Inference with Few Heterogeneous Clusters*

Rustam Ibragimov
Imperial College Business School
Imperial College London

Ulrich K. Müller
Economics Department
Princeton University

April 2015

Abstract

Suppose estimating a model on each of a small number of potentially heterogeneous clusters yields approximately independent, unbiased and Gaussian parameter estimators. We make two contributions in this set-up. First, we show how to compare a

*The authors would like to thank three anonymous referees and seminar and conference participants at various universities and conferences for helpful comments. Ibragimov gratefully acknowledges partial support via NSF grant SES-0820124 and grants from the GDN-SEE and CIS Research Competition, the Russian Ministry of Education and Science (Innopolis University) and the Russian Government Program of Competitive Growth of Kazan Federal University (Higher Institute of Information Technologies and Information Systems), and Müller gratefully acknowledges support by the NSF via grant SES-0518036. We are indebted to Nail Bakirov (1952-2010) and Daniyar Mushtari (1945-2013) for inspiring discussions, comments and attention to our work. We are also thankful to Aprajit Mahajan for sharing data and useful discussions, and to Chenchuan (Mark) Li for outstanding research assistance.

scalar parameter of interest between treatment and control units using a two-sample t-statistic, extending previous results for the one-sample t-statistic. Second, we develop a test for the appropriate level of clustering, which tests the null hypothesis that clustered standard errors from a much finer partition are correct. We illustrate the approach by revisiting empirical studies involving clustered, time series and spatially correlated data.

JEL classification: C12, C14, C32

Keywords: Dependence, clustered standard errors, two-sample t-statistic, variance heterogeneity, difference-in-difference, structural breaks

1 Introduction

The use of clustered standard errors has become widespread in empirical economics. For instance, Bertrand, Duflo and Mullainathan (2004) stress the importance of allowing for time series correlation in panel difference-in-difference applications. The usual asymptotic justification for the use of clustered standard errors requires the number of clusters to go to infinity, so that standard errors can be consistently estimated. In a number of contexts, though, there are only few clusters that reliably provide independent information about the parameter of interest. It is then not possible to estimate the correct standard errors precisely, and the variability in the standard error estimator has to be taken into account when conducting inference.

In a time series context, the asymptotic framework of Kiefer and Vogelsang (2002, 2005) provides a model for the variability of such standard error estimators: Even asymptotically, the denominator of their t-statistics remains random. But its asymptotic distribution is known (at least up to a scaling constant that cancels in the overall fraction), so that an appropriate critical value can be computed. Similarly, in a panel context, Hansen (2007) and Donald and Lang (2007) derive asymptotically justified inference in which the variability of the standard error is explicitly taken into account. Closely related approaches are developed in Müller (2007, 2014), Stock and Watson (2008), Sun, Phillips and Jin (2008), Bester, Conley and Hansen (2011) and Sun (2013, 2014).

An important limitation of these approaches is that the asymptotic distribution of the standard error estimator needs to be fully known, at least up to a scaling constant. This requires strong homogeneity assumptions, ruling out clusters of different size or with substantially different design matrices, and, in a time series context, deterministic or stochastic trends in second moments.

In general, allowing for variance heterogeneity leads to test statistics whose distribution depends on the relative variances from each cluster. These nuisance parameters cannot be consistently estimated, given that the point of clustering standard errors is to remain agnostic about the form of intra-cluster correlations. With a finite number of clusters, also bootstrap or subsampling methods have no theoretical justification. In Monte Carlo experiments, Cameron, Gelbach and Miller (2008) found good performance of the percentile-t wild cluster bootstrap even with a small number of clusters, although these experiments focussed on relatively homogeneous designs. We consider explicitly heterogenous designs below and find that the method does not generally control size under cluster heterogeneity. Further analytical progress can be made by deriving bounds for the appropriate quantile of the test statistic that hold for any value of the cluster variances.[1]

Bakirov and Székely (2005) establish the following remarkable small sample result: The usual student-t critical values are valid for the t-test about the mean of q independent and Gaussian observations, even if the variances are heterogeneous, at least at conventional

[1] See Imbens and Kolesár (2012), Carter, Schnepel and Steigerwald (2013), Webb (2014), MacKinnon and Webb (2014) and Canay, Romano and Shaikh (2014) for some recent alternative suggestions for inference with a small number of clusters.

significance levels.[2] In a previous paper, Ibragimov and Müller (2010), we rely on this result to derive asymptotically valid inference about a scalar parameter of interest β as follows: Partition the data into $q \geq 2$ groups that provide approximately independent information about β. Estimate the model on each of the groups to obtain estimators $\hat{\beta}_j$, $j = 1, \ldots, q$ (the model may contain additional parameters beyond β, which are estimated along with $\hat{\beta}_j$ but then discarded). Then test the null hypothesis $H_0 : \beta = \beta_0$ with the usual t-test using the q observations $\{\hat{\beta}_j\}_{j=1}^q$ and $q - 1$ degrees of freedom.[3] Given the result of Bakirov and Székely (2005), this test is asymptotically valid as long as the $\hat{\beta}_j$'s are asymptotically independent, unbiased and Gaussian of possibly different variances. Even severe heterogeneity in the variability of the $\hat{\beta}_j$'s can thus be accommodated, enabling valid inference with very few and potentially heterogenous clusters. As discussed in more detail in Ibragimov and Müller (2010) natural group choices in a time series or spatial setting lead to asymptotic independence of the group estimators under conventional weak dependence assumptions, so that this approach may be applied in a wide range of settings.

