Inference for Iterated GMM Under Misspecification and Clustering

Bruce E. Hansen*
University of Wisconsin

Seojeong Lee†
University of New South Wales

Preliminary: Do Not Cite

Abstract

This paper proposes simple new inference methods for over-identified Generalized Method of Moments (GMM) estimation which correct the standard error bias which arises when moments are possibly misspecified. We focus on the iterated GMM estimator, providing the first rigorous demonstration of existence, and the first distribution theory for iterated GMM under moment misspecification. Our distribution theory is asymptotic, allowing for either independent or clustered samples. Our simulation results show that our methods successfully remove the large biases in inference due to moment 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 finding further supports the conclusion of the former but is in sharp contrast to that of the latter.

*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) advovated 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 should be used as 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 deriviatives, 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 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. The asymptotic theory developed in this paper explicitly allows for clustered dependence. We allow for quite general forms of clustered dependence, allowing for heterogenous 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. Most of the current literature focuses on the linear regression model. The only previous papers of which we are aware which explored a distribution theory for GMM allowing for clustered dependence are Hwang (2016) and Hansen and Lee (2017).

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, we focus on the iterated GMM estimator, which simplifies the analysis by removing the dependence on the specified iteration. This requires a development of a new fixed-point theory which hasn't been used previously in the GMM literature but provides additional insight.

In empirical practice it is conventional to use iterated GMM, meaning that the weight matrix is iterated until convergence. Implicitly, this assumes that an iterated limit exists, or equivalently that GMM iteration is a contraction. To our knowledge the validity of this assertion has never been investigated. We show that under smoothness conditions if the population version of the iteration is a contraction then the sample version is also a contraction so the conventional assertion (that the iterated limit exists) is indeed valid. The condition on the population version holds under correct specification, and it also holds under mild misspecification.

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

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 shows that the iterated GMM estimator defined with the efficient weight matrix is invariant to the weight matrix being constructed with re-centering. Section 5 presents the asymptotic distribution of the GMM estimator. Section 6 discusses covariance matrix estimation. Section 7 discusses the GMM test if over-identifying restrictions. Section 8 describes the results for the linear model. Section 9 presents simulation evidence of the finite sample distributions. Section 10 is an application to the data and model of Acemoglu, Johnson, Robinson, and Yared (2008). Formal proofs are presented in the Appendix.

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

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

Given this notation, the iteration sequence can be written as

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

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

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

Our distribution theory will allow weight matrices which take the form of sample averages

$$\overline{W}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} W(X_i, \theta) \tag{5}$$

for some $l \times l$ function $W(x, \theta)$. A constant weight matrix is obtained by setting $W(x, \theta) = W$. The 2SLS weight matrix is obtained by setting $W(X_i, \theta) = Z_i Z_i'$ for an instrument vector Z_i. The standard efficient weight matrix is obtained by setting $W(x, \theta) = m(x, \theta)m(x, \theta)'$.

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

$$m_n(\theta) = E\left(\overline{m}_n(\theta)\right) \tag{6}$$

$$W_n(\theta) = E\left(\overline{W}_n(\theta)\right). \tag{7}$$

Notice that we write the expectations (6) and (7) as functions of n. This allows heterogeneous distributions, and additionally under cluster sampling with non-homogeneous cluster sizes the weight

matrix (7) 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):

$$g_n(\phi) = \arg\min_{\theta\in\Theta} m_n(\theta)' W_n(\phi)^{-1} m_n(\theta). \tag{8}$$

Definition (8) 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 (8) 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

$$g_n(\theta_n) = \theta_n \tag{9}$$

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 (7) may vary with the sample under cluster sampling.

The existence of the fixed points of (4) and (9) 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 $\|A\| = \sqrt{\lambda_{\max}(A'A)}$ denote the spectral norm.

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

1. Θ *is compact.*
2. $\inf_{\theta\in\Theta} \lambda_{\min}(W_n(\theta)) \geq C^{-1}$.
3. $\inf_{\phi\in\Theta} \lambda_{\min}(D_n(g_n(\phi), \phi)) \geq C^{-1}$.
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 $n \to \infty$*

$$\sup_{\theta\in\Theta} \|\overline{m}_n(\theta) - m_n(\theta)\| \to_p 0 \tag{10}$$

$$\sup_{\theta\in\Theta} \left\|\overline{Q}_n(\theta) - Q_n(\theta)\right\| \to_p 0 \tag{11}$$

$$\sup_{\theta\in\Theta} \left\|\overline{R}_n(\theta) - R_n(\theta)\right\| \to_p 0 \tag{12}$$

$$\sup_{\theta\in\Theta} \left\|\overline{W}_n(\theta) - W_n(\theta)\right\| \to_p 0 \tag{13}$$

$$\sup_{\theta\in\Theta} \left\|\overline{S}_n(\theta) - S_n(\theta)\right\| \to_p 0 \tag{14}$$

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. $\sup_{\phi\in\Theta} \|\widehat{g}(\phi) - g_n(\phi)\| \to_p 0$.
2. *With probability tending to one, the map $\widehat{g}(\phi)$ 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 $\widehat{g}(\phi)$ is uniformly consistent (established in part 1), and similarly its derivative, then $\widehat{g}(\phi)$ 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 $\widehat{g}(\phi)$ and its derivative.

4 Weight Matrix Invariance

Consider efficient weight matrices which set

$$W(x,\theta) = m(x,\theta)m(x,\theta)'$$

or

$$W(x,\theta) = (m(x,\theta) - Em(X_i,\theta))\,(m(x,\theta) - Em(X_i,\theta))'\,.$$

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 non-clustered efficient weight matrix.

Theorem 3. *Suppose that Assumptions 1 and 2 hold for all weight matrices. The pseudo-true value θ_n and iterated GMM estimate $\widehat{\theta}$ are unaffected if the weight matrix uses $W(x,\theta) = m(x,\theta)m(x,\theta)'$ or $W(x,\theta) = (m(x,\theta) - Em(X_i,\theta))\,(m(x,\theta) - Em(X_i,\theta))'$*

Theorem 3 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 estimate is unaffected the criterion function and LR-type test statistics are affected by the choice. For this reason it may be advisable to use the re-centered versions of the weight matrices as these remain variance estimates under moment misspecification under random sampling, and therefore improve the performance of GMM test statistics based on the GMM criterion.

It is important to understand that Theorem 3 applies only to the GMM estimator when iterated until convergence. It does not apply to the s-step estimator.

