Robust IV Inference with Clustering Dependence*

Jianfei Cao

The University of Chicago
Booth School of Business
5807 S. Woodlawn, Chicago, IL 60637, USA
e-mail: jcao0@chicagobooth.edu

November 6, 2020

Abstract: Linear IV models with clustering dependence are widely used in empirical studies, although the common solution, the *cluster covariance estimator*, often produces undesirable inferential results, especially with weak instruments. In this paper, I propose a method that is robust to both weak IV and (potentially heterogeneous) clustering dependence. The proposed method is based on the idea of Fama-MacBeth estimation, with group-level estimators being a truncated version of the unbiased IV estimator. Asymptotic validity is shown under both strong and weak IV sequences, as well as under general requirements. Simulation results indicate the method has good finite-sample performance in both size and power. The proposed method is applied to study the effect of city compactness on population density.

Key Words: Weak dependence; Weak instruments; Fama-MacBeth method; t-test

1. Introduction

In linear IV models, accounting for clustering dependence has been a standard procedure in conducting statistical inference in empirical research. A common solution is to use the cluster covariance estimator (CCE), which is often referred to as the "clustered standard error" method. In linear models CCE methods are shown to deliver valid inference under either strong homogeneity across groups (the *large-homogeneous-group* approach, e.g., Bester

*I am grateful to Christian Hansen, Max Farrell, Tetsuya Kaji, Panos Toulis, Azeem Shaikh, Stéphane Bonhomme, and Alexander Torgovitsky for helpful comments and suggestions.

et al., 2011) or lots of small groups (the *many-small-group* approach, e.g., Hansen and Lee, 2019). Those results provide theoretical justification for the usage of CCE methods in the linear IV model under strong IV.[1] This paper concerns statistically inference in the linear IV model with clustering dependence.

However, for many common settings, whether either the large-homogeneous-group or many-small-group approach can justify the usage of standard clustering methods is not clear. Specifically, Bester et al. (2011) assume all groups are similar in size and the same design-matrix limit, which does not hold in many settings. Hansen and Lee (2019) require that $\max_g n_g^2/n \to 0$, where n_g is the number of observations in group g and n is the sample size. In the case of equal-sized groups, this requirement implies $n/G^2 \to 0$, where G is the number of groups. MacKinnon and Webb (2017) conduct simulation studies and show the empirical rejection can be as high as 0.1073 at a level-0.05 test, when $(n, G) = (2000, 50)$, thus with $n/G^2 = 0.8$, and group sizes are proportional to population of the 50 states in the US. A non-exhaustive search in recent empirical research suggests n/G^2 is often large, for example, Coibion et al. (2017) with $n/G^2 = 0.46$ or 0.34 in Table 3, Dell (2012) with n/G^2 ranging from 1.21 to 16.65 in Table 7, and Deryugina et al. (2019) with $n/G^2 = 2.43$ in Table 2. For a discussion on the poor asymptotic approximation of inference methods based on asymptotic theory, see Ferman and Pinto (2019), Ferman (2019), MacKinnon and Webb (2017), and Young (2019).

Moreover, standard methods suffer from size distortion when weak IV is a concern. Although robust inference methods such as the Anderson-Rubin test (AR, Anderson and Rubin, 1949) work under standard assumptions described in the previous paragraph, whether those methods have good inferential properties when standard assumptions break is not well understood. In the simulation section, we show the extension of AR with the standard error calculated by CCE methods can result in size distortion under imbalanced group sizes, with sizes being as high as 0.116 at a level-0.05 test.

Alternatively, Fama-MacBeth methods (Fama and MacBeth, 1973; Ibragimov and Müller, 2010), sometimes referred to as mean group estimation (Pesaran and Smith, 1995; Pesaran et al., 1999), provide another inferential approach that exploits the clustering dependence structure. Those methods first perform group-level estimation for each group and consider a weighted average of all group-level estimators. Under a wide variety of circumstances, the resulting average has well-understood properties, and a simple procedure such as a t-test

[1] Hansen and Lee (2019) cover both OLS and IV, whereas Hansen (2007) considers only the OLS case, but the results can be extended to the strong IV model.

can be used to attain valid inference. In this paper I introduce a group-based inference method that is built on Fama-MacBeth methods, in order to simultaneously solve clustering dependence and potentially weak IV.

In this paper I study robust inferential methods to overcome the practical issues mentioned above, based on the idea of Fama-MacBeth estimation. Because the Fama-MacBeth approach calculates the group-level estimator using only the data in a certain group, a potential finite-sample problem may arise in the IV estimation. To account for that possibility, I propose a truncated version of the unbiased IV estimator introduced by Andrews and Armstrong (2017) in calculating the group-level estimator. I show this estimator is nearly unbiased, and that using it in the Fama-MacBeth approach produces valid inference. The proposed method allows for a moderate number of moderate-sized groups (e.g., 30 groups of around 30 observations as in the simulation section) and is robust to both weak IV and heterogeneous clustering dependence. Table 1 summarizes whether a certain aforementioned method is robust to a non-conventional set-up.

TABLE 1
Robustness of Inferential Methods in Linear IV Models with Clustering Dependence

	$n/G^2 \gg 0$	Heterogeneous Groups	Weak IV
CCE (large-G)	NO	YES	NO
CCE (small-G)	YES	NO	NO
AR-CCE (large-G)	NO	YES	YES
AR-CCE (small-G)	YES	NO	YES
Fama-MacBeth	YES	YES	NO
Proposed method	YES	YES	YES

Notes: This table roughly summarizes whether a candidate inferential method is robust to a certain non-conventional set-up. "YES" means it generally delivers correct size and "NO" means it does not. "Large-G" stands for the *many-small-group* approach and "small-G" stands for the *large-homogeneous-group*" one. "AR-CCE" is the natural extension of the Anderson-Rubin method to the case with clustering dependence (described in Section 4.2). "Proposed method" is the Fama-MacBeth approach with truncated unbiased estimators proposed in this paper.

Both an unbiased group-level estimator and the truncation are important in implementing the Fama-MacBeth approach in this setting. Without the former, the group-level IV estimator may lead to substantial finite-sample bias and cause size distortion under the null. The latter guarantees the group-level estimators have finite second moments such that the test has power. Simulation studies show direct usage of Andrews and Armstrong (2017) produces far less power, and the proposed method is robust to many settings and has good power properties.

Throughout, I assume one endogenous variable and focus on the case of one instrument.

Cases with multiple instruments can be dealt with using the averaging method introduced by Andrews and Armstrong (2017). Similar to Andrews and Armstrong (2017), to implement the proposed method, the sign of the first-stage parameter is assumed to be known. This assumption is often a weak one in empirical studies, because the sign of the instrument is typically embedded in the reasoning of instrument validity and comes in before the discussion of the strength of the instrument. Mills (2019) shows 82.35% of the papers published in the American Economic Review from 2014 to 2018 and with "instrument" in the abstract claim the first-stage sign is known. Additionally, Mills (2019) shows exploiting information of the first-stage sign may help improve test power. Another underlying assumption I assume throughout is the group-level normal model (see Section 3.1), for which a sufficient assumption would be weak dependence as in the large-homogeneous-group approach (Bester et al., 2011).

The paper contributes to two streams of literature. First, the proposed method fills a gap in the literature on cluster-based inferential methods. Although those methods are extensively studied under standard assumptions such as the large-homogeneous-group case and the many-small-group case (Bertrand et al., 2004; Hansen, 2007; Bester et al., 2011; Cameron and Miller, 2015; Hansen and Lee, 2019), the properties of those methods outside the standard assumptions are largely unknown. I show through simulation that existing methods can break under many circumstances. I advocate the usage of the proposed Fama-MacBeth approach with truncated unbiased IV estimation and show its validity.

Second, this paper complements the recent literature on the Fama-MacBeth approach and shows its usefulness. Fama and MacBeth (1973) introduced this approach, but it was only recently theoretically justified by Ibragimov and Müller (2010). Ibragimov and Müller (2010), Canay et al. (2017), Cao et al. (2019), and Hagemann (2019a,b) have documented the robustness and good power properties of this approach. Many of their results can be either applied or extended to the strong IV case, but extension to allowing for weak IV is non-trivial.

The remainder of the paper is organized as follows. Section 2 introduces a truncated version of the unbiased IV estimator with known first-stage sign. Section 3 proposes the inferential method that applies the truncated unbiased IV estimator. In addition, the primitive conditions for both strong and weak IV asymptotics are listed. Simulation studies are presented in section 4. In section 5, I apply the proposed method to study the effect of city compactness on population density. Section 6 concludes. Proofs are relegated to the appendix.

2. Truncated Unbiased IV with Known First-Stage Sign

We first consider a simple linear IV model. Let X, Y, and Z be $n \times 1$ data vectors for three scalar variables. The reduced-form formulation of the linear IV model is

$$\begin{cases} Y = Z\pi\beta + U, \\ X = Z\pi + V, \end{cases} \tag{2.1}$$

where π and β are both scalars. We are interested in the structural equation parameter β. Assume the sign of π is known, and, without loss of generality, let $\pi > 0$. Assume the vector of reduced-form and first-stage estimators follows

$$\widehat{\psi} = \begin{pmatrix} \widehat{\gamma} \\ \widehat{\pi} \end{pmatrix} = \begin{pmatrix} (Z'Z)^{-1}Z'Y \\ (Z'Z)^{-1}Z'X \end{pmatrix} \sim N(\mu, \Sigma), \tag{2.2}$$

where

$$\mu = \begin{pmatrix} \pi\beta \\ \pi \end{pmatrix}, \ \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}.$$

The usual IV estimator is $\widehat{\beta}_{IV} = \widehat{\gamma}/\widehat{\pi}$. We assume through-out that Σ is known and positive definite. Our analysis relies heavily on (2.2), which applies to cases where normality is a good approximation of the reduced-form and first-stage coefficients $\widehat{\psi}$.

