta_2 \in \Theta$ *and* $f(x,\theta) = m(x,\theta)$, $Q(x,\theta)$, $R(x,\theta)$, $v(x,\theta)$, *and* $U(x,\theta)$

$$\|f(x,\theta_1) - f(x,\theta_2)\| \leq A(x)h\left(\|\theta_1 - \theta_2\|\right)$$

where $h(u) \downarrow 0$ *as* $u \downarrow 0$, *and (i)* $\sup_i EA(X_i) \leq C$ *for* $f(x,\theta) = R(x,\theta)$, *and (ii)* $\sup_i EA(X_i)^2 \leq C$ *for* $f(x,\theta) = m(x,\theta)$, $Q(x,\theta)$, *and* $U(x,\theta)$, *and (iii)* $\sup_i EA(X_i)^4 \leq C$ *for* $f(x,\theta) = v(x,\theta)$

6. *The observations are grouped in independent clusters of size* n_g

7. *If the weight matrix is unclustered (14) then*

   (a) $n^{-1}\sum_{g=1}^G n_g^2 \leq C$

   (b) $\max_{g\leq G} n_g^2/n \to 0$

8. *If the weight matrix is clustered (15) then*

   (a) $n^{-1}\sum_{g=1}^G n_g^4 \leq C$

   (b) $\max_{g\leq G} n_g^4/n \to 0$

   (c) $\|W_n(\theta)\| \leq C$ *and* $\|S_n(\theta)\| \leq C$

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

$$\left(H_n^{-1}\Omega_n H_n^{-1\prime}\right)^{-1/2}\sqrt{n}\left(\widehat{\theta} - \theta_n\right) \xrightarrow{d} N\left(\mathbf{0}, I_k\right) \tag{23}$$

where Ω_n *and* H_n *are defined in (21) and (22), respectively.*

Theorem 3 provides a simple characterization of the asymptotic distribution of the GMM estimator under possible moment misspecification and cluster sampling.

The asymptotic variance in Theorem 3 differs from the classical formula

$$\left(Q'W^{-1}Q\right)^{-1}\left(Q'W^{-1}\Omega_{11}W^{-1}Q\right)\left(Q'W^{-1}Q\right)^{-1} \tag{24}$$

where Ω_{11} is (21) with $\mu_n = 0$, in two ways. First, the matrix H_n defined in (22) is a function of the curvature in $Q_n(\theta)$ and $W_n(\theta)$ through the matrix derivatives R_n and S_n. Larger curvature implies larger distortions. Second, the asymptotic covariance matrix Ω_n defined in (21) of the vector ψ_g is

an augmented version of the classic covariance matrix. Ω_n is augmented by the variation in $\widetilde{Q}$ and $\widetilde{W}_g$. Larger variance in these variables implies larger distortions. All of these differences disappear when $\mu_n = 0$ (correct specification) but appear when $\mu_n \neq 0$.

Theorem 3 is agnostic about whether or not the model is correctly specified, and thus provides valid covariance matrix estimates and standard errors without sensitivity to specification. This is a more robust distribution theory, and also important in studying test power and bootstrap distributions.

The asymptotic distribution in Theorem 3 is similar to that obtained by Hall and Inoue (2003) and they are equivalent when $W_n = W_n(\theta)$ does not depend on θ in the i.i.d. case. An important distinction is that Theorem 3 allows $W_n(\theta)$ to depend on θ and thus includes the iterated GMM estimator. Theorem 3 is the first distribution theory which formally covers the iterated GMM estimator, both under correct specification and misspecification, and to allow for cluster sampling which includes random sampling as a special case.

Assumption 3.1 is a full-rank condition on the effective Hessian H_n; it simplifies to the full column rank condition of Q_n if $\mu_n = 0$. Assumption 3.2 excludes singular covariance matrices. Assumption 3.4 impose uniform integrability on specific powers of the moments and their derivatives. In the i.i.d. case Assumption 3.4 simplifies to moment bounds. Assumption 3.5 are Lipschitz bounds on the same functions.

Clustered sampling is permitted by Assumptions 3.6-3.8. Assumptions 3.7-3.8 control the degree of heterogeneity in the cluster sizes n_g. Condition (b) implies $G \to \infty$ so we are in the "large number of clusters" asymptotic framework. Assumption 3.7 is used when the weight matrix is unclustered (as in 2SLS). The assumptions allow for mildly growing and heterogeneous cluster sizes. Assumption 3.8 is used when the weight matrix is unclustered (as when the optimal weight matrix is used in a clustered sample). Conditions (a) and (b) are more restrictive versions of the conditions in Assumption 3.7. Condition (c) requires that the weight matrix and its derivative is bounded. This will hold (under the other assumptions) if the cluster sizes n_g are bounded, but may not hold otherwise. This assumption may be important (may not simply be a technical condition) since if the weight matrix is unbounded as the sample size diverges, then it is unclear if the GMM criterion will be well behaved. The conditions in Assumptions 3.7-3.8 are probably not the weakest possible.

The asymptotic distribution (23) implies that the approximate scaled variance matrix is $H_n^{-1}\Omega_n H_n^{-1\prime}$. It does not require, however, that scaled variance matrix converges to a constant. This is important for clustered data as Ω_n may not converge even after re-scaling.

6 Covariance Matrix Estimation

It is straightforward to calculate an estimate of the covariance matrix

$$V_n = H_n^{-1}\Omega_n H_n^{-1\prime}$$

from Theorem 3. Construct the derivatives $\widehat{Q} = \overline{Q}_n(\widehat{\theta})$, $\widehat{R} = \overline{R}_n(\widehat{\theta})$, $\widehat{S} = \overline{S}_n(\widehat{\theta})$, $\widehat{W} = \overline{W}_n(\widehat{\theta})$, $\widehat{\mu} = \overline{m}_n(\widehat{\theta})$ and

$$\widehat{H} = \overline{H}_n(\widehat{\theta}) = \widehat{Q}'\widehat{W}^{-1}\widehat{Q} + (\widehat{\mu}'\widehat{W}^{-1} \otimes I_k)\widehat{R} - (\widehat{\mu}'\widehat{W}^{-1} \otimes \widehat{Q}'\widehat{W}^{-1})\widehat{S}. \tag{25}$$

Construct the cluster sum

$$\widehat{\psi}_g = \widehat{Q}'\widehat{W}^{-1}\widetilde{m}_g(\widehat{\theta}) + \widetilde{Q}_g(\widehat{\theta})'\widehat{W}^{-1}\widehat{\mu} - \widehat{Q}'\widehat{W}^{-1}\widetilde{W}_g(\widehat{\theta})\widehat{W}^{-1}\widehat{\mu} \tag{26}$$

and the cluster variance estimator

$$\widehat{\Omega} = \frac{1}{n}\sum_{g=1}^{G} \widehat{\psi}_g\widehat{\psi}_g' \tag{27}$$

and

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

The standard errors for $\widehat{\theta}$ can be obtained by taking the square roots of the diagonal elements of $n^{-1}\widehat{V}$. In the case of no clustering, set $G = n$ and $n_g = 1$.

We now establish that $\widehat{V}$ is consistent for V and that replacement in Theorem 3 of V_n by $\widehat{V}$ does not affect the asymptotic distribution.

Theorem 4. *Under Assumptions 1 and 3,*

$$\left\| V_n^{-1/2}\widehat{V}V_n^{-1/2} - I_k \right\| \to_p 0 \tag{29}$$

and

$$\widehat{V}^{-1/2}\sqrt{n}\left(\widehat{\theta} - \theta_n\right) \xrightarrow{d} N\left(\mathbf{0}, I_k\right) \tag{30}$$

as $n \to \infty$.

Equation (29) shows that $\widehat{V}$ is consistent in a sense appropriate when the variance matrix may not be converging with n. Equation (30) 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.

While Theorem 4 is quite straightforward to establish in the case of i.i.d. observations, it is quite tedious to establish in the case of clustered observations with a clustered weight matrix. In particular, the clustered weight matrix poses special challenges. This is because $\widehat{\Omega}$ includes terms $\widetilde{W}_g\widetilde{W}_g'$ which are effectively the fourth power of a cluster sum. This required an extension of the uniform convergence theory of Hansen and Lee (2017, Theorem 6). The proof method is similar but the details are somewhat tedious, so we present the proof in the Appendix.

To emphasize, Theorem 4 shows that robust t-statistics and Wald statistics have conventional asymptotic distributions, but this requires that the covariance matrix has been calculated with our new robust estimator which accounts for possible misspecification. The result fails if conventional covariance matrix estimators are used, as they are in general inconsistent for the correct variance.

7 Weight Matrix Invariance

If the inverse of the weight matrix is consistent for the covariance matrix of the moment function, the asymptotic variance of the GMM estimator is minimized under correct specification, which leads to efficient GMM. Under independent sampling, efficient weight matrices are constructed using the moment function $m(X_i, \theta)$ or its recentered version, $m(X_i, \theta) - \overline{m}_n(\theta)$. An open question is whether recentering affects the pseudo-true value θ_n or the estimate $\widehat{\theta}$. Recentering is known to not affect the continuous-updating GMM estimator (see Newey and Smith (2004), footnote 2) but its impact on the iterated GMM estimator is unknown. We now show that recentering has no effect for the efficient weight matrix, either clustered or non-clustered.

The covariance matrix of $\sqrt{n}\overline{m}_n(\theta)$ is

$$\begin{aligned}
\Omega_n(\theta) &= E\left(n\left(\overline{m}_n(\theta) - E\overline{m}_n(\theta)\right)\left(\overline{m}_n(\theta) - E\overline{m}_n(\theta)\right)'\right) \\
&= \frac{1}{n}\sum_{g=1}^{G} E\left(\left(\widetilde{m}_g(\theta) - E\widetilde{m}_g(\theta)\right)\left(\widetilde{m}_g(\theta) - E\widetilde{m}_g(\theta)\right)'\right) \\
&= \frac{1}{n}\sum_{g=1}^{G} E\widetilde{m}_g(\theta)\widetilde{m}_g(\theta)' - \frac{1}{n}\sum_{g=1}^{G} E\widetilde{m}_g(\theta)E\widetilde{m}_g(\theta)'.
\end{aligned}$$

If we further assume that $m(\theta) \equiv Em(X_i, \theta)$ does not vary across observations, then

$$\Omega_n(\theta) = \frac{1}{n}\sum_{g=1}^{G} E\widetilde{m}_g(\theta)\widetilde{m}_g(\theta)' - \left(\frac{1}{n}\sum_{g=1}^{G} n_g^2\right) m(\theta)m(\theta)'.$$

Depending on the user's assumption whether the model is correctly specified or possibly misspecified, the clustered efficient weight matrices are defined as either

$$\overline{W}_n(\theta) = \frac{1}{n}\sum_{g=1}^{G} \widetilde{m}_g(\theta)\widetilde{m}_g(\theta)' \tag{31}$$

or

$$\overline{W}_n^*(\theta) = \frac{1}{n}\sum_{g=1}^{G} \widetilde{m}_g(\theta)\widetilde{m}_g(\theta)' - \left(\frac{1}{n}\sum_{g=1}^{G} n_g^2\right) \overline{m}_n(\theta)\overline{m}_n(\theta)'. \tag{32}$$

Under independent sampling, we set $n = G$ and $n_g = 1$ so that (31) and (32) become

$$\overline{W}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} m(X_i, \theta)m(X_i, \theta)' \tag{33}$$

and

$$\overline{W}_n^*(\theta) = \frac{1}{n}\sum_{i=1}^{n} m(X_i, \theta)m(X_i, \theta)' - \overline{m}_n(\theta)\overline{m}_n(\theta)', \tag{34}$$

respectively.