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

$$\begin{aligned}\frac{\partial}{\partial\theta'}\overline{F}_n(\theta) &= \overline{Q}_n(\theta)'\overline{W}_n(\theta)^{-1}\overline{Q}_n(\theta)+\left(\overline{m}_n(\theta)'\overline{W}_n(\theta)^{-1}\otimes I_k\right)\overline{R}_n(\theta) \\ &\quad -\left(\overline{m}_n(\theta)'\overline{W}_n(\theta)^{-1}\otimes\overline{Q}_n(\theta)'\overline{W}_n(\theta)^{-1}\right)\overline{S}_n(\theta) \\ &\equiv \overline{H}_n(\theta).\end{aligned} \tag{15}$$

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

$$0=\overline{F}_n(\widehat{\theta})\simeq\overline{F}_n(\theta_n)+\overline{H}_n(\theta_n)\left(\widehat{\theta}-\theta_n\right). \tag{16}$$

Thus

$$\sqrt{n}\left(\widehat{\theta}-\theta_n\right)\simeq-\overline{H}_n(\theta_n)^{-1}\sqrt{n}\,\overline{F}_n(\theta_n). \tag{17}$$

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

$$\sqrt{n}\,\overline{F}_n(\theta_n)=\sqrt{n}\widetilde{F}_n(\theta_n)\left(1+o_p(1)\right) \tag{18}$$

where

$$\sqrt{n}\widetilde{F}_n(\theta_n)=Q_n'W_n^{-1}\sqrt{n}\left(\overline{m}_n-\mu_n\right)+\sqrt{n}\left(\overline{Q}_n-Q_n\right)W_n^{-1}\mu_n-Q_n'W_n^{-1}\sqrt{n}\left(\overline{W}_n-W_n\right)W_n^{-1}\mu_n.$$

Our distribution theory will be derived allowing for either independent samples or clustered sampling. 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 are $n=\sum_{g=1}^{G}n_g$. 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 $n_g=1$.

Given this notation, we write $\widetilde{F}_n(\theta_n)$ as a sum across the cluster sums

$$\begin{aligned}\widetilde{m}_g &= \sum_{j=1}^{n_g} m(X_{gj},\theta_n)\\ \widetilde{Q}_g &= \sum_{j=1}^{n_g} Q(X_{gj},\theta_n)\\ \widetilde{W}_g &= \sum_{j=1}^{n_g} W(X_{gj},\theta_n)\end{aligned}$$

so that $\overline{m}_n = \frac{1}{n}\sum_{g=1}^G \widetilde{m}_g$, $\overline{Q}_n = \frac{1}{n}\sum_{g=1}^G \widetilde{Q}_g$, and $\overline{W}_n = \frac{1}{n}\sum_{g=1}^G \widetilde{W}_g$. Define

$$\widetilde{\psi}_g = Q_n' W_n^{-1}\left(\widetilde{m}_g - E\widetilde{m}_g\right) + \left(\widetilde{Q}_g - E\widetilde{Q}_g\right) W_n^{-1}\mu_n - Q_n' W_n^{-1}\left(\widetilde{W}_g - E\widetilde{W}_g\right) W_n^{-1}\mu_n.$$

We then have

$$\sqrt{n}\widetilde{F}_n(\theta_n) = \frac{1}{\sqrt{n}}\sum_{g=1}^G \widetilde{\psi}_g. \tag{19}$$

Together, we have the representation

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

The CLT can be applied to (19), which has variance

$$\Omega_n = \frac{1}{n}\sum_{g=1}^G E\left(\widetilde{\psi}_g\widetilde{\psi}_g'\right). \tag{20}$$

This leads to an asymptotic distribution theory for $\widehat{\theta}$. We now provide regularity conditions and a formal statement. Define

$$H_n = H_n(\theta_n) = Q_n' W_n^{-1} Q_n + \left(\mu_n' W_n^{-1} \otimes I_k\right) R_n - \left(\mu_n' W_n^{-1} \otimes Q_n' W_n^{-1}\right) S_n. \tag{21}$$

Also define $S(x,\theta) = \frac{\partial}{\partial\theta'}\text{vec}W(x,\theta)$.

Assumption 3. *For some $0 < C < \infty$ and $2 \le r < \infty$*

1. $\lambda_{\min}(H_n) \ge C^{-1}$
2. $\lambda_{\min}(\Omega_n) \ge C^{-1}$
3. *For each $\theta \in \Theta$, and $f(x) = \|m(x,\theta)\|^r$, $\|Q(x,\theta)\|^r$, $\|W(x,\theta)\|^r$, $\|R(x,\theta)\|$, and $\|S(x,\theta)\|$*

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

4. *For each* $\theta_1, \theta_2 \in \Theta$ *and* $f(x,\theta) = m(x,\theta)$, $Q(x,\theta)$, $W(x,\theta)$, $R(x,\theta)$ *and* $S(x,\theta)$

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

where $h(u) \downarrow 0$ *as* $u \downarrow 0$, *and* $EA(X_i) \leq C$

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

6. $\frac{1}{n}\left(\sum_{g=1}^{G} n_g^r\right)^{2/r} \leq C$

7. $\max_{g \leq G} n_g^2/n \to 0$

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

$$\left(H_n^{-1}\Omega_n H_n^{-1\prime}\right)^{-1/2}\sqrt{n}\left(\widehat{\theta} - \theta_n\right) \xrightarrow{d} N\left(\mathbf{0}, I_k\right) \tag{22}$$

where Ω_n *and* H_n *are defined in (20) and (21), respectively.*

Theorem 4 provides the asymptotic distribution of the GMM estimator under possible misspecification.

Theorem 4 provides a simple characterization of the asymptotic distribution of the GMM estimator. The asymptotic variance in (22) is quite different from the standard textbook formula

$$(Q'W^{-1}Q)^{-1}\left(Q'W^{-1}\Omega_{11}W^{-1}Q\right)\left(Q'W^{-1}Q\right)^{-1} \tag{23}$$

where Ω_{11} is (20) with $\mu_n = 0$. The difference is due to the terms S and R. The covariance matrix in (22) equals the conventional formula (23) under correct specification but in general they differ.

The asymptotic variance in Theorem 4 differs from the classical formula (23) in two ways. First, the matrix H_n defined in (21) 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 (20) 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 $\mu_n \neq 0$.

Unlike conventional GMM distributional theory Theorem 4 incorporates the effect of weight matrix estimation and the effect of estimation of the derivative matrix Q_n. This appears in the asymptotic covariance matrix through the derivative matrices S_n and R_n. Theorem 4 is also 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 4 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 iid case. An important distinction is that Theorem 4 allows $W_n(\theta)$ to depend on θ and thus includes the iterated GMM

estimator. Theorem 4 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.

Clustered sampling is permitted by Assumptions 3.5-3.7. Assumptions 3.6-3.7 control the degree of heterogeneity allowed in the cluster sizes n_g. They are satisfied if the cluster sizes are bounded or are growing modestly with sample size. The assumptions imply that $G \to \infty$, so this is the "large G" asymptotic framework. The assumptions exclude any single cluster from asymptotically dominating the sample.