This paper extends this approach in two dimensions. First, we establish a corresponding result for the comparison of a scalar parameter across two types of groups, such as treatment

[2] For a two-sided t-test, the result holds at the 8.3% level and below for all values of $q \geq 2$, and it also holds at the 10% level for $q \leq 14$.

[3] The idea of using group estimates for a t-test goes back to Brillinger (1973); it is also known as the "batch mean method" in the analysis of Markov Chain Monte Carlo output and as the "Fama and MacBeth (1973) method" in finance. The contribution of Ibragimov and Müller (2010) is to demonstrate its validity even with a small number of *heterogeneous* groups.

and control groups, or pre- and post-structural break data with known break date. The small sample problem here is the analysis of the usual two sample t-statistic when the underlying observations in the two samples are independent and Gaussian, but of potentially heterogeneous variance within and across the two samples. We prove that the critical value of a student-t distribution with degrees of freedom equal to the smaller sample size minus one lead to valid tests at conventional significance levels. This result then allows us to derive asymptotically valid inference about a scalar parameter $\beta = \delta_1 - \delta_2$, where δ_1 and δ_2 describe two different populations: Let $\{\hat{\delta}_{1,j}\}_{j=1}^{q_1}$ and $\{\hat{\delta}_{2,j}\}_{j=1}^{q_2}$ be the parameter estimates from the two types of groups with population values δ_1 and δ_2, respectively, where $q_1, q_2 \geq 2$. The null hypothesis $H_0 : \beta = \beta_0$ can then be tested with the usual two-sample t-test using the observations $\{\hat{\delta}_{1,j}\}_{j=1}^{q_1}$ and $\{\hat{\delta}_{2,j}\}_{j=1}^{q_2}$, and a critical value from a student-t distribution with $\min(q_1, q_2) - 1$ degrees of freedom.

Second, we develop a test for the appropriate level of clustering. A researcher entertains the null hypothesis that a fine level of clustering is appropriate, with the alternative that only a coarser level of clustering (few groups with corresponding estimators $\{\hat{\beta}_j\}_{j=1}^{q}$) actually provides approximately independent information about the parameter of interest. For example, in an analysis with a large panel of countries, a fine level of clustering might cluster on countries, while a coarser level only imposes independence across a few ($= q$) larger regions. We approximate the fine clustering by asymptotics where the number of clusters goes to infinity, so that under the null hypothesis, the asymptotic variance σ_j^2 of each of the $\hat{\beta}_j$ can be consistently estimated. In the example, $\hat{\beta}_j$ is the parameter estimator using data from region j only, and σ_j^2 is estimated using the usual clustered standard error in the estimation of $\hat{\beta}_j$, where the clustering is on countries. The suggested test then compares the sample variance computed from the q observations $\{\hat{\beta}_j\}_{j=1}^q$ with what one would expect if the $\hat{\beta}_j$ where normal of variance proportional to the estimated value of σ_j^2, as would be the case asymptotically under the null hypothesis. The test can also be applied in the context of comparisons between two populations, as described. Rejections of the test suggest that usual inference with clustered standard errors using the fine level of clustering is invalid, so that instead, the methods based on group estimators $\hat{\beta}_j$ should be applied.

The remainder of this paper is organized as follows. Section 2 provides evidence on the failure of Cameron, Gelbach, and Miller's (2008) percentile-t wild cluster bootstrap, as well as Bester, Conley, and Hansen's (2011) approach, to reliably control size under cluster heterogeneity. Section 3 discusses inference about comparisons across two populations in detail. Section 4 develops the test for the level of clustering, and provides some Monte Carlo evidence on its small sample properties. Section 5 illustrates the new tests in four empirical applications.

2 Validity of Inference with Few Heterogeneous Clusters

As an initial motivation, consider a linear regression

$$y_{j,i} = X_{j,i}'\theta + \varepsilon_{j,i} \tag{1}$$

where $y_{j,i}$ and $X_{j,i}$ are the ith of n_j observations from cluster j, $j = 1, \ldots q$, $X_{j,i}$ is a non-random $k \times 1$ regressor and $\varepsilon_{j,i}$ is mean zero normal and uncorrelated across clusters $E[\varepsilon_{j,i}\varepsilon_{l,k}] = 0$ for $j \neq l$, but not necessarily within clusters. Suppose we are interested in inference about the first element of θ, $\beta = \iota_1'\theta$ with $\iota_1 = (1, 0, \ldots, 0)'$. Specifically, we seek to test the null hypothesis $H_0 : \beta = \beta_0$ against the two-sided alternative $H_1 : \beta \neq \beta_0$.

The usual OLS estimator $\hat{\theta}^{OLS}$ can be written as

$$\hat{\theta}^{OLS} = \theta + \left(\sum_{j=1}^{q} \Gamma_j\right)^{-1} \sum_{j=1}^{q} Z_j \tag{2}$$

where $\Gamma_j = \sum_{i=1}^{n_j} X_{j,i}X_{j,i}' = X_j'X_j$, and $Z_j = \sum_{i=1}^{n_j} X_{j,i}\varepsilon_{j,i}$ are independent $\mathcal{N}(0, \Psi_j)$ with $\Psi_j = \mathrm{Var}[\sum_{i=1}^{n_j} X_{j,i}\varepsilon_{j,i}]$. The point of clustering is to remain agnostic about the value of $\{\Psi_j\}_{j=1}^{q}$ while conducting inference about β.