Comment 2.1. The model (2.1) has an equivalent structural formulation. Generalization of (2.1) to models with multiple instruments and/or other control variables can be done in standard methods. For multiple instruments, we can use a weighted average of the proposed estimators of a single instrument, because a weighted average of (nearly) unbiased estimators is still (nearly) unbiased. Including other control variables can be done through projection on the null space by the Frisch-Waugh-Lovell theorem.

Comment 2.2. The assumption that π has known sign is often weak in empirical studies. According to a survey by Mills (2019) on papers published in the American Economic Review from 2014 to 2018, 14 out of 17 papers with "instrument" in the abstract claim to have known first-stage sign.

Comment 2.3. The normal model (2.2) is common in the literature on IV inference that is robust to weak instruments (see, e.g., Andrews et al., 2006; Andrews and Mikusheva, 2016; Kleibergen, 2002; Moreira, 2003; Moreira and Moreira, 2019; Staiger and Stock, 1997). One motivation is that the vector $(\pi\beta, \beta)$ can be considered a regular parameter and well estimated under mild regularity conditions, whereas β itself is only weakly regular in the

case of weak instruments (Kaji, 2020). As a result, the least-squares estimator for $(\pi\beta, \beta)$ can often be approximated by a normal distribution. One simple example for the model (2.2) to hold is the case where Z is fixed and the rows of $[U, V]$ are i.i.d. or stationary. In this case, the covariance matrix of $\widehat{\psi}$ is

$$\Sigma = (I_2 \otimes (Z'Z)^{-1}Z')\text{Var}[(U', V')'](I_2 \otimes (Z'Z)^{-1}Z')' \tag{2.3}$$

and can be consistently estimated. See Andrews et al. (2019) for a review on the normal approximation to the distribution of $(\widehat{\gamma}, \widehat{\pi})$.

We follow Andrews and Armstrong (2017) and define the unbiased IV estimator. Let

$$\widehat{\delta} = \widehat{\delta}(\widehat{\psi}, \Sigma) = \widehat{\gamma} - \frac{\sigma_{12}}{\sigma_2^2}\widehat{\pi}$$

and

$$\widehat{\tau} = \widehat{\tau}(\widehat{\psi}, \Sigma) = \frac{1}{\sigma_2}\frac{1 - \Phi(\widehat{\pi}/\sigma_2)}{\phi(\widehat{\pi}/\sigma_2)} = \frac{1}{\sigma_2}\Psi(\widehat{\pi}/\sigma_2),$$

where $\Psi(x) = (1 - \Phi(x))/\phi(x)$, and $\Phi(\cdot)$ and $\phi(\cdot)$ are cdf and pdf for the standard normal distribution, respectively. The unbiased IV estimator is

$$\widehat{\beta}_U = \widehat{\beta}_U(\widehat{\psi}, \Sigma) = \widehat{\delta}\widehat{\tau} + \frac{\sigma_{12}}{\sigma_2^2}.$$

It is shown that $E[\widehat{\beta}_U] = \beta$ when $\pi > 0$.

Comment 2.4. The main idea of $\widehat{\beta}_U$ is to use the fact that $\widehat{\tau}$ is an unbiased estimator for $1/\pi$ (Voinov and Nikulin, 1993). Because $\widehat{\delta}$ can be considered the projection of $\widehat{\gamma}$ on the null space of $\widehat{\pi}$, $\widehat{\delta}$ is independent of $\widehat{\pi}$, and thus of $\widehat{\tau}$ as a function of $\widehat{\pi}$. Those facts lead to $E[\widehat{\beta}_U] = \beta$ (Andrews and Armstrong, 2017).

Define the truncated version of the unbiased IV estimator by

$$\tilde{\beta} = \widehat{\delta}\tilde{\tau} + \frac{\sigma_{12}}{\sigma_2^2},$$

where

$$\tilde{\tau} = \frac{1}{\sigma_2}\Psi\left(\frac{\max\{\widehat{\pi}, \pi^*\}}{\sigma_2}\right),$$

and π^* is some truncation parameter. That is, we "winsorize" the unbiased IV estimator according to $\widehat{\pi}$ by the threshold π, when $\widehat{\pi}$ is too small. We do so because $\Psi(\cdot)$ is positive and strictly decreasing on $\mathbb{R}$, and $\Psi(x) \to \infty$ as $x \to -\infty$, which causes $\widehat{\beta}_U$ to have an unbounded second moment. By truncation, we eliminate extreme values of $\widehat{\beta}_U$, which is important in conducting inference.

Example 1. We visualize the truncation in Figure 1. Consider a simple case where $\widehat{\psi} = (\widehat{\gamma}, \widehat{\pi})' \sim N(\psi, I_2)$. Then, the unbiased IV estimator for β is $\widehat{\beta}_U = \widehat{\delta}\widehat{\tau}$, where $\widehat{\delta} = \widehat{\gamma}$ and $\widehat{\tau} = (1 - \Phi(\widehat{\pi}))/\phi(\widehat{\pi})$. Define $\widehat{\pi}_U = 1/\widehat{\tau}$, then $\widehat{\beta}_U = \widehat{\gamma}/\widehat{\pi}_U$; that is, $\widehat{\beta}_U$ is the slope of the line through $(\widehat{\pi}_U, \widehat{\delta})$ and the origin. Then, the proposed truncated estimator $\tilde{\beta}$ is the slope of the line through $(\tilde{\pi}, \widehat{\gamma}) = (\max\{\widehat{\pi}_U, \pi^*\}, \widehat{\delta})$ and the origin.

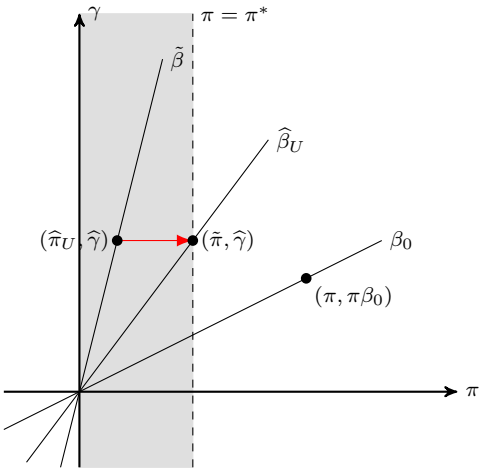

Fig 1: $\tilde{\beta}$ is obtained through winsorizing $\widehat{\pi}_U$ in the gray area.

The following result shows the truncated estimator is nearly unbiased when the truncation is appropriate.

Proposition 1. *Assume β is fixed. Suppose (i) $|\sigma_{12}/\sigma_2^2| < \infty$, (ii) $\pi^*/\sigma_2 \to -\infty$, and (iii) $\pi\pi^*/\sigma_2^2 \to -\infty$. Then, $E[\tilde{\beta}] - \beta \to 0$.*

Comment 2.5. Proposition 1 gives guidance on when the proposed estimator $\tilde{\beta}$ is approximately unbiased. A trivial example is where $\pi^* \to -\infty$ and everything else is constant, in which case, $\tilde{\beta}$ is approaching the unbiased estimator $\widehat{\beta}_U$. Under either the common strong IV asymptotics where $\sigma_2 = O(1/\sqrt{n})$ and π is constant, or the common weak IV asymptotics where $\sigma_2 = O(1/\sqrt{n})$ and $\pi = O(1/\sqrt{n})$, (ii) and (iii) require π^* to be negative and not to shrink as fast as $1/\sqrt{n}$.

3. Fama-MacBeth Inference with Truncated Unbiased IV

Consider a triangular array $\{\{(X_{n,i}, Y_{n,i}, Z_{n,i})\}_{i=1}^{n}\}_{n\geq 1}$ that follows the linear IV model (2.1),

$$\begin{cases} Y_{n,i} = Z_{n,i}\pi_n\beta + U_{n,i}, \\ X_{n,i} = Z_{n,i}\pi_n + V_{n,i}, \end{cases} \tag{3.1}$$

and a sequence of clustering dependence structures $\{\mathcal{C}_n\}_{n\geq 1}$ with $\mathcal{C}_n = \{I_{n,g}\}_{g=1}^{G_n}$ such that $G_n \to \infty$ as $n \to \infty$. That is, for any fixed n, observations are independent across groups but may be dependent within a group. As in section 2, (X, Y, Z) is considered fixed and (U, V) is considered random. The parameter of interest is β, which does not vary with the sample size n. Our goal is to make inferential statement on the hypothesis $H_0 : \beta = \beta_0$. The first-stage coefficient π_n is allowed to change with n but stays the same across groups for each fixed n.[2] In the following presentation, we suppress n for simplicity. All variables and parameters (except β) should be considered a function of n.