The asymptotic distribution (22) 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 4. Construct the derivatives

$$\begin{aligned}
\widehat{Q} &= \frac{\partial}{\partial \theta'}\overline{m}_n(\widehat{\theta}) \\
\widehat{R} &= \frac{\partial}{\partial \theta'}\mathrm{vec}(\overline{Q}_n(\widehat{\theta})') \\
\widehat{S} &= \frac{\partial}{\partial \theta'}\mathrm{vec}(\overline{W}_n(\widehat{\theta}))
\end{aligned}$$

and define the estimates $\widehat{W} = \overline{W}_n(\widehat{\theta})$, $\widehat{\mu} = \overline{m}_n(\widehat{\theta})$ and

$$\widehat{H} = \overline{H}_n(\widehat{\theta}) = \widehat{Q}'\widehat{W}^{-1}\widehat{Q} + (\widehat{\mu}'\widehat{W}^{-1} \otimes I_k)\widehat{R} - (\widehat{\mu}'\widehat{W}^{-1} \otimes \widehat{Q}'\widehat{W}^{-1})\widehat{S}. \tag{24}$$

Construct the cluster sums

$$\widehat{\psi}_g = \sum_{j=1}^{n_g}\left(\widehat{Q}'\widehat{W}^{-1}m(X_{gj},\widehat{\theta}) + Q(X_{gj},\widehat{\theta})'\widehat{W}^{-1}\widehat{\mu} - \widehat{Q}'\widehat{W}^{-1}W(X_{gj},\widehat{\theta})\widehat{W}^{-1}\widehat{\mu}\right),$$

and the cluster variance estimator

$$\widehat{\Omega} = \frac{1}{n}\sum_{g=1}^{G}\widehat{\psi}_g\widehat{\psi}_g' \tag{25}$$

and

$$\widehat{V} = \widehat{H}^{-1}\widehat{\Omega}\widehat{H}^{-1\prime}. \tag{26}$$

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 22 of V_n by $\widehat{V}$ does not affect the asymptotic distribution. We require a slight strengthening of the Lipschitz condition in Assumption 3.4.

Assumption 4. *Assumption 3.4 holds for $m(x,\theta)$, $Q(x,\theta)$, and $W(x,\theta)$ with $EA(X_i)^2 \leq C$.*

Theorem 5. *Under Assumptions 1, 3, and 4*

$$\left\| V_n^{-1/2} \widehat{V} V_n^{-1/2} - I_k \right\| \to_p 0 \tag{27}$$

and

$$\widehat{V}^{-1/2} \sqrt{n} \left(\widehat{\theta} - \theta_n \right) \xrightarrow{d} N\left(\mathbf{0}, I_k\right) \tag{28}$$

as $n \to \infty$.

Equation (27) shows that $\widehat{V}$ is consistent in a sense appropriate when the variance matrix may not be converging with n. Equation (28) implies that test statistics constructed with $\widehat{V}$ have standard asynmptotic distributions. In particular, t-statistics are asymptotically standard normal, and Wald statistics have asymptotic chi-square distributions.

To emphasize, Theorem 5 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 mis-specification. The result fails if conventional covariance matrix estimators are used, as they are in general inconsistent for the correct variance.

7 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 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

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

If the observations are clustered, then the only motification 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 with a non-clustered weight matrix. The latter takes the form

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

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 $\widehat{e}_i = y_i - x_i'\widehat{\theta}$. Set $\widehat{W} = \overline{W}_n(\widehat{\theta})$.

The asymptotic covariance matrix estimate under the iid case takes the form

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

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

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

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

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

8 Simulation

In this section, we provide simulation results of the iterated GMM for linear and nonlinear models with i.i.d. observations. We report the ratio of mean standard errors relative to the standard deviation of $\widehat{\theta}$ using our new robust standard error ($s_r(\widehat{\theta})/s.d.$) and using the conventional standard error ($s(\widehat{\theta})/s.d.$), and the empirical size of the t test at 5% significance level based on the robust t

statistic (Size(t_r)) and the conventional t statistic (Size(t)). We also report the median number of iterations required to obtain GMM convergence. The number of observations is set $n = 100$, 500, and 2500. The number of Monte Carlo repetitions are $r = 5000$.

The linear model is specified as

$$\begin{aligned} y_i &= x_i\theta_0 + (z_{1i} - z_{2i})\alpha_0 + e_i, \\ x_i &= z_{1i} + z_{2i} + u_i, \\ \begin{pmatrix} z_{1i} \\ z_{2i} \end{pmatrix} &\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right), \\ \begin{pmatrix} e_i \\ u_i \end{pmatrix} &\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & .5\sqrt{8} \\ .5\sqrt{8} & 8 \end{bmatrix}\right). \end{aligned}$$

The variance of u_i is set so that the theoretical first stage R^2 equals 0.2. α_0 measures the degree of misspecification because a nonzero α_0 implies that the instruments are invalid. Using z_{1i} and z_{2i} as instruments, θ_0 is estimated by iterated GMM and 2SLS. This model is designed so that the pseudo-true values of iterated GMM and 2SLS equal the true value. Evaluated at θ_0, the moment condition is

$$E\left(\begin{pmatrix} z_{1i} \\ z_{2i} \end{pmatrix}(y_i - x_i\theta)\right) = \begin{pmatrix} \alpha_0 \\ -\alpha_0 \end{pmatrix} = \mu_0.$$

We set $\theta_0 = 1$. The degree of misspecification is set to vary from 0 to 1.

Our nonlinear model is a modification of the linear model obtained by replacing θ_0 with $\exp(\theta_0)$. All other specifications are identical. The pseudo-true value of iterated GMM is again equal to the true value.

The results for the linear model estimated by iterated GMM are provided in Table 1, those for the same model estimated by 2SLS presented in Table 2, and those for the nonlinear model are in Table3. The results in the three tables are quite similar, with the main difference being that iterated GMM has greater distortions for the smallest sample size ($n = 100$).

The second and third columns report the ratio of mean standard errors relative to the standard deviation of $\widehat{\theta}$ using our new robust standard error and using the conventional standard error, respectively. The results show that the conventional standard errors are highly biased, and this bias does not decrease with the sample size. The bias is sufficiently severe that the mean standard error can be less than one-half the true standard deviation. In contrast, the results show that our new robust standard errors have quite modest bias. For most cases the mean standard deviations are close to the true standard deviations. The largest deviation is only 10%.