Let $\hat{e}_j = Z_j - \Gamma_j(\hat{\theta}^{OLS} - \theta)$. Then the usual clustered and degree of freedom corrected standard error of $\hat{\beta}^{OLS} = \iota_1'\hat{\theta}^{OLS}$ is $\hat{\sigma}_\beta$, where

$$\hat{\sigma}_\beta^2 = \frac{q}{q-1}\iota_1' \left(\sum_{j=1}^{q} \Gamma_j\right)^{-1} \left(\sum_{j=1}^{q} \hat{e}_j\hat{e}_j'\right) \left(\sum_{j=1}^{q} \Gamma_j\right)^{-1} \iota_1 \tag{3}$$

and the corresponding t-statistic is

$$t^{\text{cluster}} = \frac{\hat{\beta}^{OLS} - \beta_0}{\hat{\sigma}_\beta}. \tag{4}$$

Ibragimov and Müller's (2010) (IM in the following) suggestion is to estimate the parameter of interest from each cluster, and then apply a t-test to the q estimates. If Γ_j is invertible, the OLS estimator of θ from cluster j is $\hat{\theta}_j = \theta + \Gamma_j^{-1} Z_j$, so that the cluster j estimator of β is $\hat{\beta}_j = \beta + \iota_1' \Gamma_j^{-1} Z_j$. Thus, IM's suggestion is to reject H_0 when the absolute value of

$$t^{IM} = \sqrt{q} \frac{\bar{\hat{\beta}} - \beta_0}{S} \tag{5}$$

is larger than the usual critical value cv from a student-t with $q-1$ degrees of freedom, where $\bar{\hat{\beta}} = q^{-1} \sum_{j=1}^q \hat{\beta}_j$ and $S^2 = \frac{1}{q-1} \sum_{j=1}^q (\hat{\beta}_j - \bar{\hat{\beta}})^2$, yielding a confidence interval for β with endpoints $\bar{\hat{\beta}} \pm \text{cv}\, S/\sqrt{q}$. Since $\hat{\beta}_j \sim \mathcal{N}(\beta, \iota_1' \Gamma_j^{-1} \Psi_j \Gamma_j^{-1} \iota_1)$ independent across j, the result of Bakirov and Székely (2005) described in the introduction ensures that this inference remains valid for any value of the Ψ_j's at significance level 8.3% and below.

Cameron, Gelbach and Miller (2008) (CGM in the following) instead consider a wild bootstrap to approximate the null quantiles of t^{cluster}. In the bootstrap world, the Γ_j's are as in the actual sample, but the Z_j's are replaced by $U_j^* \hat{e}_j^R$, where the U_j are i.i.d. random variables with $P(U_j^* = 1) = P(U_j^* = -1) = 1/2$ and $\hat{e}_j^R$ are the estimates of Z_j under the null hypothesis, that is with R the last $k-1$ columns of I_k, $\hat{e}_j^R = Z_j - \Gamma_j R (\sum_{i=1}^q R' \Gamma_i R)^{-1} \sum_{i=1}^q R' Z_i$. Note that this bootstrap distribution consists of (at most) 2^q distinct points. CGM find in Monte Carlo simulations that under homogeneous clusters ($\Gamma_i \approx \Gamma_j$ and $\Psi_i \approx \Psi_j$ for all i, j), this procedure works well even for fairly small q.

Alternatively, Bester, Conley and Hansen (2011) (BCH in the following) suggest relying on t^{cluster} with a critical value from a student-t with $q-1$ degrees of freedom. Under homogeneity of $\Gamma_j = X_j'X_j$ across clusters ($\Gamma_i = \Gamma_j$), this results in valid inference, because t^{cluster} then reduces to IM's statistic t^{IM} via $\bar{\hat{\beta}} = \hat{\beta}^{OLS}$.

Little is known about the validity of CGM's and BCH's method under general cluster heterogeneity for finite q (validity under $q \to \infty$ follows from standard arguments). Both methods implicitly define a critical region CR, the subset of values of $\{Z_j\}_{j=1}^q$ for which the null hypothesis $H_0 : \beta = \beta_0$ is rejected. The critical region depends on the observed matrices $\{\Gamma_j\}_{j=1}^q$, $\text{CR} = \text{CR}_{\{\Gamma_j\}_{j=1}^q}$. In this notation, the null rejection probability simply becomes $P(\{Z_j\}_{j=1}^q \in \text{CR}_{\{\Gamma_j\}_{j=1}^q})$, which is a function of $\{\Psi_j\}_{j=1}^q$ via $Z_j \sim \mathcal{N}(0, \Psi_j)$. As noted before, the point of clustering is to remain agnostic about the value of $\{\Psi_j\}_{j=1}^q$. So for a given value of $\{\Gamma_j\}_{j=1}^q$, the size of these methods is usefully defined as

$$\sup_{\{\Psi_j\}_{j=1}^q} P(\{Z_j\}_{j=1}^q \in \text{CR}_{\{\Gamma_j\}_{j=1}^q}), \tag{6}$$

the largest rejection probability that can be induced by varying $\{\Psi_j\}_{j=1}^q$. It is computationally difficult to determine this quantity, as the space of q covariance matrices of dimension $k \times k$ is large unless both k and q are very small. To get some sense for the reliability of the CGM and BCH methods we compute their rejection probability for a relatively small set of values of $\{\Psi_j\}_{j=1}^q$ at the edge of the parameter space, as detailed in the online appendix. The largest of these null rejection probabilities is, by construction, a lower bound on actual size (6).