3.1. General results

We consider a Fama-MacBeth-type procedure. Namely, we estimate a truncated unbiased IV estimator $\tilde{\beta}_g$ for each group $g \in \{1, \ldots, G\}$, using only $\{(X_i, Y_i, Z_i)\}_{i\in I_g}$. Thus, we obtain a set $\{\tilde{\beta}_g\}_{g=1}^{G}$ of nearly unbiased IV estimators with bounded second moments. Define group-level quantities $\{n_g, \widehat{\psi}_g, \widehat{\delta}_g, \widehat{\tau}_g, \pi_g^*\}_{g=1}^{G}$ accordingly. As in section 2, we assume the group-level reduced-form and first-stage coefficients follow a normal distribution with known covariance Σ_g such that

$$\widehat{\psi}_g = \begin{pmatrix} \widehat{\gamma}_g \\ \widehat{\pi}_g \end{pmatrix} \sim N(\mu_g, \Sigma_g),$$

from which the group-level truncated unbiased IV estimator $\tilde{\beta}_g$ is constructed.[3] Therefore, either the errors (U, V) follow normal distribution or at least a moderate number of observations are in each group. Also, define $\{\sigma_{1,g}, \sigma_{2,g}, \sigma_{12,g}, \mu_{\delta,g}, \sigma_{\delta,g}\}$ such that

$$\Sigma_g = \begin{pmatrix} \sigma_{1,g}^2 & \sigma_{12,g} \\ \sigma_{12,g} & \sigma_{2,g}^2 \end{pmatrix},$$

$$\mu_{\delta,g} = \pi(\beta - \sigma_{12,g}/\sigma_{2,g}^2),$$

$$\sigma_{\delta,g}^2 = \sigma_{1,g}^2 - \sigma_{12,g}^2/\sigma_{2,g}^2.$$

[2] This assumption is made here for simplicity. In principle, we do not need to assume π is the same across different groups, because of the nature of group-level estimation.

[3] In practice, $\{\Sigma_g\}_{g=1}^{G}$ can be estimated by model-based or HAC-type estimators.

For the set of group-level estimates $\{\tilde{\beta}_g\}_{g=1}^G$, define the Fama-MacBeth estimator

$$\bar{\beta} = \frac{1}{G}\sum_{g=1}^{G}\tilde{\beta}_g$$

and the standard error

$$\text{se} = \sqrt{\frac{1}{G(G-1)}\sum_{g=1}^{G}(\tilde{\beta}_g - \bar{\beta})^2}.$$

The corresponding t-statistic is

$$t = \frac{\bar{\beta} - \beta_0}{\text{se}}.$$

We show t is asymptotically normal when the estimator is properly truncated.

Assumption 1. *(i)* $\limsup_n \sup_g |\sigma_{12,g}/\sigma_{2,g}^2| < \infty$*;*

(ii) $\sup_g \pi_g^*/\sigma_{2,g} \to -\infty$*, as* $n \to \infty$*;*

(iii) $\sup_g \pi\pi_g^*/\sigma_{2,g}^2 \to -\infty$*, as* $n \to \infty$*.*

Define $M = \sup_g \Psi(\pi_g^*/\sigma_{2,g})/\sigma_{2,g}$. Conceptually, M guides the overall level of truncation across groups with respect to $\widehat{\tau}_g$. The reason is that $\Psi(\cdot)$ is a strictly decreasing one-to-one map such that $\widehat{\pi}_g \geq \pi_g^*$ if and only if $\widehat{\tau}_g \leq \Psi(\pi_g^*/\sigma_{2,g})/\sigma_{2,g}$.

Assumption 2. *The truncation parameter* M *satisfies*

$$M = o\left(\frac{B}{\bar{\sigma}_\delta(\kappa G)^{1/3}}\right),$$

where

$$B^2 = \sum_{g=1}^{G} E[(\tilde{\beta}_g - E[\tilde{\beta}_g])^2],$$

$$\bar{\sigma}_\delta = \max_g \sigma_{\delta,g},$$

$$\kappa = \max_g K\left(-\frac{3}{2}, \frac{1}{2}; -\frac{\mu_{\delta,g}^2}{2\sigma_{\delta,g}^2}\right)$$

and $K(a, b; z)$ *is Kummer's confluent hypergeometric function.*

Comment 3.1. Assumptions 1 and 2 are high-level conditions that allow for many IV configurations. Both a fixed π (strong IV) or a local drifting sequence that shrinks at the rate of $n^{-1/2}$ (weak IV) are discussed below. Assumption 1 is generally weak. Assumption 1(i) implies σ_{12} and $\sigma_{2,g}^2$ are approximately of the same scale. This assumption is reasonable because they are typically $O(1/n_g)$ with weak dependence. Assumptions 1(ii) & (iii) require both $\pi_g^*/\sigma_{2,g}$ and $\pi\pi_g^*/\sigma_{2,g}^2$ to go to $-\infty$, uniformly. Those assumptions are weak under strong IV as long as π_g^* is negative and bounded away from zero. Under weak IV where

$\pi = O(1/\sqrt{n})$, 1(ii) is weak and 1(iii) holds when $\inf_g n_g/\sqrt{n}$ does not go to zero too fast; that is, the number of groups increases too fast. Assumption 2 puts restrictions on the truncation parameter. Practical suggestions of how the truncation parameters are chosen are given in Appendix A.

Theorem 1. *Under Assumption 1 and 2,* $t \xrightarrow{d} N(0,1)$.

Comment 3.2. This result implies the test $\psi = \mathbb{1}\{|t| > z_{\alpha/2}\}$ delivers an asymptotically correct size at level α, where $z_{\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution. In practice, some quantities in constructing the t-statistic need to be estimated. The implementation details are in Appendix A.

3.2. Strong IV asymptotics

In this subsection, I give the primitive assumptions under which the proposed method delivers valid inference under strong IV.

Define

$$\begin{cases} \bar{\sigma}_2 = \max_g \sigma_{2,g} \\ \underline{\sigma}_2 = \min_g \sigma_{2,g} \end{cases}. \tag{3.2}$$

Assumption S1. (i) $\liminf_n \pi > 0$;

(ii) $\limsup_n \sup_g |\sigma_{12,g}/\sigma_{2,g}^2| < \infty$;

(iii) $\underline{\sigma}_2 M \to \infty$ and $\bar{\sigma}_2 = O(1)$.

Assumption S2. (i) $\bar{\sigma}_2/\underline{\sigma}_2 = O(1)$;

(ii) $M = o(BG^{-1/3})$.

Comment 3.3. S1(i) implies a strong IV sequence and includes the case where π is fixed as $n \to \infty$. S1(ii) is the same as Assumption 1(i). The first half of S1(iii) together with S2(ii) provides guidance on the choice of M. The second half of S1(iii) is weak as long as groups are not diminishing. S2(i) requires that no severe size imbalance exists across groups. Together with the assumptions under the weak IV asymptotics in section 3.3, these assumptions have implications on the selection of the truncation parameter. Practical suggestions are given in Appendix A.

Proposition 2. *Under Assumption S1 and S2 (strong IV sequence), Assumptions 1 and 2 hold.*

3.3. Weak IV asymptotics

In this subsection, I give the primitive assumptions under which the proposed method delivers valid inference under weak IV, where the first-stage strength parameter π follows a drifting sequence towards 0 at the rate of $n^{-1/2}$.

Let $\bar{\sigma}_2$ and $\underline{\sigma}_2$ be defined in equation (3.2). Similarly, define $\bar{\sigma}_\delta = \max_g \sigma_{\delta,g}$ and $\underline{\sigma}_\delta = \min_g \sigma_{\delta,g}$.

Assumption W1. (i) $\pi = \pi_0/\sqrt{n}$;

(ii) $\sup_n \sup_g |\sigma_{12,g}/\sigma^2_{2,g}| < \infty$;

(iii) $n^{-1/2}\Psi^{-1}(\underline{\sigma}_2 M)/\bar{\sigma}_2 \to -\infty$.

Assumption W2. (i) $\pi^2/\underline{\sigma}^2_\delta \to 0$;

(ii) $M = o(B\bar{\sigma}^{-1}_\delta G^{-1/3})$.

Comment 3.4. W1(i) is standard in the weak IV literature (e.g., Staiger and Stock, 1997). In the case of weak dependence with approximately balanced groups, $\sigma_{2,g} = O(n_g^{-1/2})$, so W1(iii) implies $\Psi^{-1}(\underline{\sigma}_2 M)/\sqrt{G} \to -\infty$; $\underline{\sigma}_\delta = O(1/\min_g n_g)$, so W2(i) implies $\max_g n_g/n \to 0$ (cf. $\max_g n_g^2/n \to 0$ in Hansen and Lee, 2019).

Proposition 3. *Under Assumptions W1 and W2 (weak IV sequence), Assumptions 1 and 2 hold.*

4. Simulation

In this section, we study the finite-sample performance of the proposed estimator. In all the following settings, the data generating process follows the linear IV model (3.1), where $n = 900$ and $G = 30$ such that $n/G^2 = 1$, which deviates from the usual asymptotics. The null hypothesis is $H_0 : \beta = 0$. For each setting, 1,000 replications are conducted to calculate the empirical rejection rate.