The fourth and fifth columns report the size of the new robust 5% t-test and the conventional 5% t-test, respectively. The conventional t-test can be quite highly biased, with rejection rates exceeding 40% in some cases, and the bias does not diminish with sample size. In contrast the robust t-statistics have considerably mildly size distortions. The iterated GMM estimator has modest distortions for the small sample size (rejection rates as high as 14% for $n = 100$) but these

n	α_0	$s_r(\hat{\theta})/s.d$	$s(\hat{\theta})/s.d$	Size(t_r)	Size(t)	Med. Iter
100	0	0.9718	0.9389	0.0654	0.0672	3
	0.1	0.9884	0.9025	0.0622	0.0762	3
	0.2	0.9674	0.7793	0.0640	0.1064	5
	0.3	0.9981	0.6951	0.0746	0.1460	6
	0.4	0.9681	0.6017	0.0954	0.2074	7
	0.5	0.9585	0.5421	0.1020	0.2664	8
	0.6	0.9430	0.5088	0.1080	0.2966	9
	0.7	0.9273	0.4651	0.1290	0.3454	10
	0.8	0.9360	0.4533	0.1264	0.3652	10
	0.9	0.9064	0.4315	0.1414	0.3916	11
	1	0.9210	0.4273	0.1382	0.4098	11
500	0	1.0083	1.0021	0.0492	0.0508	2
	0.1	0.9966	0.9365	0.0502	0.0644	3
	0.2	1.0064	0.8289	0.0470	0.0994	5
	0.3	0.9950	0.7005	0.0552	0.1634	6
	0.4	0.9984	0.6109	0.0616	0.2314	8
	0.5	0.9938	0.5450	0.0604	0.2800	9
	0.6	0.9726	0.4844	0.0712	0.3428	10
	0.7	0.9761	0.4501	0.0808	0.3640	11
	0.8	1.0090	0.4371	0.0768	0.3874	12
	0.9	0.9776	0.4104	0.0860	0.4206	13
	1	0.9845	0.3972	0.0836	0.4448	14
2500	0	0.9819	0.9808	0.0570	0.0572	2
	0.1	1.0204	0.9637	0.0458	0.0578	3
	0.2	1.0011	0.8263	0.0508	0.1092	4
	0.3	0.9985	0.7050	0.0532	0.1666	6
	0.4	0.9898	0.6042	0.0526	0.2460	7
	0.5	0.9925	0.5368	0.0520	0.2878	8
	0.6	1.0162	0.4980	0.0524	0.3226	10
	0.7	0.9825	0.4451	0.0602	0.3792	11
	0.8	1.0020	0.4262	0.0540	0.4118	12
	0.9	1.0016	0.4032	0.0558	0.4398	13
	1	1.0128	0.3928	0.0588	0.4268	14

Table 1: Monte Carlo Results for Linear Model - Iterated GMM

n	α_0	$s_r(\hat{\theta})/s.d$	$s(\hat{\theta})/s.d$	Size(t_r)	Size(t)
100	0	0.9846	0.9623	0.0612	0.0636
	0.1	1.0018	0.9422	0.0580	0.0688
	0.2	0.9571	0.8256	0.0530	0.0856
	0.3	1.0016	0.7724	0.0506	0.1078
	0.4	0.9802	0.6854	0.0540	0.1468
	0.5	0.9659	0.6258	0.0618	0.1826
	0.6	0.9713	0.5987	0.0556	0.2092
	0.7	0.9604	0.5545	0.0570	0.2308
	0.8	0.9687	0.5391	0.0608	0.2516
	0.9	0.9557	0.5164	0.0584	0.2738
	1	0.9778	0.5140	0.0516	0.2798
500	0	1.0081	1.0040	0.0480	0.0482
	0.1	1.0004	0.9595	0.0498	0.0610
	0.2	1.0109	0.8848	0.0448	0.0798
	0.3	0.9915	0.7783	0.0554	0.1280
	0.4	1.0016	0.7124	0.0496	0.1542
	0.5	1.0101	0.6601	0.0454	0.1892
	0.6	0.9808	0.5986	0.0510	0.2340
	0.7	0.9900	0.5739	0.0530	0.2508
	0.8	0.9897	0.5513	0.0522	0.2694
	0.9	0.9872	0.5313	0.0562	0.2918
	1	0.9974	0.5210	0.0524	0.3060
2500	0	0.9828	0.9820	0.0566	0.0566
	0.1	1.0174	0.9789	0.0474	0.0552
	0.2	1.0076	0.8849	0.0480	0.0824
	0.3	0.9997	0.7878	0.0522	0.1230
	0.4	1.0011	0.7133	0.0490	0.1614
	0.5	0.9905	0.6481	0.0552	0.1978
	0.6	1.0248	0.6271	0.0452	0.2122
	0.7	0.9979	0.5786	0.0526	0.2640
	0.8	0.9882	0.5493	0.0522	0.2772
	0.9	0.9979	0.5362	0.0532	0.2938
	1	1.0165	0.5315	0.0484	0.2900

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

n	α_0	$s_r(\hat{\theta})/s.d$	$s(\hat{\theta})/s.d$	Size(t_r)	Size(t)	Med. Iter
100	0	0.9654	0.9299	0.0540	0.0560	3
	0.1	0.9221	0.8444	0.0710	0.0830	3
	0.2	0.9856	0.7909	0.0780	0.1070	4
	0.3	0.9973	0.7017	0.0760	0.1490	5
	0.4	1.0020	0.6147	0.0810	0.1850	7
	0.5	0.9739	0.5548	0.1040	0.2290	7
	0.6	0.9034	0.4785	0.1200	0.2990	8
	0.7	0.9616	0.4877	0.1160	0.3260	9
	0.8	0.9179	0.4539	0.1350	0.3380	10
	0.9	0.8745	0.4212	0.1480	0.3830	10
	1	0.8751	0.4360	0.1410	0.3750	10
500	0	1.0056	0.9992	0.0490	0.0520	2
	0.1	0.9871	0.9260	0.0510	0.0680	3
	0.2	1.0047	0.8243	0.0540	0.1130	4
	0.3	0.9984	0.7048	0.0500	0.1690	6
	0.4	0.9848	0.6051	0.0630	0.2340	7
	0.5	0.9794	0.5318	0.0780	0.2910	8
	0.6	0.9819	0.4944	0.0640	0.3380	9
	0.7	0.9958	0.4658	0.0730	0.3610	10
	0.8	0.9559	0.4163	0.0930	0.4310	11
	0.9	0.9785	0.4158	0.0800	0.4320	11.5
	1	0.9816	0.3906	0.1010	0.4450	12
2500	0	1.0051	1.0038	0.0490	0.0500	2
	0.1	0.9907	0.9356	0.0550	0.0670	3
	0.2	0.9882	0.8138	0.0550	0.1130	4
	0.3	0.9975	0.7043	0.0550	0.1640	5
	0.4	0.9338	0.5699	0.0680	0.2720	6
	0.5	1.0681	0.5759	0.0420	0.2610	8
	0.6	1.0008	0.4880	0.0600	0.3370	9
	0.7	0.9942	0.4512	0.0550	0.3830	10
	0.8	0.9862	0.4224	0.0590	0.4150	11
	0.9	0.9929	0.4010	0.0510	0.4360	11
	1	1.0296	0.3992	0.0450	0.4670	12

Table 3: Monte Carlo Results for Nonlinear Model - Iterated GMM

are eliminated in large samples (rejection rates no larger than 6% for $n = 2500$). The results show clearly that the conventional t-statistic cannot control size, while the new robust t-statistics have excellent size control, at least in large samples. Table 2 shows that for the 2SLS estimator the new robust t-statistics have excellent size control for all sample sizes (rejection rates no larger than 6%), indicating that the small sample distortions of Table 1 are probably due to standard challenges in weight matrix estimation, and not specific to our estimator.