<< Table 1 about here >>

Since size depends on $\{\Gamma_j\}_{j=1}^q$, we computed this lower bound for 100 independent draws of $\{\Gamma_j\}_{j=1}^q$, where Γ_j are distributed i.i.d. Wishart with $2k$ degrees of freedom and scale matrix I_k. Table 1 reports summary statistics of these 100 draws for various values of k and q. One can see that both methods are very seriously oversized, at least for some values of $\{\Gamma_j\}_{j=1}^q$. The one exception is CGM's method for $k = 1$ and $q > 4$, for which we found no evidence of size distortions. For $k = 1$ and $q = 4$ CGM's method seems to result in an empty critical region, i.e. it never rejects. With $q = 4$ the bootstrap distribution has only $2^q = 16$ points of support, and for $k = 1$, the realized value of test statistic t^{cluster} apparently always falls between 2.5% and 97.5% quantiles of this distribution.

For computational reasons, we only considered the values 1, 2 and 3 for the number of regressors k. Note, however, that k can be thought of as the number of non-cluster specific regressors. This follows from standard Frisch-Waugh logic: Let $W_{j,i}$ be regressors that are specific to one group, that is each element of $W_{j,i}$ is nonzero only for one cluster j, let $\tilde{X}_{j,i}$ be the $k \times 1$ non-cluster specific original regressors, and let $\tilde{\varepsilon}_{j,i}$ be the original disturbances. Now define $X_{j,i}$ and $\varepsilon_{j,i}$ as the residuals of a linear regression of $\tilde{X}_{j,i}$ and $\tilde{\varepsilon}_{j,i}$ on $W_{j,i}$, respectively. Then $\varepsilon_{j,i}$ are still uncorrelated across clusters, equations (2)-(4) still hold, and both CGM's and BCH's method behave as described in Table 1. For instance, if a regression analysis contains cluster fixed-effects and a single non-cluster specific regressor of interest, then the $k = 1$ results of Table 1 apply.

One might argue that this linear regression design with normal errors and fixed regressors is fairly special. But consider asymptotics where the number of observations in each cluster n_j is some positive fraction of n, and $n \to \infty$. A law of large numbers and a central limit theorem applied to cluster averages then yields $n^{-1}\Gamma_j = n^{-1}\sum_{i=1}^{n_j} X_{j,i}X'_{j,i} \xrightarrow{p} G_j$ and $n^{-1/2}Z_j = \sum_{i=1}^{n_j} X_{j,i}\varepsilon_{j,i} \Rightarrow \mathcal{N}(0, \Psi_j)$ independent across j (see IM for additional details and a generalization to GMM models). The distributional assumption of treating the Γ_j as fixed and Z_j as independent mean-zero normals then arises naturally. The numbers in Table 1 are therefore also lower bounds on the the asymptotic size of the CGM and BCH method under such asymptotics, and IM's method controls asymptotic size no matter the value of $\{\Psi_j\}_{j=1}^q$. Consequently, the results here point to IM's method as a generally more reliable procedure to conduct inference with few heterogeneous clusters.

Note, however, that in order to implement IM's t-statistic (5), it must be possible to estimate the parameter β from each cluster. This rules out parameters of interest β that are only identified from across cluster variation, rendering the Γ_j non-invertible. A particularly important example is inference about the difference of a linear regression coefficient between two populations, with the first q_1 clusters from one population, and $q_2 = q - q_1 > 0$ independent clusters from the second population. With a scalar regressor $x_{j,i}$, this corresponds in the above notation to inference about the first element of θ in (1) with $X_{j,i} = (x_{j,i}, x_{j,i})'$ for $j \leq q_1$ and $X_{j,i} = (0, x_{j,i})'$ for $j = q_1 + 1, \ldots, q_1 + q_2$, leading to 2×2 matrices $\Gamma_j = X'_j X_j$

of the form

$$\Gamma_j = \begin{pmatrix} \gamma_j & \gamma_j \\ \gamma_j & \gamma_j \end{pmatrix} \text{ for } j \leq q_1 \text{ and } \Gamma_j = \begin{pmatrix} 0 & 0 \\ 0 & \gamma_j \end{pmatrix} \text{ for } j = q_1 + 1, \ldots, q_1 + q_2 \tag{7}$$

with $\gamma_j = \sum_{i=1}^{n_j} x_{j,i}^2 > 0$, and $Z_j = \sum_{i=1}^{n_j} X_{j,i}\varepsilon_{j,i} \sim \mathcal{N}(0, \Psi_j)$ with

$$\Psi_j = \begin{pmatrix} \psi_j & \psi_j \\ \psi_j & \psi_j \end{pmatrix} \text{ for } j \leq q_1 \text{ and } \Gamma_j = \begin{pmatrix} 0 & 0 \\ 0 & \psi_j \end{pmatrix} \text{ for } j = q_1 + 1, \ldots, q_1 + q_2$$

for some $\psi_j \geq 0$. As before, these expressions also remain valid in the presence of additional cluster specific regressors $W_{j,i}$ once $x_{j,i}$ and $\varepsilon_{j,i}$ are defined as residuals of a linear regression of the original scalar regressor of interest $\tilde{x}_{j,i}$ and the original disturbance $\tilde{\varepsilon}_{j,i}$ on the cluster specific regressors.