Theorem 5. *Suppose that Assumptions 1 and 2 hold for all the weight matrices defined above. The pseudo-true value θ_n and iterated GMM estimate $\widehat{\theta}$ are invariant to the choice of (i) either* (31) *or* (32) *for the clustered weight matrix, and (ii) either* (33) *or* (34) *for the non-clustered weight matrix.*

Theorem 5 shows that recentering the weight matrix does not alter the pseudo-true value θ_n and iterated GMM estimate $\widehat{\theta}$ when the weight matrix takes the efficient form. However, while the estimator itself is unaffected the covariance matrix estimate, the criterion function, and LR-type test statistics are affected by the choice. Their relative finite sample performance deserves more research.

It is important to understand that Theorem 5 applies only to the GMM estimator when iterated until convergence. It does not apply to the s-step estimator.

8 Over-identifying Restrictions Test

It is conventional to report the over-identifying restrictions test (the J test) statistic along with the GMM point estimates and standard errors. The J statistic is the GMM criterion evaluated at the estimator with the efficient weight matrix:

$$\overline{J}_n = n \cdot \overline{m}_n(\widehat{\theta})' \widehat{W}^{-1} \overline{m}_n(\widehat{\theta}) \tag{35}$$

where $\widehat{W}$ is either (32) or (31) under cluster sampling and either (34) or (33) under independent sampling, evaluated at the estimator. The null hypothesis is that the moment condition is correctly specified, $E(m(X_i, \theta)) = 0$.

It is straightforward to show that (35) is consistent for the chi-square distribution with the degrees of freedom $l - k$ under the null under independent sampling. Under cluster sampling, the clustered efficient weight matrix should be used because the conventional non-clustered efficient weight matrix is no longer consistent for the asymptotic covariance matrix of the moment in general. Theorem 17 Equation (68) of Hansen and Lee (2017), which shows consistency of the J test under cluster sampling, holds with the iterated GMM estimator.

Both the uncentered and re-centered efficient weight matrices can be used for the test statistic and the resulting two J statistics are closely related. Indeed, from the proof of Theorem 5, we can deduce that

$$\overline{J}_n = \frac{1}{1 + (C_n/n)\,\overline{J}_n^*} \overline{J}_n^* \tag{36}$$

where $\overline{J}_n^*$ is the J test statistic using the centered weight matrix $\widehat{W}^*$ and $C_n = 1$ when (33) and (34) are used and $C_n = \sum_{g=1}^G n_g^2/n$ when (31) and (32) are used. Since $(C_n/n)\,\overline{J}_n^* \geq 0$,

$$\overline{J}_n \leq \overline{J}_n^*. \tag{37}$$

So the J statistics are rank ordered and are monotonic functions of the other, which implies that they are identical *tests*, holding size constant. For example, if the bootstrap p-values are calculated they will be identical.

In practice, the centered version is larger so will reject more frequently based on asymptotic critical values. Since J tests tend to over-reject in finite samples, this could result in spurious over-rejection. On the other hand, the centered version may have better power. Hall (2000) shows that the J test based on the two-step GMM using the centered heteroskedasticity and autocorrelation consistent covariance (HACC) weight matrix is more powerful than the other in larger samples.

9 Linear Model

Consider the linear model $y_i = x_i'\theta + e_i$ with moment condition $E(z_i e_i) = 0$. We consider two possible weight matrices, corresponding to 2SLS and conventional efficient GMM. For each weight matrix we describe the covariance matrix estimator under independent and clustered dependence.

First, take the case of 2SLS. The estimator is

$$\widehat{\theta} = (X'Z\left(Z'Z\right)^{-1}Z'X)^{-1}(X'Z\left(Z'Z\right)^{-1}Z'Y).$$

where X, Y, Z are stacked data matrices. No iteration is required. The residuals are $\widehat{e}_i = y_i - x_i'\widehat{\theta}$ and let $\widehat{e} = Y - X\widehat{\theta}$.

The asymptotic covariance matrix estimate for $\widehat{\theta}$ takes the form

$$\begin{aligned}
\widehat{V} &= \widehat{H}^{-1}\widehat{\Omega}\widehat{H}^{-1} \\
\widehat{H} &= \frac{1}{n}X'Z\left(Z'Z\right)^{-1}Z'X \\
\widehat{\Omega} &= \frac{1}{n}\widehat{\Psi}'\widehat{\Psi}.
\end{aligned}$$

If the observations are independent (not clustered) then $\widehat{\Psi}$ is an $n \times k$ matrix whose i^{th} row is $\widehat{\psi}_i'$ where

$$\widehat{\psi}_i = -X'Z\left(Z'Z\right)^{-1}z_i\widehat{e}_i - x_i z_i'\left(Z'Z\right)^{-1}Z'\widehat{e} + X'Z\left(Z'Z\right)^{-1}z_i z_i'\left(Z'Z\right)^{-1}Z'\widehat{e}.$$

If the observations are clustered, then the only modification is that $\widehat{\Psi}$ is an $G \times k$ matrix whose g^{th} row is $\widetilde{\psi}_g'$ where

$$\widetilde{\psi}_g = \sum_{j=1}^{n_g}\widehat{\psi}_i$$

Second, take the case of efficient GMM. Under independent sampling, a non-clustered weight matrix takes the form

$$\overline{W}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} z_i z_i'\left(y_i - x_i'\theta\right)^2.$$

If the observations are clustered, then a clustered weight matrix takes the form

$$\overline{W}_n(\theta) = \frac{1}{n}\sum_{g=1}^{G} Z_g'(Y_g - X_g\theta)(Y_g - X_g\theta)'Z_g$$

where Z_g, Y_g, and X_g are $n_g \times l$, $n_g \times 1$, and $n_g \times k$ stacked data matrices for cluster g.

Given a preliminary estimate $\widehat{\theta}_0$ the s-step GMM estimator is defined by

$$\widehat{\theta}_s = (Z'X\overline{W}_n(\widehat{\theta}_{s-1})^{-1}X'Z)^{-1}(Z'X\overline{W}_n(\widehat{\theta}_{s-1})^{-1}X'Y).$$

The iterated GMM estimator $\widehat{\theta}$ is this limit iterated until convergence. The residuals are $\widehat{e}_i = y_i - x_i'\widehat{\theta}$. Let $\widehat{e} = Y - X\widehat{\theta}$ and $\widehat{e}_g = Y_g - X_g\widehat{\theta}$. Set $\widehat{W} = \overline{W}_n(\widehat{\theta})$.

The asymptotic covariance matrix estimate under independent sampling (not clustered) takes the form

$$\begin{aligned}
\widehat{V} &= \widehat{H}^{-1}\widehat{\Omega}\widehat{H}^{-1\prime} \\
\widehat{H} &= \frac{1}{n^2}X'Z\widehat{W}^{-1}Z'X - \frac{2}{n^3}X'Z\widehat{W}^{-1}\sum_{i=1}^{n} z_i x_i'\left(\widehat{e}_i z_i'\widehat{W}^{-1}Z'\widehat{e}\right)
\end{aligned}$$

where $\widehat{\Psi}$ is an $n \times k$ matrix whose i^{th} row is $\widehat{\psi}_i'$ where

$$\widehat{\psi}_i = -\frac{1}{n}X'Z\widehat{W}^{-1}z_i\widehat{e}_i - \frac{1}{n}x_i z_i'\widehat{W}^{-1}Z'\widehat{e} + \frac{1}{n^2}X'Z\widehat{W}^{-1}z_i z_i'\widehat{e}_i^2\widehat{W}^{-1}Z'\widehat{e}.$$

If the observations are clustered, then $\widehat{H}$ is

$$\widehat{H} = \frac{1}{n^2}X'Z\widehat{W}^{-1}Z'X - \frac{1}{n^3}X'Z\widehat{W}^{-1}\sum_{g=1}^{G}\left(Z_g'\widehat{e}_g\widehat{e}'Z\widehat{W}^{-1}Z_g'X_g + Z_g'X_g\left(\widehat{e}'Z\widehat{W}^{-1}Z_g'\widehat{e}_g\right)\right)$$

and $\widehat{\Psi}$ is an $G \times k$ matrix whose g^{th} row is $\widetilde{\psi}_g'$ where

$$\widetilde{\psi}_g = -\frac{1}{n}X'Z\widehat{W}^{-1}Z_g'\widehat{e}_g - \frac{1}{n}X_g'Z_g\widehat{W}^{-1}Z'\widehat{e} + \frac{1}{n^2}X'Z\widehat{W}^{-1}Z_g'\widehat{e}_g\widehat{e}_g'Z_g\widehat{W}^{-1}Z'\widehat{e}.$$

10 Simulation

In this section, we illustrate the performance of the methods proposed in this paper in two simulation experiments, one for i.i.d. observations and one for clustered dependence. In both experiments we investigate the performance of inference methods for 2SLS and iterated efficient GMM estimation, using both conventional and our recommended robust standard errors. We find large and important improvements in performance by using our recommended methods.

Our first experiment concerns a simple linear instrumental variable regression with a single

endogenous regressor. The model to be estimated is

$$\begin{aligned} y_i &= x_i\theta_0 + e_i \\ E\left(z_i e_i\right) &= 0 \end{aligned} \tag{38}$$

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 2SLS and iterated efficient GMM, and calculate standard errors using the conventional heteroskedasticity-robust formula and using our misspecification-robust formula. All results are reported using 5000 Monte Carlo replications.