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 our theorem of Section 3 which shows that a sufficient condition for GMM iteration to be a contraction is mild mis-specification. As mis-specification increases the contraction property weakens and thus iterative convergence slows.

It is also quite interesting to point out the behavior of the statistics when there is no mis-specification ($\alpha_0 = 0$). 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 not estimating unnecessary term. However, in all three tables the robust standard error has less bias than the conventional standard error, and the robust t-statistic has slightly less size distortion than the conventional t-statistic. This finding points out that there is no apparent cost of using our new robust standard errors, even in the context of no mis-specification.

9 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 on democracy.

The model estimated by AJRY is

$$d_{it} = \alpha d_{i,t-1} + \gamma y_{i,t-1} + \mu_t + \delta_i + u_{it}, \tag{29}$$

where d_{it} is a measure of democracy, $y_{i,t-1}$ is the lagged value of log income per capita, μ_t's are a set of time effects, and δ_i's are a set of country fixed effects. 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. By taking time difference, we can eliminate the country fixed effects:

$$\Delta d_{it} = \alpha \Delta d_{i,t-1} + \gamma \Delta y_{i,t-1} + \Delta \mu_t + \Delta u_{it}. \tag{30}$$

The model is estimated by the difference GMM of Arellano and Bond (1991). The dataset is an unbalanced panel. For each country i, the length of available time series is T_i. Due to differencing, only $T_i - 2$ time series observations are used for estimation. Thus, the number of countries N corresponds to the number of clusters G, and the time series observations used for estimation $T_i - 2$ corresponds to the number of individuals in the cluster n_g.

In the tables below, results in columns I and III are based on the one-step estimator and results in columns II and IV are based on the iterated estimator. For each point estimate, we report the robust standard errors ("robust") and the conventional standard errors ("conventional") in parentheses. We also report the number of instruments used, the number of total observations $\sum_{i=1}^{N}(T_i - 2)$, and the number of countries N.

Table 4 presents the replication result of Table 2 of AJRY. The main finding of AJRY is presented in columns I and III. The results based on the iterated GMM are presented in columns II and IV. The findings are interesting. The iterated GMM point estimates of the effect of income on democracy are smaller (in absolute value). The robust standard error of the iterated GMM is almost half of that of the conventional one-step GMM for five-year data. Overall, the iterated GMM estimation results provide strong evidence of no causal effect of income on democracy.

AJRY Table 2	Column (4) five-year data		Column (8) ten-year data	
	One-step I	Iterated II	One-step III	Iterated IV
Democracy_{t-1}	0.489	0.744	0.227	0.288
robust	(0.095)	(0.128)	(0.125)	(0.146)
conventional	(0.085)	(0.043)	(0.123)	(0.111)
Income_{t-1}	-0.129	**-0.009**	-0.318	**-0.280**
robust	(0.088)	**(0.039)**	(0.183)	**(0.202)**
conventional	(0.076)	(0.040)	(0.180)	(0.170)
Hansen J Test	[0.01]	[0.04]	[0.02]	[0.09]
# of Instruments	55		15	
Observations	838		338	
Countries	127		118	

Table 4: Income and Democracy

The use of the iterated GMM and the robust standard error are strongly recommended because the underlying moment condition is potentially misspecified. First of all, the instruments may violate the exclusion restrictions. AJRY show that the likely sources of correlation between the instruments and the error term are not present, so it is unlikely that the instruments are invalid at large. However, unless one is fully confident in the validity of instruments, it is reasonable to take into account for a small/moderate violation of the assumption on estimation and inference. Second, the effect of income on democracy may be heterogeneous across countries. Cervellati, Jung, Sunde, and Vischer (2014, CJSV hereinafter) argue that the effect is heterogeneous across former colonies and non-colonies, and even within former colonies. Bonhomme and Manresa (2015) find evidence of grouped patterns of unobserved heterogeneity in the same dataset.

One implication of parameter heterogeneity is that the underlying moment condition may be

misspecified. To fix ideas, consider a simple regression model with endogeneity

$$y_i = x_i\beta + \varepsilon_i, \tag{31}$$

where y_i and x_i are scalars and $Ex_i\varepsilon_i \neq 0$. With valid instruments z_{1i} and z_{2i} that satisfy $Ez_{1i}\varepsilon_i = Ez_{2i}\varepsilon_i = 0$, β can be consistently estimated by using the over-identified moment conditions

$$0 = E[z_{1i}(y_i - x_i\beta)] = E[z_{2i}(y_i - x_i\beta)]. \tag{32}$$

Now suppose heterogeneity in the regression equation

$$y_i = x_i\beta + x_ic_i\gamma + u_i, \tag{33}$$

where c_i is a dummy variable and $Ez_{1i}u_i = Ez_{2i}u_i = 0$. If the true DGP is (33), but we use the moment conditions (32) then it imposes

$$\frac{Ez_{1i}y_i}{Ez_{1i}x_i} = \frac{Ez_{2i}y_i}{Ez_{2i}x_i}. \tag{34}$$

By plugging in (33), we have

$$\beta + \gamma\frac{Ez_{1i}x_ic_i}{Ez_{1i}x_i} = \beta + \gamma\frac{Ez_{2i}x_ic_i}{Ez_{2i}x_i}. \tag{35}$$

This equation holds if $\gamma = 0$ or $Ez_{ji}x_ic_i = Ez_{ji}x_iEc_i$ for $j = 1, 2$, but does not hold in general. A GMM estimator based on (32) will converge in probability to a linear combination of the both sides of (35), which is a value that depends on the weight matrix. Evaluated at the probability limit, the moment condition (32) does not hold so the model is misspecified.