<< Table 2 about here>>

Table 2 reports summary statistics of lower bounds on size (6) of the CGM and BCH methods in this "Two-Sample" design for various values of q_1 and q_2. As in Table 1, for each pair of (q_1, q_2), we generated 100 draws of $\{\gamma_j\}_{j=1}^q$ with γ_j i.i.d. Chi-squared with 2 degrees of freedom. For each such realization of $\{\gamma_j\}_{j=1}^q$, we compute the largest null rejection probability over a finite set of values of $\{\psi_j\}_{j=1}^q$ detailed in the online appendix. As can be seen from the table, neither of the two methods yields reliable inference. This motivates the development of a version of IM's method that guarantees valid inference in the two-sample design, which we pursue in the next section.

3 Comparisons between Two Populations

3.1 Small Sample Result

Let $Y_{i,j}$ be independent random variables with distribution $Y_{i,j} \sim \mathcal{N}(\mu_i, \sigma^2_{i,j})$, $j = 1, \ldots, q_i$, $i = 1, 2$, where $q_i \geq 2$. Define the statistics $\bar{Y}_i = q_i^{-1} \sum_{i=1}^{q_i} Y_{i,j}$ and $s_i^2 = (q_i - 1)^{-1} \sum_{i=1}^{q_i} (Y_{i,j} - \bar{Y}_i)^2$ for $i = 1, 2$. The parameter of interest is $\Delta = \mu_1 - \mu_2$, so we seek to test $H_0 : \Delta = \Delta_0$ against $H_1 : \Delta \neq \Delta_0$. The usual two sample t-statistic is given by

$$t = \frac{\bar{Y}_1 - \bar{Y}_2 - \Delta_0}{\sqrt{\frac{s_1^2}{q_1} + \frac{s_2^2}{q_2}}} \tag{8}$$

and the null hypothesis is rejected for large values of $|t|$.[4] In the case of homogenous samples with $\sigma_{i,j} = \sigma_i > 0$, the null distribution of t only depends on the nuisance parameter σ_1/σ_2, and Mickey and Brown (1966) show that the quantiles of t are bounded by the appropriate quantiles from a t-distribution with $\min(q_1, q_2) - 1$ degrees of freedom. This bound is sharp, since it is obtained as either $\sigma_1/\sigma_2 \to 0$ or $\sigma_1/\sigma_2 \to \infty$.

The following theorem provides a corresponding result under heterogeneity within the individual samples, where the nuisance parameter space involves in addition the $q_1 + q_2 - 2$ ratios $\sigma_{i,j}/\sigma_{i,1}$, $j = 2, \ldots, q_i$, $i = 1, 2$.

Theorem 1 *Let* $\mathrm{cv}(\alpha, m)$ *be the* $1 - \alpha/2$ *quantile of the student-t distribution with* m *degrees of freedom. Under the null hypothesis of* $\Delta = \Delta_0$,

$$\sup_{\{\sigma_{1,j}\}_{j=1}^{q_1}, \{\sigma_{2,j}\}_{j=1}^{q_2}} P(|t| > \mathrm{cv}(\alpha, \min(q_1, q_2) - 1)) = \alpha$$

[4]We define t to be zero if $s_1^2 = s_2^2 = \bar{Y}_1 - \bar{Y}_2 = 0$, a zero probability event if $\max_{i,j} \sigma_{i,j} > 0$.

for $2 \leq q_1, q_2 \leq 50$ *and* $\alpha \in \{0.001, 0.002, \ldots, 0.083\}$, *and also for* $\alpha \in \{0.083, 0.084, \ldots, 0.10\}$ *if* $2 \leq q_1, q_2 \leq 14$.

Theorem 1 is a new probabilistic result of potentially independent interest in the literature of small sample properties of t-statistics and quadratic forms in symmetric and normal variates (cf. Efron (1969), Benjamini (1983), Dufour (1991), Dufour and Hallin (1993), Bakirov (1989a, 1989b, 1995) and Bakirov and Székely (2005)). The most closely related work is Bakirov (1998), who studies the behavior of the two-sample t-statistic with the "pooled" variance estimator in the denominator under variance heterogeneity. Bakirov (1998) shows that the student-t critical value with $\min(q_1, q_2) - 1$ degrees of freedom yields a valid test when $\min(q_1, q_2) \geq 7$ and for very low levels of α (much smaller than 1% for most values of (q_1, q_2), and always less than 1%). The proof of Theorem 1 is involved. It relies in part on the approach of Bakirov (1998), the insights of Bakirov and Székely (2005), and a number of additional arguments. See the online appendix for details.

One step of the proof requires comparisons of a (large but finite) set of quantities that depend on α, q_1 and q_2. We performed these comparisons for the values indicated in the theorem, but we would expect the result to go through also for additional values of α and $q_1, q_2 > 50$. Under $\min(q_1, q_2) \to \infty$ and $\max_{i,j,k,l} \sigma_{i,j}/\sigma_{k,l} < \infty$ the validity of the t-test follows, of course, from standard asymptotic arguments.