Our data-generating process is

$$\begin{aligned} y_i &= x_i\theta_0 + \alpha_0\left(z_{1i} - z_{2i} + z_{3i} - z_{4i}\right) + e_i, \\ x_i &= \pi_0\left(z_{1i} + z_{2i} + z_{3i} + z_{4i}\right) + u_i, \\ z_i &\sim N\left(0, I_4\right), \\ \begin{pmatrix} e_i \\ u_i \end{pmatrix} &\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix}\right). \end{aligned} \tag{39}$$

We set $\theta_0 = 1$, vary α_0 from 0 to 1 in steps of 0.1, and set the first-stage coefficient π_0 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_0. At $\alpha_0 = 0$, the model is correctly specified. For $\alpha_0 \neq 0$ we find

$$E\left(z_i e_i\right) = \begin{pmatrix} \alpha_0 \\ -\alpha_0 \\ \alpha_0 \\ -\alpha_0 \end{pmatrix} = \mu_0 \neq 0$$

so the moment condition (38) fails to hold. The pseudo-true value θ_0, however is invariant to α_0.

We first investigate the performance of inference based on 2SLS estimation. 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 conventional heteroskedasticity-robust standard errors relative to the standard deviation of $\widehat{\theta}$. We can see that the standard errors are unbiased for $\alpha_0 = 0$, but highly biased for $\alpha_0 \neq 0$. 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 $\alpha_0 = 0$, with approximately a 10% downwards bias for $\alpha_0 = 0.1$ under strong identification, and a 30-40% downward bias under weak identification. The bias is increasing in α_0, and does not improve with sample size. Indeed the worst case arises for $\alpha_0 = 1$, $R^2 = 0.02$, 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 six and seven we report the size of nominal asymptotic 5% t-tests for $H_0 : \theta_0 = 1$ against $H_1 : \theta_0 \neq 1$. Column six reports the size of tests using our proposed misspecification-robust standard errors. We can see that under strong identification the tests have excellent size performance. Under weak identification and the smaller sample size we can see that there is mild size distortion (rejection rates range from 0.049 to 0.078) but the distortion disappears when the sample size is increased. Column seven 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 severe even for the mildest departures from correct specification in the presence of weak instruments. Indeed, the size of the t-test is 23% for $\alpha = 0.1$ and $n = 2500$, equals 46% for $\alpha = 0.2$, and exceeds 70% for $\alpha \geq 0.7$.

In the final column 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.

We second investigate the performance of inference based on iterated efficient GMM. The results are reported in Table 2. The same statistics are reported as in Table 1, plus the median number of iterations required to obtain convergence.

Overall, the results in Table 2 are similar to those in Table 1. The main difference is 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%), but this disappears as the sample size increases.

The final column reports the median number of iterations required to obtain GMM 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.

On a final note it is also quite interesting to point out the behavior of the statistics when there is no misspecification ($\alpha_0 = 0$). 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, in both tables 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. This finding points out that there is no apparent cost of using our new robust standard errors, even in the context of no misspecification.

R^2	n	α_0	$s_r(\hat{\theta})/s.d$	$s(\hat{\theta})/s.d$	Size(t_r)	Size(t)	Reject(J)
0.2	250	0	1.0209	0.9964	0.0568	0.0616	0.0584
		0.1	1.0054	0.9187	0.0580	0.0792	0.7212
		0.2	0.9869	0.7818	0.0576	0.1266	0.9990
		0.3	0.9910	0.6879	0.0522	0.1764	1.0000
		0.4	1.0050	0.6290	0.0518	0.2126	1.0000
		0.5	0.9774	0.5688	0.0604	0.2516	1.0000
		0.6	0.9994	0.5508	0.0512	0.2730	1.0000
		0.7	0.9892	0.5256	0.0548	0.2902	1.0000
		0.8	0.9931	0.5127	0.0514	0.3108	1.0000
		0.9	0.9987	0.5040	0.0542	0.3200	1.0000
		1.0	0.9712	0.4832	0.0564	0.3304	1.0000
	2500	0	1.0132	1.0107	0.0474	0.0478	0.0484
		0.1	0.9969	0.9260	0.0510	0.0686	1.0000
		0.2	0.9912	0.7946	0.0506	0.1196	1.0000
		0.3	1.0107	0.7039	0.0464	0.1662	1.0000
		0.4	1.0006	0.6256	0.0514	0.2176	1.0000
		0.5	1.0052	0.5812	0.0508	0.2530	1.0000
		0.6	1.0171	0.5553	0.0470	0.2742	1.0000
		0.7	0.9909	0.5198	0.0518	0.3036	1.0000
		0.8	0.9849	0.5017	0.0570	0.3250	1.0000
		0.9	1.0127	0.5036	0.0466	0.3236	1.0000
		1.0	1.0024	0.4901	0.0490	0.3294	1.0000
0.02	250	0	1.2982	1.0149	0.0784	0.1172	0.0552
		0.1	1.1479	0.6974	0.0662	0.1898	0.6098
		0.2	1.2335	0.5374	0.0610	0.2806	0.8784
		0.3	1.0764	0.4568	0.0558	0.3538	0.8980
		0.4	1.0923	0.4344	0.0580	0.4176	0.9104
		0.5	1.0842	0.4191	0.0490	0.4442	0.9160
		0.6	1.0688	0.4120	0.0524	0.4684	0.9164
		0.7	1.1297	0.4106	0.0582	0.4710	0.9198
		0.8	1.0245	0.3980	0.0520	0.4870	0.9220
		0.9	1.0509	0.3988	0.0594	0.4974	0.9164
		1.0	1.0715	0.3929	0.0568	0.5148	0.9228
	2500	0	1.0494	1.0188	0.0476	0.0534	0.0506
		0.1	1.0104	0.5963	0.0492	0.2284	0.9998
		0.2	0.9983	0.3724	0.0496	0.4590	1.0000
		0.3	0.9966	0.2907	0.0524	0.5834	1.0000
		0.4	1.0020	0.2552	0.0478	0.6542	1.0000
		0.5	1.0044	0.2339	0.0468	0.6756	1.0000
		0.6	1.0119	0.2237	0.0486	0.6976	1.0000
		0.7	1.0054	0.2139	0.0528	0.7202	1.0000
		0.8	1.0184	0.2104	0.0508	0.7274	1.0000
		0.9	1.0197	0.2065	0.0504	0.7234	1.0000
		1.0	1.0113	0.2027	0.0510	0.7414	1.0000

Table 1: Monte Carlo Results for Linear Model - 2SLS

R^2	n	α_0	$s_r(\hat{\theta})/s.d$	$s(\hat{\theta})/s.d$	Size(t_r)	Size(t)	Reject(J)	Med. Iter
0.2	250	0	1.0119	0.9749	0.0596	0.0658	0.0578	3
		0.1	1.0011	0.8716	0.0618	0.0938	0.7138	4
		0.2	0.9777	0.6926	0.0648	0.1638	0.9992	7
		0.3	0.9944	0.5864	0.0738	0.2370	1.0000	9
		0.4	1.0015	0.5124	0.0838	0.3046	1.0000	11
		0.5	0.9657	0.4498	0.0954	0.3736	1.0000	12
		0.6	0.9728	0.4278	0.0982	0.4002	1.0000	14
		0.7	0.9890	0.4098	0.1032	0.4242	1.0000	15
		0.8	0.9783	0.3903	0.1086	0.4530	1.0000	16
		0.9	0.9457	0.3786	0.1244	0.4766	1.0000	16
		1.0	0.9539	0.3661	0.1210	0.4870	1.0000	17
	2500	0	1.0115	1.0079	0.0484	0.0484	0.0486	2
		0.1	0.9954	0.8918	0.0530	0.0832	1.0000	4
		0.2	0.9974	0.7216	0.0498	0.1610	1.0000	6
		0.3	1.0164	0.6007	0.0486	0.2334	1.0000	8
		0.4	0.9904	0.4990	0.0552	0.3244	1.0000	11
		0.5	0.9885	0.4437	0.0594	0.3894	1.0000	13
		0.6	1.0087	0.4165	0.0554	0.4082	1.0000	14
		0.7	0.9836	0.3832	0.0608	0.4508	1.0000	15
		0.8	0.9846	0.3666	0.0616	0.4732	1.0000	16
		0.9	0.9995	0.3610	0.0586	0.4920	1.0000	17
		1.0	1.0000	0.3515	0.0578	0.4964	1.0000	18
0.02	250	0	1.2766	0.9924	0.0832	0.1250	0.0536	4
		0.1	1.1371	0.6791	0.0742	0.2002	0.6020	5
		0.2	1.2603	0.5280	0.0738	0.2930	0.8730	6
		0.3	1.0416	0.4492	0.0782	0.3828	0.8938	7
		0.4	1.0266	0.4247	0.0760	0.4456	0.9052	8
		0.5	1.0291	0.4105	0.0744	0.4710	0.9080	8
		0.6	1.0044	0.4045	0.0740	0.4878	0.9120	9
		0.7	1.0620	0.4022	0.0772	0.4900	0.9146	9
		0.8	0.9611	0.3919	0.0726	0.5102	0.9180	9
		0.9	1.0047	0.3932	0.0806	0.5172	0.9118	9
		1.0	0.9827	0.3842	0.0796	0.5396	0.9198	9
	2500	0	1.0476	1.0157	0.0478	0.0542	0.0506	3
		0.1	1.0105	0.5875	0.0512	0.2356	0.9998	4
		0.2	1.0030	0.3662	0.0552	0.4726	1.0000	7
		0.3	0.9969	0.2841	0.0586	0.5974	1.0000	9
		0.4	0.9946	0.2492	0.0642	0.6648	1.0000	10
		0.5	1.0123	0.2298	0.0606	0.6942	1.0000	12
		0.6	1.0092	0.2189	0.0612	0.7164	1.0000	12
		0.7	1.0128	0.2098	0.0636	0.7234	1.0000	13
		0.8	1.0195	0.2058	0.0668	0.7536	1.0000	14
		0.9	1.0259	0.2026	0.0590	0.7448	1.0000	14
		1.0	1.0078	0.1980	0.0630	0.7496	1.0000	15