CJSV argue that although AJRY found no evidence of income on democracy, if the countries are divided into two groups, former-colonies and non-colonies, there is strong evidence that income has a negative effect on democracy for former-colonies and a positive effect for non-colonies. Their model is

$$d_{it} = \alpha d_{i,t-1} + \gamma y_{i,t-1} + \phi y_{i,t-1}c_i + \mu_t + \delta_i + u_{it}, \tag{36}$$

where c_i is a time invariant country specific feature. By time differencing, we have

$$\Delta d_{it} = \alpha\Delta d_{i,t-1} + \gamma\Delta y_{i,t-1} + \phi\Delta y_{i,t-1}c_i + \Delta\mu_t + \Delta u_{it}. \tag{37}$$

This equation is used to form the moment condition. If we estimate the original model (30) while (37) is true, then the corresponding moment condition is misspecified. This justifies the use of the iterated GMM and the robust standard errors in Table 4.

We now investigate whether parameter heterogeneity is supported by data. Following CJSV, let c_i be an indicator for former colonies. Table 5 shows the estimation result of Equation (37) based on the full sample. The null hypothesis is $H_0 : \phi = 0$ that there is no heterogeneity in the effect of

	five-year data		ten-year data	
	One-step	Iterated	One-step	Iterated
	I	II	III	IV
Democracy_{t-1}	0.451	0.714	0.136	0.188
robust	(0.104)	(0.139)	(0.126)	(0.161)
conventional	(0.091)	(0.045)	(0.119)	(0.114)
Income_{t-1}	-0.076	0.005	-0.142	-0.122
robust	(0.075)	(0.028)	(0.142)	(0.170)
conventional	(0.065)	(0.032)	(0.139)	(0.128)
$\text{Income}_{t-1} \times c_i$	-0.141	**-0.039**	-0.368	**-0.317**
robust	(0.083)	**(0.049)**	(0.166)	**(0.169)**
conventional	(0.072)	(0.038)	(0.158)	(0.152)
Hansen J Test	[0.01]	[0.05]	[0.02]	[0.06]
# of Instruments	56		16	
Observations	838		338	
Countries	127		118	

Table 5: Income and Democracy with Colony Interaction Effect

income on democracy between former colonies and non-colonies. CJSV find that the estimates of ϕ are statistically significant using the one-step GMM and the conventional standard error (columns I and III), but they become insignificant using the iterated GMM and the robust standard error (columns II and IV). This is against the finding of CJSV that former colonies and non-colonies have heterogeneous effects of income on democracy.

Although estimating (37) using the full sample yields more precise estimates, it also imposes an additional restrictions that all other coefficients are the same across colonies and non-colonies. To allow for heterogeneous convergence processes and development dynamics, we also estimate (30) for the subsample of colonies and non-colonies and present the results[1] in Table 6. The finding is striking. CJSV argue that non-colonies have a positive effect of income on democracy while former colonies have a negative effect based on significant point estimates in their Table 3 columns (1) and (2). However, with the iterated GMM, the point estimates change towards zero and they are not significantly different from zero using the robust standard error. Thus, both for non-colonies and former colonies, there is no evidence of causal effect of income on democracy. This further supports the original finding of AJRY.

To further investigate heterogeneity within colonies, we estimate (37) with $c_i = 1$ if the country had historically strong institutions, and $c_i = 0$ otherwise. As a proxy for institutional quality,

[1] In Table 6 Panel A Columns I and II, the instruments are based on the double lags (Holtz-Eakin, Newey, and Rosen, 1988) rather than all the lags greater than or equal to 2 (Arellano and Bond, 1991) due to the small number of cross-section observations.

CJSV use (i) the level of constraints on the executive in 1900; (ii) the information whether the country became independent before 1900; and (iii) the information whether the colony was subject to the rule of the late colonial powers. CJSV found significant heterogeneity in the effect of income on democracy in colonies across all the three specifications.

The findings are again in sharp contrast with those of CJSV. Across all specifications, the coefficient estimate of ϕ are not significantly different from zero, which implies that the null hypothesis of no heterogeneity with respect to institutional quality is not rejected. The only exception is Table 9 Column II, where the coefficient estimate is marginally significant at 5% level. But the magnitude is still relatively small and the effect of income on democracy is not significantly different from zero. Therefore there is no evidence of heterogeneity with respect to institutional quality within colonies.

Does this imply the effect of income on democracy homogeneous across countries and there is no need to worry about model misspecification? Although insignificant due to large standard errors, the point estimates of the effect of income on democracy (γ, not ϕ) are negative and large (in absolute value) in Table 7 columns II and IV and Table 9 column IV. This suggests the presence of potential heterogeneity within former colonies with relatively weak (or "extractive", according to Acemoglu, Johnson, Robinson, and Yared, 2009) institutions. In addition, Bonhomme and Manresa (2015) divide the countries into four groups based on unobserved heterogeneity where those four groups show distinctive paths of democracy over time. Overall there is some evidence of parameter heterogeneity across countries. The iterated GMM and the robust standard error can provide reliable results.

10 Appendix

Proof of Theorem 1: To show that the map $g_n(\phi)$ is a contraction, we show that for $c = 2kC^5\delta < 1$, $\|g(\phi_1) - g(\phi_2)\| \leq c\|\phi_1 - \phi_2\|$ 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

$$0 = \frac{\partial}{\partial \theta} J_n(\theta, \phi) = 2Q_n(\theta)' W_n(\phi)^{-1} m_n(\theta) \tag{38}$$

where $Q_n(\theta) = \frac{\partial}{\partial \theta'} m_n(\theta)$. Since (38) 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

$$V_n(\phi) = \frac{\partial}{\partial \phi'} g_n(\phi) = -D_n(g_n(\phi), \phi)^{-1} B_n(g_n(\phi), \phi) \tag{39}$$

where

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

We calculate that

$$\begin{aligned} B_n(\theta, \phi) &= 2\left[m_n(\theta)' \otimes Q_n(\theta)'\right] \frac{\partial}{\partial \phi'} \text{vec}\left(W_n(\phi)^{-1}\right) \\ &= -2\left[m_n(\theta)' \otimes Q_n(\theta)'\right] [W_n(\phi)^{-1} \otimes W_n(\phi)^{-1}] S_n(\phi). \end{aligned} \tag{40}$$

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

$$\|B_n(g_n(\phi), \phi)\| \leq 2\|Q_n(g_n(\phi))\| \|S_n(\phi)\| \left\|W_n(\phi)^{-1}\right\|^2 \|m_n(g_n(\phi))\| \leq 2C^4\delta. \tag{41}$$

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,

$$\left\|[D_n(g_n(\phi), \phi)^{-1}]_j\right\| \leq \|D_n(g_n(\phi), \phi)^{-1}\|_F \leq \sqrt{k}\lambda_{\max}(D_n(g_n(\phi), \phi)^{-1}) \leq \sqrt{k}C. \tag{42}$$