For each setting, we observe $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ and a partition $\{I_g\}_{g=1}^G$ of $\{i\}_{i=1}^n$. Let consecutive observations belong to the same group; that is, $I_1 = \{1, 2, \ldots, |I_1|\}$, $I_2 = \{|I_1| + 1, \ldots, |I_1| + |I_2|\}$, and so on, where $|\cdot|$ is cardinality. The data are drawn according to the

following process:

$$
\begin{aligned}
&Y_i = Z_i\pi\beta + U_i \\
&X_i = Z_i\pi + V_i \\
&\begin{pmatrix} U_i \\ V_i \end{pmatrix} \sim N\left(0, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right), \text{ if } i = 1 + \sum_{h=1}^{g} |I_h| \text{ for some } g = 0, 1, \ldots, G-1 \\
&\begin{pmatrix} U_i \\ V_i \end{pmatrix} = 0.5 \begin{pmatrix} U_{i-1} \\ V_{i-1} \end{pmatrix} + \sqrt{1 - 0.5^2} \begin{pmatrix} \varepsilon_i^U \\ \varepsilon_i^V \end{pmatrix}, \text{ if } i \neq 1 + \sum_{h=1}^{g} |I_h| \text{ for any } g = 0, 1, \ldots, G-1 \\
&\begin{pmatrix} \varepsilon_i^U \\ \varepsilon_i^V \end{pmatrix} \sim N\left(0, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right) \text{ and is i.i.d. across } i.
\end{aligned}
$$

Also, each dimension of the k-dimensional instruments Z_i takes one draw from the distribution of $\{U_i\}_{i=1}^n$ and is fixed across replications. Thus, (U_i, V_i) within each group follows an AR(1) process and is independent across different groups. The parameters $(\beta, \pi, \{I_g\}_{g=1}^G, k)$ vary accordingly across settings.

4.1. Debiasing and truncation

We first investigate three Fama-MacBeth-type inferential procedures and show the necessity of debiasing and truncation. We consider the t-test on group-level 2SLS estimators (FM), the t-test on group-level unbiased IV estimators(FMU), and the proposed t-test on group-level truncated unbiased IV estimators (FMUT), with with truncation parameter selected as suggested in Appendix A. The full-sample 2SLS with CCE estimates of standard errors is also reported for comparison.

In this experiment, we have five instrumental ($k = 5$) and one endogenous variable. Groups are imbalanced in sizes, with five groups of 90 observations and 25 groups of 18 observations. For each group, the observations follow an AR(1) process as described before. The first-stage coefficient is $\pi = (0.1, 0.1, 0.1, 0.1, 0.1)'/\sqrt{5}$ such that $\|\pi\|_2 = 0.1$.

The power curves are reported in Figure 1. Estimators used in CCE and FM are both biased. FM has large bias between the two, because it uses group-level 2SLS estimators with much larger finite-sample bias than the full-sample estimator. FMU is less powerful with FMUT, because the unbiased IV estimator does not have a bounded second moment, such that the resulting t-statistic has a tail that is too large.

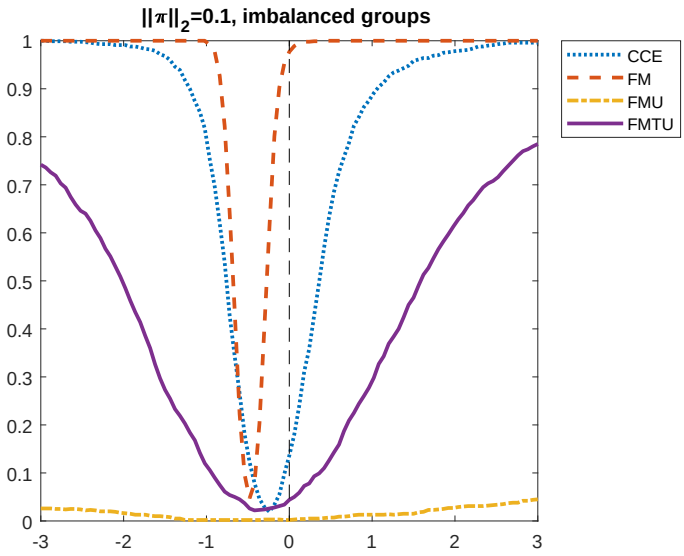

Fig 2: Power comparison among Fama-MacBeth procedures ($\alpha = 0.05$)

4.2. Comparison with other methods

Here, we compare the proposed method with the existing inferential procedure. We consider the "clustered standard error" approach (CCE) and the natural extension of Anderson-Rubin test to our settings (AR-CCE). To implement the AR-CCE method, we apply CCE to the regression of $Y - X\beta_0$ on Z, where β_0 is the hypothesized value as in $H_0 : \beta = \beta_0$. In our case, we test $H_0 : \beta = 0$, so AR-CCE is equivalent to performing CCE to test the hypothesis $H_0 : \gamma = 0$ in the regression $Y = Z\gamma + U$.

We look at several configurations. The number of instruments k varies in the set $\{1, 5, 10\}$. The first-stage strength is chosen such that $\|\pi\|_2 \in \{0.1, 0.5\}$, with $\pi = \|\pi\|_2 \iota_k/\sqrt{k}$ and ι_k being a k-vector of 1's. For example, in the case of $\|\pi\|_2 = 0.1$ and $k = 5$, we have $\pi = (0.1, 0.1, 0.1, 0.1, 0.1)'/\sqrt{5}$. We also consider both balanced and imbalanced groups. In the balanced-group case, we have 30 groups of 30 observations; in the imbalanced-group case, we have 5 groups of 90 observations and 25 groups of 18 observations.

The sizes are reported in Table 2 and the power curves are in Figures 3, 4, and 5. Among all methods, only FMUT is able to deliver a robust inference result at the null across all settings. CCE displays a noticeable bias under $\|\pi\|_2$ and over-identification. AR-CCE is robust to weak instruments, but not under group imbalance.

TABLE 2
Summary ($\alpha = 0.05$)

			Balanced Groups			Imbalanced Groups		
			Median	MAD	Size	Median	MAD	Size
$k = 1$	$\pi = 0.5$	CCE	0.002	0.050	0.040	0.001	0.049	0.062
		AR-CCE	-	-	0.040	-	-	0.060
		FMTU	-0.010	0.057	0.039	-0.035	0.087	0.052
	$\pi = 0.1$	CCE	0.010	0.254	0.043	0.007	0.251	0.048
		AR-CCE	-	-	0.040	-	-	0.060
		FMTU	0.032	0.303	0.066	0.002	0.354	0.048
$k = 5$	$\pi = 0.5$	CCE	0.011	0.051	0.051	0.010	0.052	0.063
		AR-CCE	-	-	0.037	-	-	0.092
		FMTU	-0.068	0.110	0.033	-0.078	0.141	0.034
	$\pi = 0.1$	CCE	0.194	0.256	0.119	0.202	0.256	0.136
		AR-CCE	-	-	0.037	-	-	0.092
		FMTU	0.073	0.260	0.047	0.029	0.312	0.044
$k = 10$	$\pi = 0.5$	CCE	0.022	0.051	0.076	0.021	0.050	0.082
		AR-CCE	-	-	0.063	-	-	0.116
		FMTU	-0.055	0.105	0.046	-0.074	0.131	0.037
	$\pi = 0.1$	CCE	0.277	0.284	0.259	0.279	0.285	0.267
		AR-CCE	-	-	0.063	-	-	0.116
		FMTU	0.067	0.247	0.075	0.026	0.294	0.046

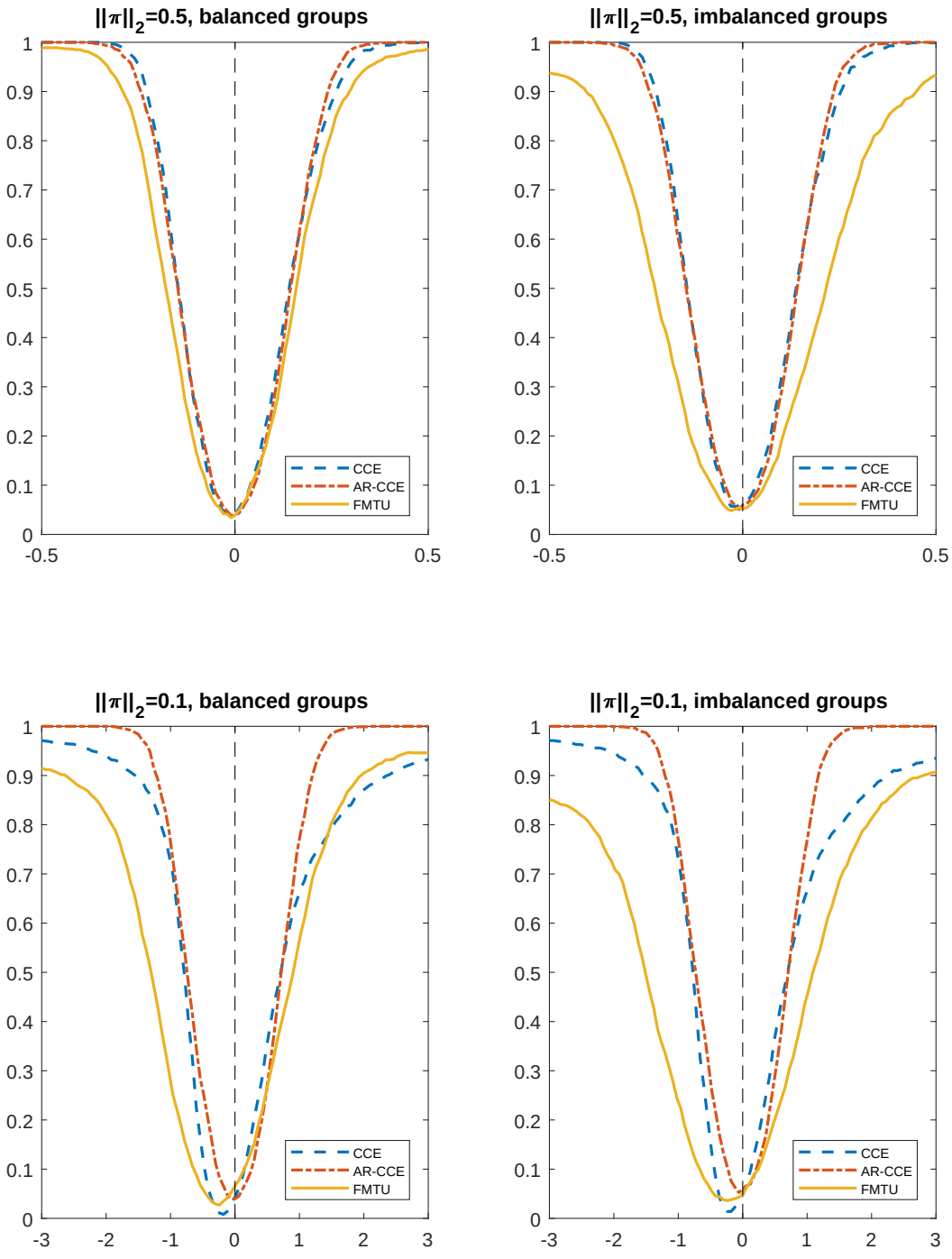

Fig 3: Power curves with nominal size $\alpha = 0.05$ and $k = 1$.