Table 2: Monte Carlo Results for Linear Model - Iterated GMM

R^2	n	G	α_0	$s_r(\hat{\theta})/s.d$	$s(\hat{\theta})/s.d$	Size(t_r)	Size(t)	Reject(J)
0.2	400	100	0	0.9930	0.9776	0.0622	0.0650	0.0538
			0.1	0.9787	0.9104	0.0562	0.0766	0.8238
			0.2	0.9967	0.8182	0.0568	0.1162	1.0000
			0.3	0.9810	0.7089	0.0582	0.1616	1.0000
			0.4	0.9974	0.6510	0.0562	0.2014	1.0000
			0.5	0.9759	0.5872	0.0562	0.2468	1.0000
			0.6	0.9927	0.5644	0.0548	0.2634	1.0000
			0.7	0.9899	0.5378	0.0584	0.2884	1.0000
			0.8	0.9854	0.5175	0.0550	0.3046	1.0000
			0.9	0.9848	0.5047	0.0550	0.3236	1.0000
			1.0	0.9889	0.4965	0.0524	0.3348	1.0000
	4000	1000	0	0.9945	0.9929	0.0554	0.0556	0.0518
			0.1	0.9946	0.9366	0.0504	0.0686	1.0000
			0.2	1.0104	0.8374	0.0504	0.1056	1.0000
			0.3	1.0023	0.7278	0.0504	0.1500	1.0000
			0.4	0.9870	0.6430	0.0534	0.2054	1.0000
			0.5	1.0001	0.6007	0.0504	0.2434	1.0000
			0.6	1.0009	0.5649	0.0538	0.2640	1.0000
			0.7	1.0033	0.5413	0.0472	0.2874	1.0000
			0.8	1.0000	0.5212	0.0488	0.3100	1.0000
			0.9	0.9947	0.5055	0.0518	0.3194	1.0000
			1.0	1.0035	0.4994	0.0476	0.3256	1.0000
0.02	400	100	0	1.2058	1.0051	0.0792	0.1100	0.0598
			0.1	1.1046	0.6834	0.0602	0.1952	0.7298
			0.2	1.0471	0.4852	0.0582	0.2998	0.9312
			0.3	1.0332	0.4100	0.0550	0.4084	0.9600
			0.4	1.0207	0.3755	0.0560	0.4714	0.9622
			0.5	1.0124	0.3596	0.0592	0.5232	0.9674
			0.6	1.0372	0.3565	0.0560	0.5496	0.9664
			0.7	1.0295	0.3461	0.0610	0.5726	0.9628
			0.8	1.0388	0.3454	0.0584	0.5756	0.9666
			0.9	1.0515	0.3438	0.0574	0.5954	0.9720
			1	0.9466	0.3313	0.0590	0.6082	0.9674
	4000	1000	0	1.0216	1.0025	0.0532	0.0574	0.0560
			0.1	1.0008	0.6286	0.0546	0.2060	1.0000
			0.2	0.9922	0.3961	0.0520	0.4274	1.0000
			0.3	1.0045	0.3040	0.0502	0.5538	1.0000
			0.4	0.9888	0.2537	0.0502	0.6374	1.0000
			0.5	1.0062	0.2335	0.0470	0.6712	1.0000
			0.6	0.9981	0.2152	0.0484	0.6974	1.0000
			0.7	1.0070	0.2066	0.0492	0.7228	1.0000
			0.8	0.9996	0.1983	0.0526	0.7290	1.0000
			0.9	0.9979	0.1938	0.0518	0.7364	1.0000
			1.0	1.0009	0.1902	0.0500	0.7352	1.0000

Table 3: Monte Carlo Results for Clustered Dependence - 2SLS

R^2	n	G	α_0	$s_r(\hat{\theta})/s.d$	$s(\hat{\theta})/s.d$	Size(t_r)	Size(t)	Reject(J)	Med. Iter
0.2	400	100	0	0.9825	0.9456	0.0644	0.0748	0.0488	4
			0.1	0.9874	0.8287	0.0668	0.1126	0.8060	6
			0.2	0.9719	0.6200	0.0760	0.2258	1.0000	11
			0.3	0.9706	0.4906	0.0878	0.3242	1.0000	16
			0.4	0.9844	0.4151	0.0944	0.4166	1.0000	21
			0.5	0.9598	0.3652	0.1136	0.4834	0.9998	25
			0.6	0.9454	0.3327	0.1346	0.5304	1.0000	28
			0.7	0.9351	0.3125	0.1384	0.5660	1.0000	30
			0.8	0.9306	0.2946	0.1488	0.5858	1.0000	32
			0.9	0.9255	0.2878	0.1450	0.6012	1.0000	34
			1.0	0.9042	0.2761	0.1532	0.6204	1.0000	35
	4000	1000	0	1.0098	1.0062	0.0470	0.0478	0.0566	2
			0.1	0.9982	0.8627	0.0526	0.0906	1.0000	5
			0.2	0.9840	0.6476	0.0612	0.2016	1.0000	9
			0.3	0.9767	0.4987	0.0560	0.3326	1.0000	15
			0.4	1.0055	0.4211	0.0544	0.4122	1.0000	20
			0.5	0.9804	0.3570	0.0630	0.4752	1.0000	24
			0.6	0.9976	0.3276	0.0598	0.5236	1.0000	28
			0.7	0.9857	0.2977	0.0610	0.5594	1.0000	32
			0.8	1.0084	0.2862	0.0608	0.5816	1.0000	35
			0.9	0.9821	0.2685	0.0662	0.5996	1.0000	37
			1.0	1.0031	0.2646	0.0646	0.6026	1.0000	39
0.02	400	100	0	1.1591	0.9284	0.0854	0.1224	0.0576	4
			0.1	1.0830	0.6335	0.0824	0.2208	0.7174	6
			0.2	1.0187	0.4543	0.0926	0.3506	0.9296	9
			0.3	0.9800	0.3908	0.1002	0.4566	0.9526	12
			0.4	1.0201	0.3747	0.1018	0.5218	0.9556	14
			0.5	0.8934	0.3416	0.1040	0.5716	0.9606	15
			0.6	0.9190	0.3381	0.1020	0.5992	0.9666	16
			0.7	0.9104	0.3303	0.1048	0.6166	0.9604	17
			0.8	0.8830	0.3232	0.1020	0.6274	0.9630	17
			0.9	0.9146	0.3295	0.0986	0.6344	0.9638	17
			1.0	0.8780	0.3249	0.1008	0.6490	0.9654	18
	4000	1000	0	1.0317	1.0106	0.0494	0.0528	0.0586	3
			0.1	1.0020	0.6038	0.0574	0.2304	1.0000	6
			0.2	0.9854	0.3715	0.0624	0.4624	1.0000	10
			0.3	0.9977	0.2830	0.0650	0.5924	1.0000	15
			0.4	1.0017	0.2382	0.0684	0.6684	1.0000	20
			0.5	0.9901	0.2169	0.0778	0.6984	1.0000	23
			0.6	0.9981	0.2026	0.0758	0.7324	1.0000	26
			0.7	1.0008	0.1925	0.0770	0.7476	1.0000	28
			0.8	0.9872	0.1856	0.0808	0.7608	1.0000	30
			0.9	1.0036	0.1827	0.0824	0.7648	1.0000	31
			1.0	0.9922	0.1786	0.0796	0.7716	1.0000	33

Table 4: Monte Carlo Results for Clustered Dependence - Iterated GMM

Our second experiment introduces clustered dependence. The model is

$$\begin{aligned} y_{gj} &= x_{gj}\theta_0 + e_{gj} \\ E\left(z_{gj}e_{gj}\right) &= 0 \end{aligned}$$

for $g = 1, 2, ..., G$ and $j = 1, 2, ..., n_g$. We set $G = (100, 1000)$ and set the cluster sizes as $n_g = 2$ for one-half of the clusters and $n_g = 6$ for the other half. Thus the total sample sizes are $n = (400, 4000)$.

Our data-generating process is similar to (39), except that the errors are generated as

$$\begin{aligned} y_{gj} &= x_{gj}\theta_0 + \alpha_0\left(z_{1gj} - z_{2gj} + z_{3gj} - z_{4gj}\right) + v_g + e_{gj}, \\ x_{gj} &= \pi_0\left(z_{1gj} + z_{2gj} + z_{3gj} + z_{4gj}\right) + u_{gj}, \\ z_{gj} &\sim N\left(0, I_4\right), \\ v_g &\sim N\left(0, 1/4\right), \\ \begin{pmatrix} e_{gj} \\ u_{gj} \end{pmatrix} &\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & \sqrt{5}/4 \\ \sqrt{5}/4 & 1 \end{bmatrix}\right). \end{aligned}$$

The parameters of the model are selected so that the equation error $v_g + e_{gj}$ has a correlation of 0.5 with the reduced form error u_{gj}, and the equation error $v_g + e_{gj}$ has a correlation of 0.2 within each cluster.

In this experiment, we report results based on conventional cluster-robust standard errors as well as our new misspecification-robust clustered standard errors.

The results for 2SLS are presented in Table 3 and those for the same model estimated by iterated GMM presented in Table 4. Overall, the results in the two tables are quite similar to those presented in Tables 1 and 2, showing unacceptable performance of the conventional cluster-robust standard errors and t tests, and excellent performance of the new robust standard errors and t tests. In the smaller sample the t-tests have somewhat more size distortion under cluster dependence than in the i.i.d. case, but the distortion diminishes with sample size. It is also interesting to note that the number of required iterations for efficient GMM is higher in Table 4 than in the i.i.d. case.

In unreported simulation results, we varied the number of instruments, the ratio of n and G, the degree of cluster size heterogeneity, and the correlation within clusters. We found that the main conclusions were largely unchanged.

11 Application: Income and Democracy

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

AJRY estimate the dynamic panel regression

$$y_{it} = \alpha y_{i,t-1} + \gamma x_{i,t-1} + \mu_t + \delta_i + u_{it}, \tag{40}$$

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. While AJRY consider several estimators we focus on the one-step GMM estimator of Arellano and Bond (1991). This is an over-identified estimator, and standard errors are clustered at the country level, so this fits exactly into the framework of our paper.

The number of countries (127 in the full data set) corresponds to the number of groups G. The sample period is 1960-2000 with observations at either 5-year or 10-year frequencies, so the number of time-series observations n_g (which is the number of observations per group) ranges up to 9, but is heterogeneous as it is an unbalanced panel.

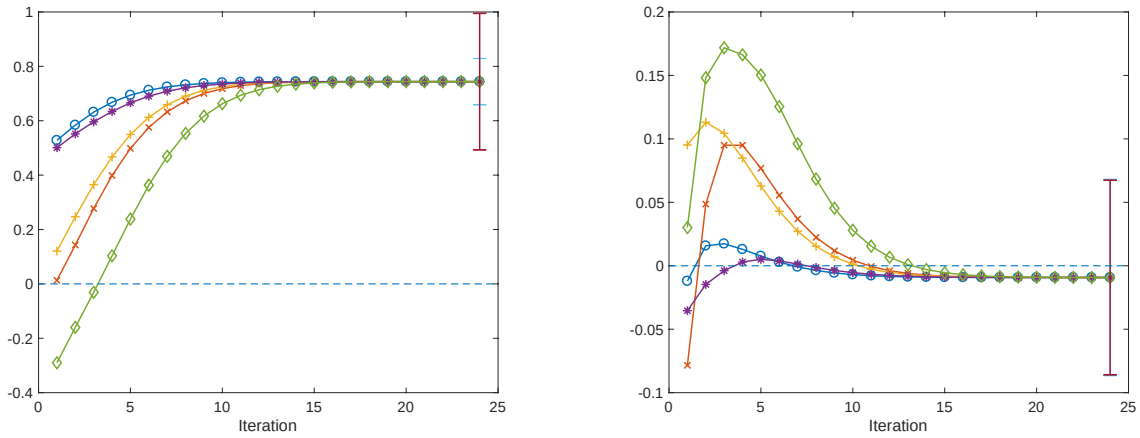
The main results of AJRY are reported in their Table 2, which are their estimates of the above dynamic panel regression model. We repeat their estimates in Table 5 below. Columns I and III are the estimates reported in AJRY (one-step GMM). Columns II and IV are iterated GMM (which are not reported in AJRY). In addition to the coefficient estimates we report Arellano-Bond standard errors (as done by AJRY) which cluster at the country level, and our new misspecification-robust standard errors, also clustered at the country level. We also 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.