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 ϕ_1 and ϕ_2, the Schwarz inequality,

(39), the Schwarz matrix inequality, (41), (42), and $c = 2kC^5\delta$,

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

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

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

the final convergence by Assumption 2 (10) and (13) plus Assumption 1.2. This implies

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

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

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

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

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

as required. ■

Proof of Theorem 2.2: The fixed point $\widehat{\theta}$ exists and is unique if $\widehat{g}(\phi)$ is a contraction mapping, in the sense that there is a $0 \leq c < 1$ such that

$$\|\widehat{g}(\phi_1) - \widehat{g}(\phi_2)\| \leq c \|\phi_1 - \phi_2\| \tag{43}$$

for all $\phi_1, \phi_2 \in \Theta$. Dominitz and Sherman (2005) Lemma 3 show that sufficient conditions for (43) 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) $\sup_\phi \|\widehat{g}(\phi) - g_n(\phi)\| \to_p 0$ (established in part 1); and (iii) $\sup_\phi \left\|\widehat{V}(\phi) - V_n(\phi)\right\| \to_p 0$ where $V_n(\phi) = \frac{\partial}{\partial \phi'} g_n(\phi)$ and $\widehat{V}(\phi) = \frac{\partial}{\partial \phi'} \widehat{g}(\phi)$. Hence it is sufficient to verify this final condition.

Recall that $V_n(\phi)$ can be expressed as (39) where $B_n(\theta, \phi)$ equals (40). 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,

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

where

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

and

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

Assumption 2 and 1.2 imply that $\widehat{B}(\theta, \phi) - B_n(\theta, \phi)$ and $\widehat{D}(\theta, \phi) - D_n(\theta, \phi)$ converge uniformly to 0. Part 1 shows that $\widehat{g}(\phi) - g_n(\phi)$ converges uniformly to 0. Together, this implies that $\widehat{V}(\phi) - V_n(\phi)$ 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 $\sup_\phi \|\widehat{g}(\phi) - g_n(\phi)\| \to_p 0$ (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 establish a slightly more general result. For any population weight matrix $W_n(\theta)$ set

$$W_n^*(\theta) = W_n(\theta) - m(\theta) m(\theta)'. \tag{44}$$

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 $\theta_n = \theta_n^*$.

By the Woodbury matrix identity,

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

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

$$\begin{aligned} m(\theta)'W_n(\theta_n^*)^{-1}m(\theta) &= m(\theta)'W_n^*(\theta_n^*)^{-1}m(\theta) - \frac{m(\theta)'W_n^*(\theta_n^*)^{-1}m(\theta_n^*)m(\theta_n^*)'W_n^*(\theta_n^*)^{-1}m(\theta)}{1 + m(\theta_n^*)'W_n^*(\theta_n^*)^{-1}m(\theta_n^*)} \\ &= (m(\theta)'W_n^*(\theta_n^*)^{-1}m(\theta))\left(1 - \rho_n(\theta,\theta_n^*)\frac{J_n^*}{1+J_n^*}\right) \end{aligned} \tag{45}$$

where

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

and $J_n^* = m(\theta_n^*)'W_n^*(\theta_n^*)^{-1}m(\theta_n^*)$. Now consider minimization of (45) over θ given fixed θ_n^*. The first term on the right-hand-side of (45) 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 (45) is minimized by maximizing $\rho_n(\theta,\theta_n^*)$ which is achieved at $\theta = \theta_n^*$ because $\rho(\theta,\theta_n^*)$ is a squared correlation. Since both terms are minimized at $\theta = \theta_n^*$ it follows that (45) is minimized at $\theta = \theta_n^*$. But the left-hand-side of (45) 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.

The same argument applies to the sample versions of weight matrices and estimators. ■

Proof of Theorem 4: We first claim that Assumption 3 implies the convergence results of Assumption 2. This follows by a 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 $\max_g n_g/n \to 0$. The uniform integrability holds by Assumption 3.3, the Lipschitz condition by Assumption 3.4 and the cluster size condition is implied by Assumption 3.7. Thus 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 (15) in the text. It is convenient to note that we can write $F = Q'W^{-1}m$ using the alternative representations

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

and recall the identity

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

The chain rule then yields (15). 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 (16) 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) imply that

$$\sup_{\theta\in\Theta}\left\|\overline{H}_n(\theta) - H_n(\theta)\right\| \to_p 0 \tag{46}$$

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

$$\|H_n^* - H_n\| \to_p 0. \tag{47}$$

We next justify equation (18) from the text. First, it follows from Assumption 3.3 that for each $\theta \in \Theta$, $2 \le r < \infty$, and $f(x) = m(x,\theta)$, $Q(x,\theta)$, and $W(x,\theta)$,

$$\sup_i E\|f(X_i)\| \le \sup_i E\|f(X_i)\|^r \le B + 1 < \infty \tag{48}$$

for a large enough B. This implies that $m_n = O(1)$, $Q_n = O(1)$, and $W_n = O(1)$. By the convergence results in Assumption 2, the left-hand side of (18) can be written as

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

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

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

the final equality since $Q_n' W_n^{-1} \mu_n = 0$. Together, these expansions lead to (18). 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 (18).

Equation (19) is an algebraic equivalence. We have established that

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

The cluster sums $\widetilde{\psi}_g$ are independent across g. Assumptions 3.2, 3.3, 3.5, 3.6 are equivalent to the assumptions for the CLT of Theorem 2 of Hansen and Lee (2017). This implies that (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 (47) imply that

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

Applying the continuous mapping theorem we find

$$\left\| H_n^{1/2} H_n^{*-1} H_n^{1/2} - I \right\| \to_p 0. \qquad (51)$$

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

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

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

Proof of Theorem 5: It is sufficient to show that

$$\left\|H_n^{1/2}\widehat{H}^{-1}H_n^{1/2}-I_k\right\| \to_p 0 \tag{52}$$

and

$$\left\|\Omega_n^{-1/2}\widehat{\Omega}\Omega_n^{-1/2}-I_k\right\| \to_p 0. \tag{53}$$

The convergence (55) is established similarly to (51) from (46), Theorem 2, and Assumption 3.1.