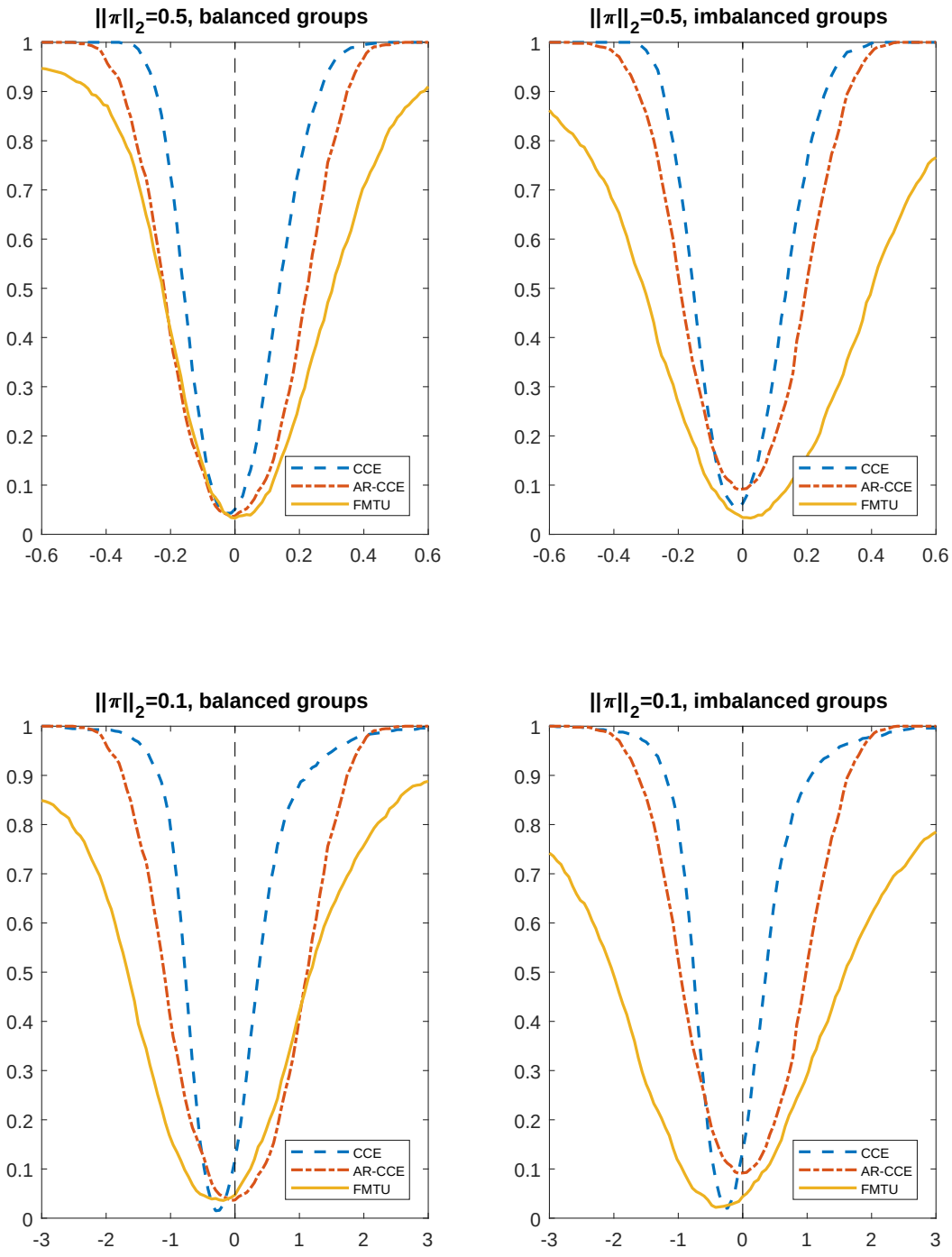

Fig 4: Power curves with nominal size $\alpha = 0.05$ and $k = 5$.

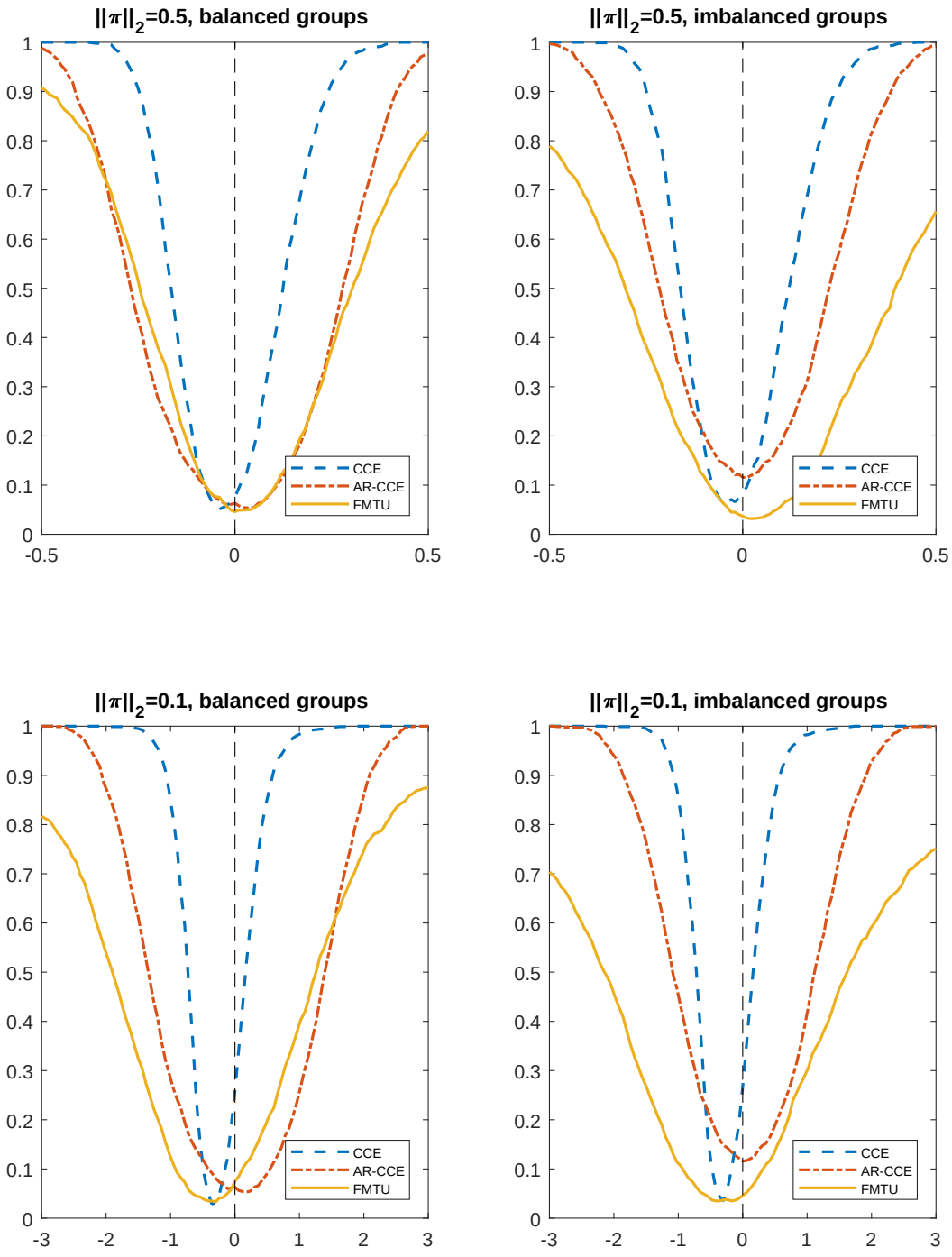

Fig 5: Power curves with nominal size $\alpha = 0.05$ and $k = 10$.

5. Empirical Application: Urban Geometry in India

In this section I use the proposed inferential method to study the effect of city shape on population density. The data used in this section were originally collected and analyzed in Harari (2020). The shape of a city affects its compactness, where compactness is measured by how convenient its residents travel for daily activities. Ideally, a compact city should look like a circle, whereas cities develop into various shapes for many reasons including geographic constraints. Compact cities are attractive to residents because their daily activities operate more efficiently than those in cities that are less compact. This argument suggests the hypothesis that more compact cities should have higher population density. However, city shape is highly endogenous because it is the outcome of economic activities. Harari (2020) proposes a solution to this edogeneity problem by utilizing geographic obstacles such as mountains and lakes as an instrument. I apply the method proposed in this paper, FMTU, in order to obtain a more robust set of empirical results.