In small samples, the t-test (8) can be quite conservative, that is its null rejection probability can be substantially below the nominal level α for some values of $\sigma_{i,j}^2$. This raises a concern about power. A natural comparison is a test based on the numerator $\bar{Y}_1 - \bar{Y}_2 - \Delta_0$

in (8) with known variances, that is a test that rejects for large values of $|z|$ with

$$z = \frac{\bar{Y}_1 - \bar{Y}_2 - \Delta_0}{\sqrt{q_1^{-2}\sum_{j=1}^{q_1}\sigma_{1,j}^2 + q_2^{-2}\sum_{j=1}^{q_2}\sigma_{2,j}^2}} \tag{9}$$

and the usual normal critical values. Figures 1 and 2 plot the rejection probabilities for some choices of q_1, q_2 and $\sigma_{i,j}^2$ of nominal 5% level tests. Note that in some scenarios, the null rejection probability of the t-test is very small; for instance, in the upper right plot of Figure 2, the null rejection probability is only 0.56% for $q_1 = 4$ and $q_2 = 16$. Remarkably, this severe underrejection does not lead to a large loss in power: under alternatives where the z test has roughly 50% power, the rejection probability of the the t-test seems almost completely determined by $\min(q_1, q_2)$, irrespective of any variance heterogeneity. As such, very substantial power losses compared to the z-test under such moderate alternatives only arise when $\min(q_1, q_2) = 4$ (where the two-sided critical value is 3.18 for the t-test, compared to 1.96 for the z-test). IM reported very similar findings for the one-sample t-statistic in their Figure 3.

<< Figures 1 and 2 about here >>

3.2 Large Sample Inference with a Finite Number of Groups

Our interest in Theorem 1 mainly stems from its application to valid large sample inference as follows. Suppose δ_i, $i = 1, 2$ are parameters of some econometric model, and we are interested in inference about $\beta = \delta_1 - \delta_2$, that is we want to test the null hypothesis $H_0 : \beta = \beta_0$. The model might be linear or nonlinear, and might involve additional parameters beyond

δ_i. Suppose the total n observations are partitioned into $q_1 + q_2$ groups, such that q_1 groups provide at least asymptotically independent information about δ_1, and the remaining q_2 groups provide asymptotically independent information about δ_2. Estimate the model q_1+q_2 times, using observations of each group only, and let $\hat{\delta}_{i,j}$, $j = 1, \ldots, q_i$ be the resulting estimators of δ_i, $i = 1, 2$. Under asymptotics in which the number $q_1 + q_2$ of groups is fixed, and each group contains more and more observations, standard results on the large sample behavior of a wide class of estimators $\hat{\delta}_{i,j}$ imply

$$\sqrt{n}(\hat{\delta}_{i,j} - \delta_i) \Rightarrow \mathcal{N}(0, \sigma_{i,j}^2),\ j = 1, \ldots, q_i,\ i = 1, 2. \tag{10}$$

What is more, by assumption about the choice of groups, $\{\hat{\delta}_{i,j}\}$ are asymptotically independent. As discussed in IM, it is not necessary that the group data itself is independent across groups for this to hold: standard weak dependence assumption in time or space induce asymptotic independence under reasonable group choices, as most of the variability of $\hat{\delta}_{i,j}$ stems from observations that are far from the "group borders".

Now define $\bar{\hat{\delta}}_i = q_i^{-1} \sum_{j=1}^{q_i} \hat{\delta}_{i,j}$ and $S_i^2 = (q_i - 1)^{-1} \sum_{j=1}^{q_i} (\hat{\delta}_{i,j} - \bar{\hat{\delta}}_i)^2$ for $i = 1, 2$, and let

$$t^{IM2} = \frac{\bar{\hat{\delta}}_1 - \bar{\hat{\delta}}_2 - \beta_0}{\sqrt{\frac{S_1^2}{q_1} + \frac{S_2^2}{q_2}}}, \tag{11}$$

the usual two sample t-statistic for the difference in means based on the two samples $\{\hat{\delta}_{1,j}\}_{j=1}^{q_1}$ and $\{\hat{\delta}_{2,j}\}_{j=1}^{q_2}$. As long as at least one of the asymptotic variances $\sigma_{i,j}^2$ is positive, $\max_{i,j} \sigma_{i,j}^2 > 0$, the continuous mapping theorem and (10) imply that

$$t^{IM2} \Rightarrow \frac{\bar{Y}_1 - \bar{Y}_2 - \Delta_0}{\sqrt{\frac{s_1^2}{q_1} + \frac{s_2^2}{q_2}}} \tag{12}$$

under the null hypothesis, where the right-hand side of (12) is as in Section 3.1 with $\Delta = \Delta_0$. Thus, Theorem 1 implies that rejecting for values of $|t^{IM2}|$ that are larger than the corresponding critical value cv of a student-t distribution with $\min(q_1, q_2) - 1$ degrees of freedom (df) results in asymptotically valid inference. Equivalently, an asymptotically valid confidence interval for β has endpoints $\bar{\hat{\delta}}_1 - \bar{\hat{\delta}}_2 \pm \text{cv}\sqrt{S_1^2/q_1 + S_2^2/q_2}$. Moreover, (12) also holds under local alternatives where $\sqrt{n}(\beta - \beta_0) \to \Delta - \Delta_0$, so that the local asymptotic power of such inference is equal to the small sample power of the two-sample t-statistic (8). As is easily seen, for more distant alternatives where $\sqrt{n}|\beta - \beta_0| \to \infty$, the test based on t^{IM2} is consistent.

Returning to the linear regression set-up with design matrices (7) of Section 2, let $\theta = (\delta_1 - \delta_2, \delta_2)'$, so that δ_1 and δ_2 are the coefficients in the two populations, and $\beta = \delta_1 - \delta_2$. Let $\hat{\delta}_{1,j} = \delta_1 + \gamma_j^{-1} Z_j$ be the estimated coefficient of a regression of $y_{j,i}$ on $x_{j,i}$ using group $j = 1, \ldots, q_1$ data only, and define $\hat{\delta}_{2,j}$ correspondingly as $\hat{\delta}_{2,j} = \delta_2 + \gamma_{q_1+1}^{-1} Z_{q_1+j}$, $j = 1, \ldots, q_2$. Then Theorem 1 implies that the test based on (11) is small sample valid under Gaussian errors $\varepsilon_{j,i}$. What is more, in the important special case where γ_j is constant across j, the power of t^{IM2} compares to the power of the (infeasible) test based on the estimator $\hat{\beta}^{OLS}$ with *known* variance just like the t-test and z-test in Figures 1 and 2 above. (When γ_j is heterogeneous, then $\hat{\beta}^{OLS}$ no longer equals $\bar{\hat{\delta}}_1 - \bar{\hat{\delta}}_2$, and relative power can go either way depending on the relationship between the heterogeneity in γ_j, and the heterogeneity in the variances. See IM for further discussion.)