Now define

$$\widehat{\psi}_g(\theta)=\sum_{j=1}^{n_g}\left(\widehat{Q}'\widehat{W}^{-1}m(X_{gj},\theta)+Q(X_{gj},\theta)'\widehat{W}^{-1}\widehat{\mu}-\widehat{Q}'\widehat{W}^{-1}W(X_{gj},\theta)\widehat{W}^{-1}\widehat{\mu}\right) \tag{54}$$

and

$$\widehat{\Omega}(\theta)=\frac{1}{n}\sum_{g=1}^{G}\widehat{\psi}_g(\theta)\widehat{\psi}_g(\theta)'$$

so that

$$\widehat{\Omega}=\widehat{\Omega}(\widehat{\theta}).$$

also define

$$\widetilde{\psi}_g(\theta)=\sum_{j=1}^{n_g}\left(Q_n'W_n^{-1}m(X_{gj},\theta)+Q(X_{gj},\theta)'W_n^{-1}\mu_n-Q_n'W_n^{-1}W(X_{gj},\theta)W_n^{-1}\mu_n\right)$$

and

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

Equation (53) then follows from $\left\|\theta_n-\widehat{\theta}\right\| \to_p 0$ and

$$\sup_{\theta\in\Theta}\left\|\Omega_n(\theta)^{-1/2}\widehat{\Omega}(\theta)\Omega_n(\theta)^{-1/2}-I_k\right\| \to_p 0. \tag{55}$$

Equation (55) follows from a ULLN for clustered variance estimators established in Theorem 6 of Hansen and Lee (2017) (which uses Assumption 4), if (54) is defined with the population moments Q_n, W_n and μ_n rather than $\widehat{Q}$, $\widehat{W}$, and $\widehat{\mu}$. Since these are all consistent (the convergence results

listed in Assumption 2 which are implied by Assumption 3 as discussed at the beginning of the proof of Theorem 4), this replacement does not affect the result in (55). ■

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. 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.

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

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

6. 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.

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

8. 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.

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

10. Hansen, Christian B. (2007). Asymptotic properties of a robust variance matrix estimator for panel data when T is large. *Journal of Econometrics*, 141, 597-620.

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

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

13. Holtz-Eakin, Douglas, Whitney K. Newey, and Harvey S. Rosen (1988). Estimating vector autoregressions with panel data. *Econometrica*, 56(6), 1371-1395.

14. Hwang, Jungbin (2016). Simple and trustworthy cluster-robust GMM inference. Working paper, University of Connecticut.

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

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

	five-year data		ten-year data	
	One-step	Iterated	One-step	Iterated
	I	II	III	IV
Panel A: Non-Colonies				
Democracy_{t-1}	0.638	0.511	0.369	0.297
robust	(0.186)	(0.334)	(0.231)	(0.170)
conventional	(0.181)	(0.146)	(0.212)	(0.121)
Income_{t-1}	0.274	**0.099**	0.443	**0.090**
robust	(0.377)	**(0.146)**	(0.316)	**(0.045)**
conventional	(0.194)	(0.079)	(0.217)	(0.034)
Hansen J Test	[0.01]	[0.12]	[0.04]	[0.25]
# of Instruments	19		15	
Observations	207		86	
Countries	27		26	
Panel B: Colonies				
Democracy_{t-1}	0.401	0.699	0.118	0.063
robust	(0.126)	(0.146)	(0.155)	(0.254)
conventional	(0.105)	(0.041)	(0.127)	(0.119)
Income_{t-1}	-0.231	**-0.033**	-0.294	**-0.352**
robust	(0.144)	**(0.054)**	(0.326)	**(0.234)**
conventional	(0.112)	(0.052)	(0.229)	(0.217)
Hansen J Test	[0.03]	[0.03]	[0.01]	[0.04]
# of Instruments	55		15	
Observations	631		252	
Countries	100		92	

Table 6: Income and Democracy in Colonies and Non-Colonies

CJSV Table 4, Panel B, Columns (1) and (2)
c_i: Constraints on the executive in 1900

	five-year data		ten-year data	
	One-step	Iterated	One-step	Iterated
	I	II	III	IV
Democracy_{t-1}	0.289	-0.423	-0.058	-0.030
robust	(0.142)	(0.380)	(0.161)	(0.159)
conventional	(0.123)	(0.039)	(0.132)	(0.111)
Income_{t-1}	-0.417	-0.337	-0.716	-0.579
robust	(0.221)	(0.289)	(0.468)	(0.394)
conventional	(0.194)	(0.116)	(0.364)	(0.299)
$\text{Income}_{t-1} \times c_i$	0.345	**0.296**	0.387	**0.081**
robust	(0.169)	**(0.309)**	(0.215)	**(0.124)**
conventional	(0.162)	(0.073)	(0.213)	(0.157)
Hansen J Test	[0.03]	[0.03]	[0.01]	[0.04]
# of Instruments	56		16	
Observations	531		216	
Countries	79		75	

Table 7: Income and Democracy within Colonies with Institutional Quality Interaction Effect using Constraints on the Executive in 1900 as a Proxy

CJSV Table 4, Panel B, Columns (3) and (4)
c_i: Independence before 1900

	five-year data		ten-year data	
	One-step	Iterated	One-step	Iterated
	I	II	III	IV
Democracy_{t-1}	0.343	0.724	0.095	-0.001
robust	(0.127)	(0.152)	(0.154)	(0.224)
conventional	(0.110)	(0.044)	(0.139)	(0.128)
Income_{t-1}	-0.270	-0.011	-0.185	-0.083
robust	(0.134)	(0.047)	(0.257)	(0.411)
conventional	(0.113)	(0.050)	(0.175)	(0.163)
$\text{Income}_{t-1} \times c_i$	0.224	**0.020**	0.157	**0.034**
robust	(0.125)	**(0.077)**	(0.215)	**(0.163)**
conventional	(0.121)	(0.037)	(0.186)	(0.148)
Hansen J Test	[0.03]	[0.04]	[0.002]	[0.005]
# of Instruments	56		16	
Observations	628		251	
Countries	99		91	

Table 8: Income and Democracy within Colonies with Institutional Quality Interaction Effect using Independence before 1900 as a Proxy

CJSV Table 4, Panel B, Columns (5) and (6)
c_i: No late colonial power

	five-year data		ten-year data	
	One-step	Iterated	One-step	Iterated
	I	II	III	IV
Democracy_{t-1}	0.355	0.666	0.085	0.059
robust	(0.115)	(0.125)	(0.134)	(0.178)
conventional	(0.101)	(0.040)	(0.125)	(0.119)
Income_{t-1}	-0.303	-0.052	-0.332	-0.262
robust	(0.122)	(0.041)	(0.242)	(0.208)
conventional	(0.110)	(0.047)	(0.203)	(0.181)
$\text{Income}_{t-1} \times c_i$	0.318	**0.111**	0.387	**0.372**
robust	(0.130)	**(0.053)**	(0.181)	**(0.314)**
conventional	(0.122)	(0.039)	(0.170)	(0.156)
Hansen J Test	[0.09]	[0.14]	[0.04]	[0.06]
# of Instruments	56		16	
Observations	631		252	
Countries	100		92	

Table 9: Income and Democracy within Colonies with Institutional Quality Interaction Effect using No Late Colonial Power as a Proxy