The primary focus of AJRY was the coefficient on lagged income and its statistical significance. We focus on two other issues. First, the difference between the one-step and iterated estimates. Second, the difference between the Arellano-Bond and misspecification-robust standard errors.

First, the difference between the one-step and iterated GMM estimates is quite large in some cases, in particular with the five-year data. For example, 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.

It is worth pointing out that the one-step and iterated estimates do not necessarily differ. For example, for the ten-year data the one-step and iterated estimates are quite similar. This shows that the sensitivity depends on the context. However, this is unknown unless both estimates are calculated.

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



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

Figure 1: Convergence path of the iterated GMM estimator in Column II of Table 5 with initial weight matices $\widehat{W}_n(\widetilde{\phi})$ 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.

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 is quite large in some cases, in particular for the iterated estimator. For example, in the five-year data the misspecification-robust standard error on lagged democracy is about three times the Arellano-Bond standard error. This shows that taking into account possible misspecification can make an enormous difference in the standard errors. In other cases, such as for lagged income in the same regression, the two standard errors are quite similar. The fact that the misspecification-robust standard errors can be substantially different suggests that ignoring misspecification bias can lead to large 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, mildly misspecified, or strongly misspecified.

It is noteworthy to point out that the p-values of the J test reported in AJRY are 0.26 and 0.07, which are different from our calculation 0.006 and 0.02 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 one-step GMM criterion with both the moment and efficient

weight matrix being evaluated at the one-step estimate. Since their main conclusions rely on the one-step GMM estimates we believe that ours is a more reasonable choice. In fact, AJRY use 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 is not a reliable test. We recommend using our J statistic (35) evaluated at the estimator of the user.

AJRY conclude that income does not have a statistically significant causal effect on democracy. Our results in Table 5 reinforce and extend this conclusion.

	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	[0.006]	[0.42]	[0.02]	[0.09]
# of Iteration	0	23	0	9
# of Instruments	55		15	
Observations	838		338	
Countries	127		118	

Standard errors clustered by country

Table 5: Extension of Acemoglu, Johnson, Robinson and Yared (2008), Table 2

What are the causes of potential misspecification? One possibility is that the dynamic structure in (40) 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 (40), 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 (40) is $\Delta u_{it} + (\gamma_i - \gamma)\Delta x_{i,t-1}$ which will be correlated with the

instrument $y_{i,t-2}$. A consequence is that the Arellano-Bond standard errors will be incorrect, but our misspecification-robust standard errors will be appropriate.

There is strong evidence for coefficient heterogeneity in equation (40). 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 heterogeneity 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 (40) 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. In this context Arellano-Bond standard errors are incorrect, and our misspecification-robust standard errors appropriate.

	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]	[0.03]	[0.38]	[0.09]	[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 6: Extension of Cervellati, Jung, Sunde, and Vischer (2014), Table 4

To highlight this issue further we examine a key table from CJSV (their Table 4) where they present Arellano-Bond estimates of model (40) 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, such that $c_i = 1$ indicates that the country had

"historically strong institutions". (Acemoglu, Johnson, Robinson and Yared (2009) make a similar distinction, describing the 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 6 below for the five-year sample. CJSV reported one-step Arellano-Bond estimates and standard errors, 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.

Our focus is on the differences between the one-step and iterated GMM estimates, and between the Arellano-Bond and misspecification-robust standard errors.

First, in many cases 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 and thus inherently arbitrary. The iterated GMM estimates are free of the arbitrary choice of initial weight matrix.

Second, in many cases the misspecification-robust standard errors are quite different from the Arellano-Bond standard errors. In some cases they are quite similar, but in other cases the misspecification-robust standard errors are three to four times as large. Inferences based on the Arellano-Bond standard errors will be misleadingly precise in such contexts.

If we examine the over-identification J tests, we find that three (four) of the six p-values are significant at the 5% (10%) level, indicating potential misspecification. Our recommendation is that in this context the misspecification-robust standard errors are more reliable choice for inference. Note that the reported p-values of the J test in CJSV are different from our calculation for the same reason given in the AJRY analysis.

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 inferences which are not fragile to specification choices.

Appendix

Proof of Theorem 1: To show that the map $g_n(\phi)$ is a contraction, we show that for $c = 2kC^5\delta < 1$, $\|g_n(\phi_1) - g_n(\phi_2)\| \leq c\|\phi_1 - \phi_2\|$ for all $\phi_1, \phi_2 \in \Theta$. By the Banach fixed point theorem this implies that the fixed point θ_n exists and is unique.

$g_n(\phi)$ minimizes $J_n(\theta, \phi)$ and thus is the θ which solves the first-order condition

$$0 = \frac{\partial}{\partial \theta} J_n(\theta, \phi) = 2Q_n(\theta)' W_n(\phi)^{-1} m_n(\theta) \tag{41}$$

where $Q_n(\theta) = \frac{\partial}{\partial \theta'} m_n(\theta)$. Since (41) is continuously differentiable under Assumptions 1.5 and 1.6, and

$$\frac{\partial}{\partial \theta'} \frac{\partial}{\partial \theta} J_n(\theta, \phi)|_{\theta = g_n(\phi)} = D_n(g_n(\phi), \phi)$$

is uniformly invertible under Assumption 1.3, it follows by the implicit function theorem that $g_n(\phi)$ exists, is continuously differentiable, and its derivative equals

$$V_n(\phi) = \frac{\partial}{\partial \phi'} g_n(\phi) = -D_n(g_n(\phi), \phi)^{-1} B_n(g_n(\phi), \phi) \tag{42}$$

where

$$B_n(\theta, \phi) = \frac{\partial}{\partial \phi'} \frac{\partial}{\partial \theta} J_n(\theta, \phi).$$

We calculate that

$$\begin{aligned} B_n(\theta, \phi) &= 2\left[m_n(\theta)' \otimes Q_n(\theta)'\right] \frac{\partial}{\partial \phi'} \text{vec}\left(W_n(\phi)^{-1}\right) \\ &= -2\left[m_n(\theta)' \otimes Q_n(\theta)'\right] [W_n(\phi)^{-1} \otimes W_n(\phi)^{-1}] S_n(\phi) \end{aligned} \tag{43}$$

where $S_n(\phi) = \frac{\partial}{\partial \phi'} \text{vec} W_n(\phi)$.

Assumptions 1.1, 1.5 and 1.6 imply that $\|Q_n(\theta)\| \leq C$ and $\|S_n(\theta)\| \leq C$ for some $C < \infty$. Assumption 1.2 implies $\left\|W_n(\phi)^{-1}\right\| \leq C$. Assumption 1.4 implies $\|m_n(g_n(\phi))\| \leq \delta$. Together these imply

$$\|B_n(g_n(\phi), \phi)\| \leq 2\|Q_n(g_n(\phi))\| \|S_n(\phi)\| \left\|W_n(\phi)^{-1}\right\|^2 \|m_n(g_n(\phi))\| \leq 2C^4\delta. \tag{44}$$

Let $[A]_j$ denote the j^{th} row of a matrix A and let $\|A\|_F = \sqrt{\text{tr}\,(A'A)}$ denote the Frobenius norm. Using the properties of the Frobenius norm and Assumption 1.3,

$$\left\|[D_n(g_n(\phi), \phi)^{-1}]_j\right\| \leq \|D_n(g_n(\phi), \phi)^{-1}\|_F \leq \sqrt{k}\lambda_{\max}(D_n(g_n(\phi), \phi)^{-1}) \leq \sqrt{k}C. \tag{45}$$

Let g_{nj} denote the j^{th} element of g_n. Using the definition of the Euclidean norm, element-by-element Taylor series expansions, where ϕ_j^* lie on the line segment joining ϕ_1 and ϕ_2, the Schwarz

inequality, (42), the Schwarz matrix inequality, (44), (45), and $c = 2kC^5\delta$,

$$\begin{aligned}
\|g_n(\phi_1) - g_n(\phi_2)\|^2 &= \sum_{j=1}^{k} |g_{nj}(\phi_1) - g_{nj}(\phi_2)|^2 \\
&= \sum_{j=1}^{k} \left| \left[V(\phi_j^*)\right]_j (\phi_1 - \phi_2) \right|^2 \\
&\leq \sum_{j=1}^{k} \left\| \left[V(\phi_j^*)\right]_j \right\|^2 \|\phi_1 - \phi_2\|^2 \\
&= \sum_{j=1}^{k} \left\| [D_n(g_n(\phi_j^*), \phi_j^*)]_j^{-1} B_n(g_n(\phi_j^*), \phi_j^*) \right\|^2 \|\phi_1 - \phi_2\|^2 \\
&\leq \sum_{j=1}^{k} \left\| [D_n(g_n(\phi_j^*), \phi_j^*)]_j^{-1} \right\|^2 \|B_n(g_n(\phi_j^*), \phi_j^*)\|^2 \|\phi_1 - \phi_2\|^2 \\
&\leq 4k^2 C^{10} \delta^2 \|\phi_1 - \phi_2\|^2 \\
&= c^2 \|\phi_1 - \phi_2\|^2
\end{aligned}$$

where $c < 1$. This establishes that the map $g_n(\phi)$ is a contraction as required. ■

Proof of Theorem 2.1: Define $\overline{J}_n(\theta, \phi) = \overline{m}_n(\theta)' \overline{W}_n(\phi)^{-1} \overline{m}_n(\theta)$. Since $g_n(\phi)$ minimizes $J_n(\theta, \phi)$, and $\overline{g}_n(\phi)$ minimizes $\overline{J}_n(\theta, \phi)$

$$\begin{aligned}
0 &\leq J_n(\overline{g}_n(\phi), \phi) - J_n(g_n(\phi), \phi) \\
&= J_n(\overline{g}_n(\phi), \phi) - \overline{J}_n(\overline{g}_n(\phi), \phi) - J_n(g_n(\phi), \phi) + \overline{J}_n(\overline{g}_n(\phi), \phi) \\
&\leq J_n(\overline{g}_n(\phi), \phi) - \overline{J}_n(\overline{g}_n(\phi), \phi) - J_n(g_n(\phi), \phi) + \overline{J}_n(g_n(\phi), \phi) \\
&\leq 2 \sup_{\phi, \theta} \left\| \overline{J}_n(\theta, \phi) - J_n(\theta, \phi) \right\| \to_p 0,
\end{aligned}$$

where the final convergence by Assumption 2 (9) and (12) plus Assumption 1. This implies

$$\sup_{\phi} |J_n(\overline{g}_n(\phi), \phi) - J_n(g_n(\phi), \phi)| \to_p 0.$$