5.1. Methodology

To facilitate quantitative analysis, Harari (2020) proposes a shape metric that is based on the average distance between any two points in a polygon, in order to measure the compactness of a city. Namely, the Shape index is defined by

$$Shape = \frac{1}{B(B-1)} \sum_{i=1}^{B} \sum_{j \neq i} d_{ij},$$

where i and j stand for two points sampled from interior points of the city, d_{ij} is the Euclidean distance between i and j, and B is the number of sampled points. We consider the *Normalized Shape* obtained by dividing *Shape* by the radius of the Equivalent Area Circle (EAC), where EAC is the circle with the same area as the city. That is, *Normalized Shape* measures how much the city shape is different from a circle.

The instrument is the *Normalized Shape* index for the projected city. Constructing the projected city is a two-step procedure. First, predict the area that a city should occupy in a given year, based on its projected historical population growth. Second, predict the shape of the city given projected area and geographic constraints. We then instrument the shape of the actual city with shape of the potential one.

I consider the same setup as Harari (2020) does. The regression of interest is

$$\Delta Population\ density = \alpha + \beta \Delta Normalized\ shape + U,$$

where the dependent variable is the change in population density from 1951 to 2011, the endogenous variable is the change in the city shape index from 1950 to 2010, and the instrument is the change in the city-shape index for the projected city expansion from 1950 to 2010. The instrument is the difference of *Normalized Shape* for projected cities. This model can be interpreted as a *difference-in-difference* design with continuous treatment and endogeneity.

I consider the potential dependence among observations by applying the framework suggested by Cao et al. (2019). Namely, I first apply k-medoids to generate a partition of cities using their geographic locations, and then use the given clustering structure to perform the proposed group-based inference method. I use this method to obtain inference results robust to spatial correlation. Factors that affect population density in a city may include climate, culture, economy, personal preferences, etc. Those factors are multidimensional and the natural administrative division[4] does not necessarily capture the underlying dependence structure. That is, cities in neighboring states may be highly correlated in factors that contribute to population density.

The idea of Cao et al. (2019) is to use k-mediods, a clustering algorithm, to generate a partition of observations that helps obtain robust results in group-based inferential methods. Cao et al. (2019) show the clustering generated by k-medoids satisfies *group-balance* and *diminishing-boundary*. The former, *group-balance*, requires there is no diminishingly small group, and the latter, *diminishing-boundary*, requires across-group dependence is approximately ignorable. The algorithmic details are repsented in Appendix E. I apply k-medoids to generate a clustering of 10 group. The resulting structure is visualized in Figure 6.

5.2. Results

Table 3 compares the original results in Table 8 of Harari (2020) with those obtained from the proposed method. Note that the first-stage t-statistic being 5.311 does not imply we can ignore the instrument strength. Lee et al. (2020) show that in order to have a level-0.05 second-stage test in a single IV model, the first-stage F-statistic needs to exceed 104.7, which translates into a t-statistic of 10.23. Comparing 2SLS and the proposed method of this paper, the estimates are qualitatively similar (-171.79 vs. -199.26), whereas the standard errors are quite different. Although the new p-value still suggests rejecting the null at some levels such as 0.1, the implication is vastly different from the original p-value, suggesting that including spatial correlation in analysis is crucial.

[4]India is a federal union comprising 28 states and 8 union territories.

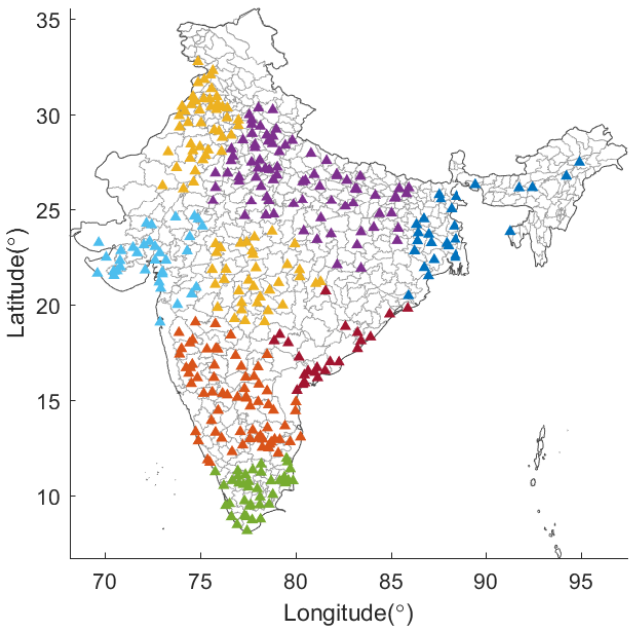

Fig 6: Partition of cities in India by k-medoids using 10 clusters. Distances are Euclidean distances based on latitude and longitude coordinates recorded at cities' centroids. Different colors correspond to different clusters in the partition. Marks are plotted at city centroids.

TABLE 3

	Δ Normalized shape	Δ Population density	
	First stage	2SLS	FMUT
	(1)	(2)	(3)
Δ Potential normalized shape	0.0996		
	(0.0188)		
Δ Normalized shape		-171.8	-199.3
		(37.32)	(88.35)
t-stat	5.311	-4.603	-2.255
p-value		0.000	0.051
Observations	351	351	351

Notes: This table reports estimates of the impacts of city shape on population. Column 1 reports the first-stage results. Column 2 reports 2SLS results with the White robust standard error. Column 3 reports results from the truncated unbiased Fama-MacBeth method. The p-value for FMUT is calculated using a Student's t-distribution of 9 degrees of freedom.

6. Conclusion

In the setting of IV regression, this paper proposes an inferential method that is based on the idea of Fama-MacBeth estimation, in order to deal with weak IV and heterogeneous clustering dependence. To overcome the finite-sample bias of IV regression, the group-level estimator is a truncated version of the unbiased IV estimator proposed by Andrews and Armstrong (2017). I give high-level conditions under which the proposed method is asymptotically valid. Asymptotic validity is also shown under both strong and weak IV sequences. Finite-sample performance is shown by simulation. The proposed method is applied to study the effect of city compactness on population density.

Appendix A: Implementation

In this section, I describe the details of implementing the proposed procedure. The idea is simply to replace each quantity by its sample analog. Section A.1 discusses the case with only one instrument. Section A.2 covers the case with more than one instrument.

A.1. One single instrument

The algorithm consists of three steps: group-level estimation, debiasing and truncation, and a t-test.

Step 1 We fix some group g and only use observations in this group. Let the corresponding group-level estimators be $\widehat{\psi}_g = (\widehat{\psi}_{1,g}, \widehat{\psi}_{2,g})'$ as in (2.2), and the residuals be $\{\widehat{U}_i, \widehat{V}_i\}_{i\in I_g}$. Let $\widehat{\Lambda}_g$ be a heteroskedasticity and autocorrelation correction estimator (HAC) of

$$Var\left[\frac{1}{\sqrt{n_g}}\sum_{i\in I_g}\begin{pmatrix} Z_iU_i \\ Z_iV_i \end{pmatrix}\right].$$

In the simulation section, we use the Newey-West estimator with $\lfloor 4(T/100)^{1/4}\rfloor$ lags (Newey and West, 1987). The estimator for $Var[\widehat{\psi}]$ is thus

$$\widehat{\Sigma}_g = \begin{pmatrix} Q_{ZZ,g}^{-1} & 0 \\ 0 & Q_{ZZ,g}^{-1} \end{pmatrix} \widehat{\Lambda}_g \begin{pmatrix} Q_{ZZ,g}^{-1} & 0 \\ 0 & Q_{ZZ,g}^{-1} \end{pmatrix},$$

where $Q_{ZZ,g} = n_g^{-1}\sum_{i\in I_g} Z_iZ_i'$. The group-level $(\widehat{\delta}, \widehat{\tau})$ is given by

$$\widehat{\delta}_g = \widehat{\delta}(\widehat{\psi}_g, \widehat{\Sigma}_g),\ \widehat{\tau}_g = \widehat{\tau}(\widehat{\psi}_g, \widehat{\Sigma}_g),$$

and the unbiased IV is $\widehat{\beta}_g = \widehat{\delta}\widehat{\tau} + \widehat{\sigma}_{12,g}/\widehat{\sigma}_{2,g}^2$.

Step 2 Consider a uniform truncation parameter where $\pi_g^* = \pi^*$ for each g. Let

$$\pi_{SIV}^* = \min_g \frac{1}{\sqrt{n_g}} \Psi^{-1}\left(c\sqrt{\frac{\bar{n}}{n_g}}\right)$$

and

$$\pi_{WIV}^* = \min_g \frac{1}{\sqrt{n_g}} \Psi^{-1}\left(\sqrt{\frac{\underline{n}}{n_g}} \Psi\left(-c\sqrt{\frac{n}{\bar{n}}}\right)\right),$$

where $\bar{n} = \max_g n_g$ and $\underline{n} = \min_g n_g$. The former is suggested by assumptions under the strong IV asymptotics in section 3.2 and the latter by the weak IV asymptotics in section 3.3. The truncation parameter is chosen to be $\pi^* = \min\{\pi_{SIV}^*, \pi_{WIV}^*\}$. In practice, I recommend using $c = 10$. Using the selected threshold π^*, we can obtain a set of group-level truncated unbiased IV estimators $\{\tilde{\beta}_g\}_{g=1}^G$.