It doesn't pose any problems if the model contains additional parameters beyond δ_i as long as $\hat{\delta}_{i,j}$ can be estimated from each cluster. In addition, note that t^{IM2} is invariant to transformations of the type $\hat{\delta}_{i,j} \to \hat{\delta}_{i,j} + m$ for any $m \in \mathbb{R}$, since m cancels in the numerator in the difference $\bar{\hat{\delta}}_1 - \bar{\hat{\delta}}_2$, and also in the expression for S_1^2 and S_2^2. Thus, the basic assumption (10) for the validity of inference based on t^{IM2} can be weakened to

$$\sqrt{n}(\hat{\delta}_{i,j} - m_n - \delta_i) \Rightarrow \mathcal{N}(0, \sigma_{i,j}^2),\ j = 1, \ldots, q_i,\ i = 1, 2 \tag{13}$$

for an unknown sequence m_n that is not required to converge. For instance, consider an intervention that has a time dimension $t = 1, \ldots, T$, so that in a linear model with time fixed effects α_t, the outcome $y_{i,j,t,l}$ in cluster $j = 1, \ldots, q_i$ of population $i = 1, 2$ for an individual $l = 1, \ldots, n_{i,j}$ with characteristics $x_{i,j,t,l}$ is

$$y_{i,j,t,l} = \delta_i + x'_{i,j,t,l}\psi + \alpha_t + u_{i,j,t,l}$$

for some conditionally mean zero error term $u_{i,j,t,l}$. Let $\hat{f}_{i,j,t}$ be the OLS estimators of the time fixed effects in a regression of $y_{i,j,t,l}$ on $x_{i,j,t,l}$ using data of cluster j from population i only (excluding an additional constant). Then $\hat{\delta}_{i,j} = T^{-1} \sum_{t=1}^{T} \hat{f}_{i,j,t}$ estimates $\delta_i + T^{-1} \sum_{t=1}^{T} \alpha_t$, and (13) holds with $m_T = m_n = T^{-1} \sum_{t=1}^{T} \alpha_t$ under sufficiently weak dependence of $u_{i,j,t,l}$ within cluster. This is true even under $T \to \infty$ asymptotics, where there is no reason to expect m_T to converge to anything. This approach can be generalized to two-way fixed effects in a diff-in-diff application, see Section 5.4 below.

For the asymptotic validity of tests based on t^{IM2}, the rate of convergence $\sqrt{n}$ in (10) (or (13)) is immaterial—any rate $a_n \to \infty$ would work, as it cancels in (11). The same approach to inference is thus also applicable in some non-regular and semi-parametric settings, as long as estimators are asymptotically unbiased and Gaussian. Furthermore, one can replace (10) by an assumption that $a_n(\hat{\delta}_{i,j} - \delta_i) \Rightarrow \sqrt{R_{i,j}} Z_{i,j}$, where $Z_{i,j} \sim iid\mathcal{N}(0,1)$, and $R_{i,j}$ are (possibly correlated) non-negative random variables that are independent of $\{Z_{i,j}\}$, as long as $\sup_{i,j} R_{i,j} > 0$ almost surely. The validity of inference based on t^{IM2} then still follows from Theorem 1 after conditioning on $\{R_{i,j}\}$. This structure allows for the presence of stochastic volatility affecting $\hat{\delta}_{i,j}$, as well as convergence of $\hat{\delta}_{i,j}$ to any distribution that can be written as a scale mixture of normals. This is a rather large class of symmetric distributions, containing all student-t distributions, the logistic distribution, the double exponential distribution and all symmetric stable distributions. Thus, after a suitable partition of a time series, the statistic t^{IM2} can also be used, say, for Chow (1960)-type tests about the change of location of a serially correlated heavy-tailed time series in the domain of attraction of a symmetric stable law, or for a Chow-test of other parameters whose estimators are known to converge to a symmetric stable law (such as, for instance, the sample autocovariances in GARCH processes and estimates of an autoregressive parameter in an AR(1) process with GARCH errors under empirically plausible assumptions, see Davis and Mikosch (1998), Mikosch and Stărică (2000), Borkovec (2001) and Cont (2001)). And, given the practical difficulty of estimating the tail index, it seems that very few alternatives modes of inference are available for such problems.

4 Testing the Level of Clustering

In applied work it can be challenging to decide on the appropriate level of clustering: fine clustering (many clusters) may rule out plausible correlations, but a coarse level of clustering (few clusters) calls into question standard inference that is based on "consistent" clustered standard errors. In this section, we develop a test φ_f of the null hypothesis that a fine level of clustering is appropriate, against the alternative that only fewer groups provide independent information about the parameter of interest.