Fix $\varepsilon > 0$. Under Assumption 1.3, $g_n(\phi)$ uniquely minimizes $J_n(\theta, \phi)$, so we can find a $\eta > 0$ such that for all ϕ, $\|g_n(\phi) - \theta\| > \varepsilon$ implies $|J_n(g_n(\phi), \phi) - J_n(\theta, \phi)| > \eta$. Thus

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

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

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

as required. ■

Proof of Theorem 2.2: 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

$$\|\overline{g}_n(\phi_1) - \overline{g}_n(\phi_2)\| \leq c \|\phi_1 - \phi_2\| \tag{46}$$

for all $\phi_1, \phi_2 \in \Theta$. Dominitz and Sherman (2005) Lemma 3 show that sufficient conditions for (46) to hold with probability tending to one as $n \to \infty$ are that (i) $g_n(\phi)$ is a contraction mapping (established in Theorem 1); (ii) $\sup_\phi \|\overline{g}_n(\phi) - g_n(\phi)\| \to_p 0$ (established in part 1); and (iii) $\sup_\phi \left\|\overline{V}_n(\phi) - V_n(\phi)\right\| \to_p 0$ where $V_n(\phi) = \frac{\partial}{\partial \phi'} g_n(\phi)$ and $\overline{V}_n(\phi) = \frac{\partial}{\partial \phi'} \overline{g}_n(\phi)$. Hence it is sufficient to verify this final condition.

Recall that $V_n(\phi)$ can be expressed as (42) where $B_n(\theta, \phi)$ equals (43). We can calculate that

$$D_n(\theta, \phi) = 2\left\{Q_n(\theta)' W_n(\phi)^{-1} Q_n(\theta) + \left(m_n(\theta)' W_n(\phi)^{-1} \otimes I\right) R_n(\theta)\right\}.$$

Similarly,

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

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

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

Assumption 1 and 2 imply that $\overline{B}_n(\theta, \phi) - B_n(\theta, \phi)$ and $\overline{D}_n(\theta, \phi) - D_n(\theta, \phi)$ converge uniformly to 0. Part 1 shows that $\overline{g}_n(\phi) - g_n(\phi)$ converges uniformly to 0. Together, this implies that $\overline{V}_n(\phi) - V_n(\phi)$ converges uniformly to 0, as required. ■

Proof of Theorem 2.3: Dominitz and Sherman (2005), Theorem 2, show that if $s(n) \to \infty$ then $\left\|\widehat{\theta}_{s(n)} - \theta_n\right\| \to_p 0$ since $g_n(\phi)$ is a contraction mapping (Theorem 1) and $\sup_\phi \|\overline{g}_n(\phi) - g_n(\phi)\| \to_p 0$ (Theorem 2.1). Combined with Theorem 2.2 we find

$$\left\|\widehat{\theta} - \theta_n\right\| \leq \left\|\widehat{\theta}_{s(n)} - \theta_n\right\| + \left\|\widehat{\theta} - \widehat{\theta}_{s(n)}\right\| \to_p 0.$$

■

Proof of Theorem 3: We first show that Assumption 3 implies the convergence results of Assumption 2.

First take $\overline{m}_n(\theta)$, $\overline{Q}_n(\theta)$, and $\overline{R}_n(\theta)$, which are sample means. (9), (10), and (11) follow by the ULLN for clustered means established by Hansen and Lee (2017, Theorem 5), which holds for random variables which are uniformly integrable, Lipschitz, and cluster sizes satisfy $\max_{g \leq G} n_g / n \to 0$. The uniform integrability holds by Assumption 3.4, the Lipschitz condition by Assumption 3.5 and the cluster size condition is implied by Assumption 3.7 or 3.8.

Second take $\overline{W}_n(\theta)$. If the weight matrix is unclustered (14) then Assumption 3.4 implies that $\|W_n(\theta)\| \leq C$ for some $C < \infty$. If the weight matrix is clustered (15) then $\|W_n(\theta)\| \leq C$ is direct from Assumption 3.8 (c). By Theorem 6 of Hansen and Lee (2017),

$$\sup_{\theta\in\Theta} \left\|\overline{W}_n(\theta) - W_n(\theta)\right\| \leq C \cdot \sup_{\theta\in\Theta} \left\|W_n(\theta)^{-1/2}\overline{W}_n(\theta)W_n(\theta)^{-1/2} - I_l\right\| \to 0.$$

This is (12).

Third, take $\overline{S}_n(\theta)$. It can be written as

$$\overline{S}_n(\theta) = \begin{cases} \frac{1}{n}\sum_{i=1}^n \left(v(X_i,\theta)\otimes U(X_i,\theta) + U(X_i,\theta)\otimes v(X_i,\theta)\right) & \text{under (14)} \\ \frac{1}{n}\sum_{i=1}^n \left(\widetilde{v}_g(\theta)\otimes \widetilde{U}_g(\theta) + \widetilde{U}_g(\theta)\otimes \widetilde{v}_g(\theta)\right) & \text{under (15)} \end{cases}$$

where $\widetilde{v}_g(\theta) = \sum_{j=1}^{n_g} v(X_{gj},\theta)$ and $\widetilde{U}_g(\theta) = \sum_{j=1}^{n_g} U(X_{gj},\theta)$. This is a subset of a larger matrix which stacks $v(X_i,\theta)$ and $\operatorname{vec} U(X_i,\theta)$. Applying the same argument as for $\overline{W}_n(\theta)$ we find (13).

The conditions of Theorem 2 are satisfied so we conclude that $\left\|\widehat{\theta} - \theta_n\right\| \to_p 0$.

We next justify the expansion (16) in the text. It is convenient to note that 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

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

The chain rule then yields (16). Similarly, we define the population analog

$$F_n(\theta) = Q_n(\theta)'W_n(\theta)^{-1}m_n(\theta)$$

and its derivative

$$\begin{aligned} \frac{\partial}{\partial\theta'}F_n(\theta) &= Q_n(\theta)'W_n(\theta)^{-1}Q_n(\theta) + \left(m_n(\theta)'W_n(\theta)^{-1}\otimes I_k\right)R_n(\theta) \\ &\quad - \left(m_n(\theta)'W_n(\theta)^{-1}\otimes Q_n(\theta)'W_n(\theta)^{-1}\right)S_n(\theta) \\ &\equiv H_n(\theta). \end{aligned}$$

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

Instead of (17) we use the exact expansion

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

where the j^{th} row of H_n^* is the j^{th} row of $\overline{H}_n(\theta_{nj})$ where θ_{nj} is on the line segment joining $\widehat{\theta}$ and

θ_n. This implies

$$\sqrt{n}\left(\widehat{\theta}-\theta_n\right) = -H_n^{*-1}\sqrt{n}\,\overline{F}_n(\theta_n).$$

The convergence results in Assumption 2 (which hold as discussed above) plus Assumption 1 imply that

$$\sup_{\theta\in\Theta}\left\|\overline{H}_n(\theta) - H_n(\theta)\right\| \to_p 0 \tag{47}$$

and that $H_n(\theta)$ is uniformly continuous in θ. Together with $\left\|\widehat{\theta}-\theta_n\right\| \to_p 0$ we obtain

$$\|H_n^* - H_n\| \to_p 0. \tag{48}$$

We next justify equation (19) from the text. First, by the convergence results in Assumption 2 and $0 = Q_n'W_n^{-1}\mu_n$, the left-hand side of (19) can be written as

$$\begin{aligned}
\sqrt{n}\,\overline{F}_n(\theta_n) &= \sqrt{n}\,\overline{Q}_n'\overline{W}_n^{-1}\overline{m}_n \\
&= \sqrt{n}\,Q_n'\overline{W}_n^{-1}\mu_n + \left(Q_n+\overline{Q}_n-Q_n\right)'\overline{W}_n^{-1}\sqrt{n}\left(\overline{m}_n-\mu_n\right) + \sqrt{n}\left(\overline{Q}_n-Q_n\right)'\overline{W}_n^{-1}\mu_n \\
&= \sqrt{n}\,Q_n'\overline{W}_n^{-1}\mu_n + Q_n'\overline{W}_n^{-1}\sqrt{n}\left(\overline{m}_n-\mu_n\right)(1+o_p(1)) + \sqrt{n}\left(\overline{Q}_n-Q_n\right)'\overline{W}_n^{-1}\mu_n \\
&= \sqrt{n}\,Q_n'\overline{W}_n^{-1}\mu_n + \left(Q_n'W_n^{-1}\sqrt{n}\left(\overline{m}_n-\mu_n\right) + \sqrt{n}\left(\overline{Q}_n-Q_n\right)'W_n^{-1}\mu_n\right)(1+o_p(1)) \\
&= \sqrt{n}\,\left(Q_n'\overline{W}_n^{-1}\mu_n + Q_n'W_n^{-1}\overline{m}_n + \overline{Q}_n'W_n^{-1}\mu_n\right)(1+o_p(1))
\end{aligned}$$

the final using the identify $Q_n'W_n^{-1}\mu_n = 0$

Second, using the identity

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

and a Taylor expansion we find

$$\begin{aligned}
\sqrt{n}\,Q_n'\overline{W}_n^{-1}\mu_n &= \sqrt{n}\left(\mu_n'\otimes Q_n'\right)\mathrm{vec}\overline{W}_n^{-1} \\
&= \sqrt{n}\left(\mu_n'\otimes Q_n'\right)\mathrm{vec}W_n^{-1} - \left(\mu_n'\otimes Q_n'\right)\left(W_n^{-1}\otimes W_n^{-1}\right)\sqrt{n}\mathrm{vec}\left(\overline{W}_n - W_n\right)(1+o_p(1)) \\
&= \sqrt{n}\,Q_n'W_n^{-1}\mu_n - Q_n'W_n^{-1}\sqrt{n}\left(\overline{W}_n - W_n\right)W_n^{-1}\mu_n\,(1+o_p(1)) \\
&= \sqrt{n}\,\left(Q_n'W_n^{-1}\mu_n - Q_n'W_n^{-1}\overline{W}_nW_n^{-1}\mu_n\right)(1+o_p(1))
\end{aligned}$$

Together, these expansions lead to (19). Note that the convergence rates of $\overline{m}_n$, $\overline{Q}_n$, and $\overline{W}_n$ are non-standard (may even be slower than $n^{-1/4}$) so that conventional expansion arguments are not appropriate to show (19).