Step 3 We apply the t-test to the set of group-level estimators $\{\tilde{\beta}_g\}_{g=1}^G$. Namely, let

$$t = \frac{\bar{\beta} - \beta_0}{\text{se}},$$

where

$$\bar{\beta} = \frac{1}{G}\sum_{g=1}^G \tilde{\beta}_g$$

and

$$\text{se} = \sqrt{\frac{1}{G(G-1)}\sum_{g=1}^G (\tilde{\beta}_g - \bar{\beta})^2}.$$

We reject the null hypothesis $H_0 : \beta = \beta_0$ if $|t| > cv$, where cv is the $(1-\alpha)$-quantile of the t-distribution of $G-1$ degrees of freedom.

A.2. Multiple instruments

When multiple instruments are available, we follow Andrews and Armstrong (2017) and use a weighted average of unbiased IV estimators with respect to all instruments. Namely, for the j-th instrument, we perform Steps 1 and 2 as in section A.1 and obtain the j-th unbiased IV estimator $\tilde{\beta}_{g,j}$. The group-level estimator is then given by

$$\tilde{\beta}_g = \sum_j w_j \tilde{\beta}_{g,j},$$

where $\{w_j\}_{j=1}^k$ is a set of weights that sum up to one. See Andrews and Armstrong (2017) for a discussion on optimal weight selection. In the simulation section, $\{w_j\}_{j=1}^k$ are simply chosen to be equal weights. Finally, we follow Step 3 in section A.1 using $\{\tilde{\beta}_g\}_{g=1}^G$.

Appendix B: Truncation Parameter Choices

In this section, we investigate the impact of the choice of the truncation parameter. As in Appendix A, we recommend using $\pi^* = \min\{\pi^*_{SIV}, \pi^*_{WIV}\}$ with $c = 10$ as the truncation parameter. We look into different choices of c in this experiment.

The data regenerating process is the same as in section 4.1. FMUT methods with three different values of c are reported. The CCE method is also reported for comparison. The power curves are shown in Figure 7. Generally, the proposed method is quite robust to the choice of the truncation parameter in terms of null rejection rate. Moreover, Figure 7 exhibits a "bias-variance" tradeoff. That is, a smaller c corresponds to high power but causes more bias.

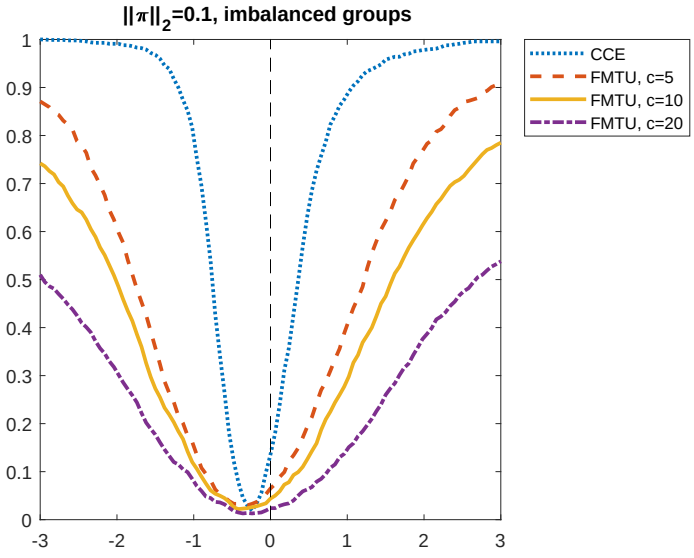

Fig 7: Power comparison among truncation-parameter choices ($\alpha = 0.05$)

Appendix C: Useful Results

In this section, I present some results that are useful for proofs in Appendix D.

Lemma 1. *$E[\widehat{\tau}\mathbb{1}\{\widehat{\pi} \geq \pi^*\}] = \eta\pi^{-1}$, where*

$$\eta = (1 - \Phi((\pi^* - \pi)/\sigma_2)) - (1 - \Phi(\pi^*/\sigma_2))\exp(\pi\pi^*/\sigma_2^2 - \pi^2/(2\sigma_2^2)).$$

Proof.

$$\begin{aligned}
& E[\widehat{\tau}\mathbb{1}\{\widehat{\pi} \geq \pi^*\}] \\
= \; & E\left[\frac{1}{\sigma_2} \cdot \frac{1-\Phi(\widehat{\pi}/\sigma_2)}{\phi(\widehat{\pi}/\sigma_2)}\mathbb{1}\{\widehat{\pi} \geq \pi^*\}\right] \\
= \; & \int_{\pi^*/\sigma_2}^{\infty} \frac{1}{\sigma_2} \cdot \frac{1-\Phi(x)}{\phi(x)}\phi(x-\pi/\sigma_2)\mathrm{d}x \\
= \; & \frac{1}{\sigma_2}\int_{\pi^*/\sigma_2}^{\infty} (1-\Phi(x))\exp(x\pi/\sigma_2 - \pi^2/(2\sigma_2^2))\mathrm{d}x \\
= \; & \frac{1}{\pi}\exp(-\pi^2/(2\sigma_2^2))\left((1-\Phi(x))\exp\left(\frac{\pi}{\sigma_2}x\right)\Big|_{\pi^*/\sigma_2}^{\infty} - \int_{\pi^*/\sigma_2}^{\infty}\exp(x\pi/\sigma_2)\mathrm{d}(1-\Phi(x))\right) \\
= \; & \frac{\eta}{\pi}.
\end{aligned}$$

The fourth equality is integration by parts. □

Kummer's confluent hypergeometric functions

$$K(a,b,z) = \sum_{k=0}^{\infty} \frac{a^{\bar{k}}}{b^{\bar{k}}}\frac{z^k}{k!},$$

where the rising factorial is defined by

$$x^{\bar{k}} = x(x+1)\ldots(x+n-1).$$

For $z < 0$,

$$K(a,b,z) = \frac{\Gamma(b)}{\Gamma(b-a)}(-z)^{-a}[1+O(|z|^{-1})],$$

where the Gamma function is

$$\Gamma(x) = \int_0^{\infty} u^{x-1}e^{-u}\mathrm{d}u.$$

See Abramowitz and Stegun (1965) for reference.

Appendix D: Proofs

Proof of Proposition 1. Let η be defined as in Lemma 1. Note

$$\begin{aligned}
|E[\tilde{\beta}] - E[\widehat{\beta}_U]| &= |E[\widehat{\delta}(\tilde{\tau}-\widehat{\tau})]| \\
&= |E[\widehat{\delta}]| \cdot |E[\tilde{\tau}-\widehat{\tau}]| \\
&= |E[\widehat{\delta}]| \cdot E[(\widehat{\tau}-\tau^*)\mathbb{1}\{\widehat{\pi} < \pi^*\}] \\
&\leq |E[\widehat{\delta}]| \cdot E[\widehat{\tau}\mathbb{1}\{\widehat{\pi} < \pi^*\}] \\
&= |\pi(\beta - \sigma_{12}/\sigma_2^2)| \cdot (\pi^{-1} - \eta\pi^{-1}) \\
&= |\beta - \sigma_{12}/\sigma_2^2| \cdot (1-\eta) \\
&\to 0.
\end{aligned}$$

The second equality uses the independence between $\widehat{\delta}$ and $\widehat{\pi}$, which implies the independence between $\widehat{\delta}$ and functions of $\widehat{\pi}$. The inequality is because Ψ is strictly decreasing, and thus $0 \leq \widehat{\tau} - \tau^* \leq \widehat{\tau}$ under the event $\widehat{\pi} < \pi^*$. The fourth equation is by Lemma 1 and the fact that $E[\widehat{\tau}] = 1/\pi$. The convergence is because $\eta \to 1$ under (i), (ii), and (iii), and σ_{12}/σ_2^2 is bounded. □

Proof of Theorem 1. By Assumption 1 and following the proof of Proposition 1,

$$\sup_g |E[\tilde{\beta}_g] - \beta| \to 0.$$

So

$$t = t^* + \frac{G^{-1}\sum_{g=1}^G (E[\tilde{\beta}_g] - \beta)}{\text{se}} = t^* + o_p(1),$$

where

$$t^* = \frac{G^{-1}\sum_{g=1}^G (\tilde{\beta}_g - E[\tilde{\beta}_g])}{\text{se}}.$$

Under Assumption 2, for some absolute constant C,

$$\begin{aligned} \sup_x |\mathrm{P}(t^* < x) - \Phi(x)| &\leq CB^{-3}\sum_{g=1}^G E[|\tilde{\beta}_g - \beta|^3] \\ &\lesssim CB^{-3}G \max_g E[|\widehat{\delta}_g|^3]E[|\tilde{\tau}_g|^3] \\ &\lesssim CB^{-3}G\bar{\sigma}_\delta^3\kappa M^3 \\ &= o(1). \end{aligned} \tag{D.1}$$

The inequality is by a Berry-Esseen bound for Student's statistic in Bentkus et al. (1996). The third line uses a representation of the third raw absolute moment of normal distribution (e.g., see Winkelbauer, 2012). Combining (D.1) with $t = t^* + o_p(1)$, we obtain $t \xrightarrow{d} N(0,1)$. □

Proof of Proposition 2. Assumption 1(i) holds automatically by Assumption S1(ii). For 1(ii), note

$$\max_g \frac{\pi_g^* - \pi}{\sigma_{2,g}} = \max_g \Psi^{-1}(\sigma_{2,g}M) - \frac{\pi}{\sigma_{2,g}} \leq \Psi^{-1}(\underline{\sigma}_2 M) \to -\infty$$

by S1(iii) and the fact that $\Psi^{-1}(x) \to -\infty$ as $x \to \infty$. For 1(iii), note

$$\max_g \frac{\pi\pi_g^*}{\sigma_{2,g}^2} = \max_g \frac{\pi\Psi^{-1}(\sigma_{2,g}M)}{\sigma_{2,g}} \leq \frac{\pi\Psi^{-1}(\underline{\sigma}_2 M)}{\bar{\sigma}_2} \to -\infty$$

by S1(iii).

To see Assumption 2, first note for some constant C_1, C_2,

$$K\left(-\frac{3}{2}, \frac{1}{2}; -\frac{\mu_\delta^2}{2\sigma_\delta^2}\right) \leq C_1\left(\frac{\mu_\delta^2}{2\sigma_\delta^2}\right)^{3/2} + C_2\left(\frac{\mu_\delta^2}{2\sigma_\delta^2}\right)^{1/2} \lesssim C_1\underline{\sigma}_\delta^{-3}$$

by properties of Kummer's confluent hypergeometric function (e.g., 13.1.5 of Abramowitz and Stegun, 1965). Therefore,

$$M = o\left(\frac{B}{\underline{\sigma}_2(\bar{\sigma}_2^{-3}G)^{1/3}}\right) = o\left(\frac{B}{\underline{\sigma}_2(\kappa G)^{1/3}}\right).$$

□

Proof of Proposition 3. Assumptions 1(i) and 1(ii) follow the same reasoning as in the proof of Proposition 2. For 1(iii), note

$$\max_g \frac{\pi\pi_g^*}{\sigma_{2,g}^2} \lesssim \frac{\Psi^{-1}(\underline{\sigma}_2 M)}{\sqrt{n}\bar{\sigma}_2} \to -\infty.$$

To see Assumption 2, for some constant C,

$$K\left(-\frac{3}{2}, \frac{1}{2}; -\frac{\mu_\delta^2}{2\sigma_\delta^2}\right) \leq 1 + C\left|-\frac{\mu_\delta^2}{2\sigma_\delta^2}\right| = 1 + O\left(\frac{\pi^2}{\underline{\sigma}_\delta^2}\right), \tag{D.2}$$

by properties of Kummer's confluent hypergeometric function. Combining (D.2) with W2(ii) gives Assumption 2. □

Appendix E: *k*-Medoids Algorithm

This section states the k-medoids algorithm used in Section 5. Let (X, d) be a metric space with a finite set of locations X and a distant metric d. For some cluster $\mathsf{C} \subseteq \mathsf{X}$ and medoid $i \in \mathsf{X}$, define the cost to be

$$\text{cost}(\mathsf{C}, i) = \sum_{j \in \mathsf{C}} d(i, j)^2.$$

Let $\mathcal{C} = \{\mathsf{C}_g\}_{g=1}^G$ be a partion of X, i.e., clustering strucutre. Define the total cost for $\mathcal{C}$ with a set of medoids $\{i_\mathsf{C}\}_{\mathsf{C}\in\mathcal{C}}$ by summing over clusters

$$\text{total cost}(\mathcal{C}, \{i_\mathsf{C}\}_{\mathsf{C}\in\mathcal{C}}) = \sum_{\mathsf{C}\in\mathcal{C}} \text{cost}(\mathsf{C}, i_\mathsf{C}).$$

Algorithm k-medoids Clustering

Input. (X, d), G.

Procedure.

1. Initialize cluster centroids $\{i_1, ..., i_G\} \subset \mathsf{X}_n$ arbitrarily.
2. While total cost decreases,
   a. For each $k \leqslant G$, for each $j \notin \{i_1, ..., i_G\}$ compute the cost with new medoids $\{i_1, ..., i_{k-1}, j, i_{k+1}, ...i_G\}$;
   b. Assign new medoids membership if the new set of medoids has less total cost.

Output. $\mathcal{C}$ with $|\mathcal{C}| = G$.

References

Abramowitz, M. and Stegun, I. A. (1965). *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*. Dover Publications.

Anderson, T. W. and Rubin, H. (1949). Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations. *The Annals of Mathematical Statistics*, 20(1):46–63.

Andrews, D. W. K., Moreira, M. J., and Stock, J. H. (2006). Optimal Two-Sided Invariant Similar Tests for Instrumental Variables Regression. *Econometrica*, 74(3):715–752.

Andrews, I. and Armstrong, T. B. (2017). Unbiased instrumental variables estimation under known first-stage sign. *Quantitative Economics*, 8(2):479–503.

Andrews, I. and Mikusheva, A. (2016). Conditional Inference With a Functional Nuisance Parameter. *Econometrica*, 84(4):1571–1612.

Andrews, I., Stock, J. H., and Sun, L. (2019). Weak Instruments in Instrumental Variables Regression: Theory and Practice. *Annual Review of Economics*, 11(1):727–753.

Bentkus, V., Bloznelis, M., and Götze, F. (1996). A Berry-Esséen bound for student's statistic in the non-i.i.d. case. *Journal of Theoretical Probability*, 9(3):765–796.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How Much Should We Trust Differences-In-Differences Estimates? *The Quarterly Journal of Economics*, 119(1):249–275.

Bester, C. A., Conley, T. G., and Hansen, C. B. (2011). Inference with dependent data using cluster covariance estimators. *Journal of Econometrics*, 165(2):137–151.

Cameron, A. C. and Miller, D. L. (2015). A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources*, 50(2):317–372.

Canay, I. A., Romano, J. P., and Shaikh, A. M. (2017). Randomization Tests Under an Approximate Symmetry Assumption. *Econometrica*, 85(3):1013–1030.

Cao, J., Hansen, C., Kozbur, D., and Villacorta, L. (2019). Inference for Dependent Data with Cluster Learning. *Working paper*.

Coibion, O., Gorodnichenko, Y., and Koustas, D. (2017). Consumption Inequality and the Frequency of Purchases. *American Economic Journal: Macroeconomics, forthcoming*.

Dell, M. (2012). Path Dependence in Development: Evidence from the Mexican Revolution. *Working paper*.

Deryugina, T., Heutel, G., Miller, N. H., Molitor, D., and Reif, J. (2019). The mortality and medical costs of air pollution: Evidence from changes in wind direction. *American Economic Review*, 109(12):4178–4219.

Fama, E. F. and MacBeth, J. D. (1973). Risk, Return, and Equilibrium: Empirical Tests. *Journal of Political Economy*, 81(3):607–636.

Ferman, B. (2019). A simple way to assess inference methods.

Ferman, B. and Pinto, C. (2019). Inference in Differences-in-Differences with Few Treated Groups and Heteroskedasticity. *The Review of Economics and Statistics*, 101(3):452–467.

Hagemann, A. (2019a). Permutation inference with a finite number of heterogeneous clusters. *arXiv preprint arXiv:1907.01049*.

Hagemann, A. (2019b). Placebo inference on treatment effects when the number of clusters is small. *Journal of Econometrics*, 213(1):190–209.

Hansen, B. E. and Lee, S. (2019). Asymptotic theory for clustered samples. *Journal of Econometrics*, 210(2):268–290.

Hansen, C. B. (2007). Asymptotic properties of a robust variance matrix estimator for panel data when T is large. *Journal of Econometrics*, 141(2):597–620.

Harari, M. (2020). Cities in Bad Shape: Urban Geometry in India. *American Economic Review*, 110(8):2377–2421.

Ibragimov, R. and Müller, U. K. (2010). t-Statistic Based Correlation and Heterogeneity Robust Inference. *Journal of Business & Economic Statistics*, 28(4):453–468.

Kaji, T. (2020). Theory of Weak Identification in Semiparametric Models.

Kleibergen, F. (2002). Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica*, 70(5):1781–1803.

Lee, D. L., McCrary, J., Moreira, M. J., and Porter, J. (2020). Valid t-ratio Inference for IV.

MacKinnon, J. G. and Webb, M. D. (2017). Wild Bootstrap Inference for Wildly Different Cluster Sizes. *Journal of Applied Econometrics*, 32(2):233–254.

Mills, B. (2019). Inference Under First-Stage Sign Information in the Instrumental Variables Model. *Working paper*.

Moreira, H. and Moreira, M. J. (2019). Optimal two-sided tests for instrumental variables regression with heteroskedastic and autocorrelated errors. *Journal of Econometrics*, 213(2):398–433.

Moreira, M. J. (2003). A Conditional Likelihood Ratio Test for Structural Models. *Econometrica*, 71(4):1027–1048.

Newey, W. K. and West, K. D. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55(3):703.

Pesaran, M. H., Pesaran, M. H., Shin, Y., and Smith, R. P. (1999). Pooled Mean Group

Estimation of Dynamic Heterogeneous Panels. *Journal of the American Statistical Association*, 94(446):621–634.

Pesaran, M. H. and Smith, R. (1995). Estimating long-run relationships from dynamic heterogeneous panels. *Journal of Econometrics*, 68(1):79–113.

Staiger, D. and Stock, J. H. (1997). Instrumental Variables Regression with Weak Instruments. *Econometrica*, 65(3):557–586.

Voinov, V. G. and Nikulin, M. S. (1993). *Unbiased Estimators and Their Applications, Vol. 1: Univariate Case*. Kluwer Academic Publishers, Dordrecht.

Winkelbauer, A. (2012). Moments and Absolute Moments of the Normal Distribution. *arXiv preprint arXiv:1209.4340*.

Young, A. (2019). Channeling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *Quarterly Journal of Economics*, 134(2):557–598.