Equation (20) is an algebraic equivalence. We have established that

$$\begin{aligned}
&\left(H_n^{-1}\Omega_n H_n^{-1}\right)^{-1/2}\sqrt{n}\left(\widehat{\theta}-\theta_n\right) \\
&= -\left(H_n^{-1}\Omega_n H_n^{-1}\right)^{-1/2} H_n^{*-1}\sqrt{n}\,\overline{F}_n(\theta_n) \\
&= -\left(H_n^{-1}\Omega_n H_n^{-1}\right)^{-1/2} H_n^{*-1}\sqrt{n}\,\widetilde{F}_n\left(1+o_p(1)\right) \\
&= -\left(H_n^{-1}\Omega_n H_n^{-1}\right)^{-1/2} H_n^{-1}\left(\frac{1}{\sqrt{n}}\sum_{g=1}^{G}\widetilde{\psi}_g\right)\left(1+o_p(1)\right) \qquad (49) \\
&- \left(H_n^{-1}\Omega_n H_n^{-1}\right)^{-1/2}\left(H_n^{*-1}-H_n^{-1}\right)\left(\frac{1}{\sqrt{n}}\sum_{g=1}^{G}\widetilde{\psi}_g\right)\left(1+o_p(1)\right). \qquad (50)
\end{aligned}$$

Note that we can write $\widetilde{\psi}_g = D_n'\widetilde{f}_g$ where

$$D_n = \begin{bmatrix} W_n^{-1}Q_n \\ W_n^{-1}\mu_n \otimes I_k \\ -W_n^{-1}\mu_n \otimes W_n^{-1}Q_n \end{bmatrix}. \qquad (51)$$

and

$$\widetilde{f}_g(\theta) = \begin{bmatrix} \widetilde{m}_g(\theta) \\ \mathrm{vec}(\widetilde{Q}_g'(\theta) \\ \mathrm{vec}(\widetilde{W}_g) \end{bmatrix}. \qquad (52)$$

The cluster sums $\widetilde{f}_g$ are independent across g. Assumptions 3.2, 3.4, 3.6, 3.8 imply the assumptions for the CLT of Corollary 1 of Hansen and Lee (2017). Thus (49) converges in distribution to $N(0, I_k)$.

The proof is completed by showing that (50) is $o_p(1)$. $\lambda_{\min}(H_n) \geq C^{-1}$ and (48) imply that

$$\left\|H_n^{-1/2}H_n^*H_n^{-1/2} - I_k\right\| = \left\|H_n^{-1/2}\left(H_n^* - H_n\right)H_n^{-1/2}\right\| \leq C\left\|H_n^* - H_n\right\| \xrightarrow{p} 0.$$

Applying the continuous mapping theorem we find

$$\left\|H_n^{1/2}H_n^{*-1}H_n^{1/2} - I_k\right\| \xrightarrow{p} 0. \qquad (53)$$

The CLT also shows that

$$\Omega_n^{-1/2}\frac{1}{\sqrt{n}}\sum_{g=1}^{G}\widetilde{\psi}_g \to_d N(0, I_k)$$

and is thus $O_p(1)$. Thus (50) is bounded by $O_p(1)$ multiplied by

$$\begin{aligned}
&\left\| \left(H_n^{-1}\Omega_n H_n^{-1}\right)^{-1/2} \left(H_n^{*-1} - H_n^{-1}\right) \Omega_n^{1/2} \right\| \\
&= \left\| \left(H_n^{-1}\Omega_n H_n^{-1}\right)^{-1/2} H_n^{-1/2} \left(H_n^{1/2} H_n^{*-1} H_n^{1/2} - I_k\right) H_n^{-1/2}\Omega_n^{1/2} \right\| \\
&= \left\| \left(H_n^{-1}\Omega_n H_n^{-1}\right)^{-1/2} H_n^{-1}\Omega_n H_n^{-1} \left(H_n^{-1}\Omega_n H_n^{-1}\right)^{-1/2} \right\|^{1/2} \left\| H_n^{1/2} H_n^{*-1} H_n^{1/2} - I_k \right\| \\
&= o_p(1).
\end{aligned}$$

Hence (50) is $o_p(1)$. This completes the proof. ■

Proof of Theorem 4: Given Assumptions 3.1 and 3.2 and (53), it is sufficient to show that

$$\left\| \widehat{\Omega} - \Omega_n \right\| \to_p 0. \tag{54}$$

Define $\widetilde{f}_g(\theta)$ as in (52),

$$\begin{aligned}
\widehat{D} &= \begin{bmatrix} \widehat{W}^{-1}\widehat{Q} \\ \widehat{W}^{-1}\widehat{\mu} \otimes I_k \\ -\widehat{W}^{-1}\widehat{\mu} \otimes \widehat{W}^{-1}\widehat{Q} \end{bmatrix} \\
\widetilde{G}(\theta) &= \frac{1}{n}\sum_{g=1}^{G} \widetilde{f}_g(\theta)\widetilde{f}_g(\theta)'
\end{aligned} \tag{55}$$

so that $\widehat{\Omega} = \widehat{D}'\widetilde{G}(\widehat{\theta})\widehat{D}$. To establish (54) we can replace $\widehat{D}$ with D_n defined in (51) (using the convergence results in Assumption 2 which are implied by Assumption 3) and then rotate out D_n. Since $\left\| \widehat{\theta} - \theta_n \right\| \to_p 0$ it is sufficient to show that

$$\sup_{\theta \in \mathcal{N}} \left\| \widetilde{G}(\theta) - E\widetilde{G}(\theta) \right\| \to_p 0 \tag{56}$$

for a neighborhood $\mathcal{N}$ of θ_n.

In the non-clustered weight matrix case, $\widetilde{f}_g(\theta)$ is a vector of cluster sums. (The third component of $\widetilde{f}_g(\theta)$ is $\text{vec}(\widetilde{W}_g) = \sum_{j=1}^{n_g} v(X_{gj},\theta) \otimes v(X_{gj},\theta)$.) Thus we can appeal to the ULLN for clustered variances of Hansen and Lee (2017, Theorem 6), for which the assumptions listed in Assumption 3 are sufficient, establishing (56).

In the clustered weight matrix case, the third component of $\widetilde{f}_g(\theta)$ is $\text{vec}(\widetilde{W}_g) = \widetilde{v}_g(\theta) \otimes \widetilde{v}_g(\theta)$ which is not a cluster sum but rather a product of cluster sums, so the results of Hansen and Lee (2017, Theorem 6) do not apply. Andrews (1992, Theorem 3) shows that (56) holds if for all $\theta \in \mathcal{N}$

$$\left\| \widetilde{G}(\theta) - G_n(\theta) \right\| \to_p 0, \tag{57}$$

and for all θ_1, $\theta_2 \in \mathcal{N}$,

$$\left\| \widetilde{f}_g(\theta_1)\widetilde{f}_g(\theta_1)' - \widetilde{f}_g(\theta_2)\widetilde{f}(\theta_2)' \right\| \leq A_g h(\|\theta_1 - \theta_2\|) \tag{58}$$

with $h(u) \downarrow 0$ as $u \downarrow 0$ and $\frac{1}{n}\sum_{g=1}^{G} EA_g \leq A < \infty$. We now establish (57) and (58).

Take (57). Fix $\theta \in \mathcal{N}$. For brevity, suppress the dependence of $\widetilde{f}_g(\theta)$ and $\widetilde{G}(\theta)$ on θ. Fix $\delta > 0$. Set $\varepsilon = (\delta/C)^2$. Define $\widetilde{h}_g = \widetilde{f}_g 1\left(\left\|\widetilde{f}_g\right\|^2 \leq n\varepsilon\right)$. Then

$$\widetilde{G} = \frac{1}{n}\sum_{g=1}^{G} \widetilde{h}_g\widetilde{h}_g' + \frac{1}{n}\sum_{g=1}^{G} \widetilde{f}_g\widetilde{f}_g' 1\left(\left\|\widetilde{f}_g\right\|^2 > n\varepsilon\right).$$

By the triangle inequality

$$E\left\|\widetilde{G} - E\widetilde{G}\right\| = \frac{1}{n} E\left\|\sum_{g=1}^{G}\left(\widetilde{h}_g\widetilde{h}_g' - E\widetilde{h}_g\widetilde{h}_g'\right)\right\| \tag{59}$$

$$+ \frac{2}{n}\sum_{g=1}^{G} E\left(\left\|\widetilde{f}_g\right\|^2 1\left(\left\|\widetilde{f}_g\right\|^2 > n\varepsilon\right)\right). \tag{60}$$

Take (59). Assumption 3.4 and the C_r inequality allow us to deduce that $E\widetilde{v}_g^4 \leq Cn_g^4$ and $E\widetilde{f}_g^2 \leq Cn_g^4$ for some $C < \infty$. Using Jensen's inequality, the assumption the clusters are independent and thus uncorrelated, the bounds $\left\|\widetilde{h}_g\right\|^2 \leq n\varepsilon$ and $\left\|\widetilde{h}_g\right\|^2 \leq \left\|\widetilde{f}_g\right\|^2$, and the definition of ε, we obtain that (59) is bounded by

$$\frac{1}{n}\left(\sum_{g=1}^{G} E\left\|\widetilde{h}_g\right\|^4\right)^{1/2} \leq \varepsilon^{1/2} C^{1/2}\left(\frac{1}{n}\sum_{g=1}^{G} n_g^4\right)^{1/2} \leq \delta.$$

Take (60). Write $\widetilde{q}_g = \left(\widetilde{m}_g, \operatorname{vec}\widetilde{Q}_g\right)$. Using the inequality

$$(A+B)1(A+B>\varepsilon) \leq 2A1(A>\varepsilon/2) + 2B1(B>\varepsilon/2),$$

we find that (60) is bounded by 8 times

$$\frac{1}{n}\sum_{g=1}^{G}\left(E\left(\|\widetilde{q}_g\|^2 1\left(\|\widetilde{q}_g\|^2 > \frac{n\varepsilon}{2}\right)\right) + E\left(\|\widetilde{v}_g\|^4 1\left(\|\widetilde{v}_g\|^4 > \frac{n\varepsilon}{2}\right)\right)\right). \tag{61}$$

Lemma 1 of Hansen and Lee (2017) implies that $\left\|n_g^{-1}\widetilde{q}_g\right\|^2$ and $\left\|n_g^{-1}\widetilde{v}_g\right\|^4$ are uniformly integrable, given Assumption 3.4. This means we can pick B sufficiently large so that

$$\sup_g E\left(\left\|n_g^{-1}\widetilde{q}_g\right\|^2 1\left(\left\|n_g^{-1}\widetilde{q}_g\right\|^2 > B\right)\right) \leq \frac{\delta}{C}$$

and

$$\sup_g E\left(\left\|n_g^{-1}\widetilde{v}_g\right\|^4 1\left(\left\|n_g^{-1}\widetilde{\nu}_g\right\|^2 > B\right)\right) \leq \frac{\delta}{C}.$$

Pick n large enough so that

$$\max_{g\leq G}\frac{n_g}{n^{1/2}} \leq \max_{g\leq G}\frac{n_g^2}{n^{1/2}} \leq \frac{(\varepsilon/2)^{1/2}}{B}$$

which is feasible by Assumption 3.8(b). Then (61) is bounded by

$$\frac{1}{n}\sum_{g=1}^{G}\left(E\left(\left\|\widetilde{q}_g\right\|^2 1\left(\left\|n_g^{-1}\widetilde{q}_g\right\|^2 > B\right)\right) + E\left(\left\|\widetilde{v}_g\right\|^4 1\left(\left\|n_g^{-1}\widetilde{v}_g\right\|^2 > B\right)\right)\right) \leq \frac{1}{n}\sum_{g=1}^{G}\left(n_g^2 + n_g^4\right)\frac{\delta}{C} \leq 2\delta.$$

We have shown that $E\left\|\widetilde{G} - E\widetilde{G}\right\| \leq 17\delta$. Since δ is arbitrary, by Markov's inequality, (57) is shown.

Take (58). Fix any $\theta_1, \theta_2 \in \mathcal{N}$. Set $\widetilde{f}_g = \sup_{\theta\in\mathcal{N}}\left\|\widetilde{f}_g(\theta)\right\|$ and $\widetilde{v}_g = \sup_{\theta\in\mathcal{N}}\|\widetilde{v}_g(\theta)\|$. Using the triangle inequality and Assumption 3.5

$$\begin{aligned}
\|\widetilde{m}_g(\theta_2) - \widetilde{m}_g(\theta_1)\| &\leq \sum_{j=1}^{n_g} A_m(X_{gj})h\left(\|\theta_1 - \theta_2\|\right) \\
\left\|\widetilde{Q}_g(\theta_2) - \widetilde{Q}_g(\theta_1)\right\| &\leq \sum_{j=1}^{n_g} A_Q(X_{gj})h\left(\|\theta_1 - \theta_2\|\right) \\
\|\widetilde{v}_g(\theta_2) - \widetilde{v}_g(\theta_1)\| &\leq \sum_{j=1}^{n_g} A_v(X_{gj})h\left(\|\theta_1 - \theta_2\|\right).
\end{aligned}$$

Using the C_r inequality and definition (52),

$$\begin{aligned}
\left\|\widetilde{f}_g(\theta_2) - \widetilde{f}_g(\theta_1)\right\| &\leq \|\widetilde{m}_g(\theta_2) - \widetilde{m}_g(\theta_1)\| + \left\|\widetilde{Q}_g(\theta_2) - \widetilde{Q}_g(\theta_1)\right\| + 2\widetilde{v}_g\|\widetilde{v}_g(\theta_2) - \widetilde{v}_g(\theta_1)\| \\
&\leq \left(\sum_{j=1}^{n_g}\left(A_m(X_{gj}) + A_Q(X_{gj}) + 2\widetilde{v}_g A_v(X_{gj})\right)\right) h\left(\|\theta_1 - \theta_2\|\right).
\end{aligned}$$

The left-hand-side of (58) is bounded by

$$2\widetilde{f}_g\left\|\widetilde{f}_g(\theta_2) - \widetilde{f}(\theta_1)\right\| \leq 2\widetilde{f}_g\left(\sum_{j=1}^{n_g}\left(A_m(X_{gj}) + A_Q(X_{gj}) + 2\widetilde{v}_g A_v(X_{gj})\right)\right) h\left(\|\theta_1 - \theta_2\|\right).$$

Hence (58) holds with

$$A_g = 2\widetilde{f}_g\left(\sum_{j=1}^{n_g}\left(A_m(X_{gj}) + A_Q(X_{gj}) + 2\widetilde{v}_g A_v(X_{gj})\right)\right).$$

It remains to show that $\frac{1}{n}\sum_{g=1}^{G} EA_g \leq A < \infty$. Assumption 3.4 allows us to deduce that $E\widetilde{v}_g^4 \leq$

Cn_g^4 and $E\widetilde{f}_g^2 \leq Cn_g^4$. Then applying Holder's inequality

$$\begin{aligned}
EA_g &= 2\sum_{j=1}^{n_g}\left(E\left(\widetilde{f}_g A_m(X_{gj})\right) + \left(\widetilde{Ef}_g A_Q(X_{gj})\right) + 2E\left(\widetilde{f}_g\widetilde{v}_g A_v(X_{gj})\right)\right) \\
&\leq 2\sum_{j=1}^{n_g}\left(E\widetilde{f}_g^2\right)^{1/2}\left(\left(EA_m^2(X_{gj})\right)^{1/2} + \left(EA_Q^2(X_{gj})\right)^{1/2} + 2\left(E\widetilde{v}_g^4\right)^{1/4}\left(E\left(A_v^4(X_{gj})\right)\right)^{1/4}\right) \\
&\leq 2C\sum_{j=1}^{n_g}\left(2n_g^2 + n_g^3\right) \\
&\leq 6Cn_g^4.
\end{aligned}$$

Hence

$$\frac{1}{n}\sum_{g=1}^{G} EA_g \leq 6C\frac{1}{n}\sum_{g=1}^{G} n_g^4 \leq 6C^2$$

by Assumption 3.8 (a). This establishes (58).

By showing (57) and (58) we have established (56) and completes the proof. ■

Proof of Theorem 5: We establish a slightly more general result. For any population weight matrix $W_n(\theta)$ set

$$W_n^*(\theta) = W_n(\theta) - C_n \cdot m_n(\theta)m_n(\theta)' \tag{62}$$

for some constant $0 < C_n < \infty$. Let θ_n and θ_n^* be the pseudo-true values under the weight matrices $W_n(\theta)$ and $W_n^*(\theta)$. We will show that $\theta_n = \theta_n^*$.

If $m_n(\theta_n) = 0$ for some θ_n then the model is correctly specified and there is no distinction $W_n(\theta) = W_n^*(\theta)$ and Theorem 5 trivially holds. Assume $m_n(\theta_n) \neq 0$ for all $\theta \in \Theta$. By the Woodbury matrix identity,

$$W_n(\theta)^{-1} = \left[W_n^*(\theta) + C_n \cdot m_n(\theta)m_n(\theta)'\right]^{-1} = W_n^*(\theta)^{-1} - \frac{C_n W_n^*(\theta)^{-1}m_n(\theta)m_n(\theta)'W_n^*(\theta)^{-1}}{1 + C_n \cdot m_n(\theta)'W_n^*(\theta)^{-1}m_n(\theta)}.$$

Hence the population GMM criterion with $W_n(\phi)^{-1}$ evaluated at $\phi = \theta_n^*$ equals

$$\begin{aligned}
m_n(\theta)'W_n(\theta_n^*)^{-1}m_n(\theta) &= m_n(\theta)'W_n^*(\theta_n^*)^{-1}m_n(\theta) \\
&\quad - \frac{C_n \cdot m_n(\theta)'W_n^*(\theta_n^*)^{-1}m_n(\theta_n^*)m_n(\theta_n^*)'W_n^*(\theta_n^*)^{-1}m_n(\theta)}{1 + C_n \cdot m_n(\theta_n^*)'W_n^*(\theta_n^*)^{-1}m_n(\theta_n^*)} \\
&= \left(m_n(\theta)'W_n^*(\theta_n^*)^{-1}m_n(\theta)\right)\left(1 - \rho_n(\theta,\theta_n^*)\frac{C_n J_n^*}{1 + C_n J_n^*}\right)
\end{aligned} \tag{63}$$

where

$$\rho_n(\theta,\theta_n^*) = \frac{\left(m_n(\theta)'W_n^*(\theta_n^*)^{-1}m_n(\theta_n^*)\right)^2}{\left(m_n(\theta)'W_n^*(\theta_n^*)^{-1}m_n(\theta)\right)\left(m_n(\theta_n^*)'W_n^*(\theta_n^*)^{-1}m_n(\theta_n^*)\right)}$$

and $J_n^* = m_n(\theta_n^*)'W_n^*(\theta_n^*)^{-1}m_n(\theta_n^*)$.

Now consider minimization of (63) over θ given fixed θ_n^*. The first term on the right-hand-side of (63) is the GMM criterion with $W_n^*(\phi)$ evaluated at $\phi = \theta_n^*$, which is minimized at θ_n^*. The second-term on the right-hand-side of (63) is minimized by maximizing $\rho_n(\theta, \theta_n^*)$ which is achieved at $\theta = \theta_n^*$ because $\rho_n(\theta, \theta_n^*)$ is a squared correlation. Since both terms are minimized at $\theta = \theta_n^*$ it follows that (63) is minimized at $\theta = \theta_n^*$. But the left-hand-side of (63) is the GMM criterion with $W_n(\phi)$ evalauted at $\phi = \theta_n^*$, so the fact that its minimum is achieved at $\theta = \theta_n^*$ means that θ_n^* is its fixed point. But the fixed point of the GMM criterion with $W_n(\phi)$ is θ_n. Thus $\theta_n^* = \theta_n$ as claimed.

Since

$$\overline{W}_n^*(\theta) = \overline{W}_n(\theta) - C_n \cdot \overline{m}_n(\theta)\overline{m}_n(\theta)'$$

for some $0 < C_n < \infty$ for the sample weight matrices, we apply the same argument to show the invariance of the estimator. ■

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. 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

4. 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.

5. Arellano, Manuel, and Olympia Bover (1995). Another look at the instrumental variable estimation of error-components models. *Journal of Econometrics*, 68, 29-51.

6. Blundell, Richard, and Stephen Bond (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*, 87, 115-143.

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

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

9. 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.

10. Evdokimov, Kirill, and Michal Kolesár (2017). Inference in Instrumental Variable Regression Analysis with Heterogeneous Treatment Effects. Working paper.

11. Hall, Alastair R. (2000). Covariance matrix estimation and the power of the overidentifying restrictions test. *Econometrica*, 68(6), 1517-1527.

12. 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.

13. Hansen, Bruce E. and Seojeong Lee (2017). Asymptotic Theory for Clustered Samples. Working paper.

14. Hansen, Lars Peter (1982). Large sample properties of Generalized Method of Moments estimators. *Econometrica*, 50, 1029-1054.

15. 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.

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

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

18. Kolesár, Michal (2013). Estimation in an Instrumental Variables Model With Treatment Effect Heterogeneity. Working paper.

19. Lee, Seojeong (2017). A Consistent Variance Estimator for 2SLS When Instruments Identify Different LATEs. *Journal of Business & Economic Statistics*, 1-11.

20. 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.

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

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

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