The setting is similar to what is described in Ibragimov and Müller (2010) and Section 3.2 above. An econometric model involves the scalar parameter of interest β, possibly along with additional parameters. There exists a partitioning of the n total observations into q groups that provide asymptotically independent information about β even under the alternative. The number of groups q is fixed as a function of the overall sample size n. Estimation of the model on the data of each of the q groups yields the estimators $\hat{\beta}_j$, $j = 1, \ldots, q$. These estimators satisfy

$$\sqrt{n}(\hat{\beta}_j - \beta) \Rightarrow \mathcal{N}(0, \sigma_j^2) \tag{14}$$

and are asymptotically independent, both under the null and alternative hypothesis about the appropriate level of clustering.

Under the null hypothesis, also a fine level of clustering is justified. Consider asymptotics in which each of the q groups eventually contain an infinite number of (asymptotically) independent clusters. The usual clustered standard errors $\hat{\omega}_j$ computed for each of the q

estimations can then be employed to accurately estimate $\hat{\sigma}_j = \sqrt{n}\hat{\omega}_j \xrightarrow{p} \sigma_j$, so that σ_j^2 in (14) is effectively known under the null hypothesis.

Our suggestion for φ_f can now be thought of as a Hausman (1978)-type test about the (asymptotic) variance of $\sqrt{n}\overline{\hat{\beta}} = \sqrt{n}q^{-1}\sum_{j=1}^{q}\hat{\beta}_j$. Under the null hypothesis, this variance can be accurately estimated by $q^{-2}\sum_{j=1}^{q}\hat{\sigma}_j^2$. Under the alternative, a natural estimator is given by

$$V = nS^2/q, \quad S^2 = (q-1)^{-1}\sum_{j=1}^{q}(\hat{\beta}_j - \overline{\hat{\beta}})^2,$$

the (rescaled) sample variance of $\{\hat{\beta}_j\}_{j=1}^{q}$. In contrast to the usual Hausman (1978) set-up, these two estimators have different rates of convergence, though, since V has a nondegenerate (and non-Gaussian) limiting distribution, while $q^{-2}\sum_{j=1}^{q}\hat{\sigma}_j^2 \xrightarrow{p} q^{-2}\sum_{j=1}^{q}\sigma_j^2$ under the null hypothesis. The distribution theory for the comparison of the two estimators is thus dominated by the variability of V.

Under the null hypothesis, the distribution of V is very well approximated by the distribution of $V_Y = nS_Y^2/q$ with $S_Y^2 = (q-1)^{-1}\sum_{j=1}^{q}(Y_j - \bar{Y})^2$ and $\bar{Y} = q^{-1}\sum_{j=1}^{q}Y_j$, where the independent random variables Y_j have distribution $\mathcal{N}(0, \hat{\omega}_j^2)$ (conditional on the standard error estimate $\hat{\omega}_j$). Let $\mathrm{cv}_V(\alpha)$ be the $1-\alpha$ quantile of V_Y, which can easily be computed by simulation or other techniques. The test φ_f then rejects if and only if V is larger than $\mathrm{cv}_V(\alpha)$. It is easily seen that φ_f is of asymptotic level α.[5] Note that the rate $\sqrt{n}$ in (14) and in the relation $\hat{\sigma}_j = \sqrt{n}\hat{\omega}_j$ is immaterial—φ_f can be implemented by simply comparing

[5]This follows since under the null hypothesis, V and $\tilde{V}_Y$ have identical limiting distribution $(q-1)^{-1}\sum_{j=1}^{q}(Y_j - \bar{Y})^2$, where the Y_j are independent and distributed $Y_j \sim \mathcal{N}(0, \sigma_j^2)$.

$qV/n = S^2$ with the appropriate quantile of S_Y^2, neither of which involve $\hat{\sigma}_j$, or any scaling by n (see the Synopsis in Section 5).

Under the alternative, the fine level of clustering ignores correlations among the observations in the groups, and $\hat{\omega}_j$ is no longer an accurate estimator of the standard error of $\hat{\beta}_j$. In particular, inference about β based on the usual clustered standard error formula will overstate the significance if positive correlations within the q groups are ignored. In this case, V takes on larger values than one would expect if indeed $\hat{\beta}_j \sim \mathcal{N}(\beta, \hat{\omega}_j^2)$, leading to a rejection of φ_f. Formally, the asymptotic distribution of V stochastically dominates the distribution of V_Y whenever $\hat{\sigma}_j \to \underline{\sigma}_j \leq \sigma_j$ with some strict inequalities, inducing an asymptotic rejection probability of φ_f larger than α.

<< Table 3 about here >>

Table 3 reports some small sample rejection probabilities of φ_f in a simple panel setting. The null rejection probabilities are fairly close to the nominal level, even when the number of independent entities within each group is as small as 5 (where the standard error estimates $\hat{\omega}_j$ are quite imprecise).

A rejection of φ_f indicates that there are correlations across the fine clusters (but within the coarse clusters) that increase the variability of $\bar{\hat{\beta}}$ relative to what is accounted for by the fine clustering. In the presence of such correlations, valid inference is obtained by relying on IM's one-sample statistic t^{IM} in (5) and critical values from a student-t distribution with $q-1$ degrees of freedom, at least asymptotically. As is common for diagnostic tests, however, a systematic determination of the mode of inference as a function of φ_f will in general induce pre-test biases due to type-I and type-II errors. If the appropriate level of clustering is in doubt, then it makes sense to report the significance of results based on various clustering assumptions, and to interpret the resulting inference conditional on the validity of these assumptions. In this perspective the test φ_f merely provides empirical evidence on the plausibility of one particular clustering assumption.

Having said that, in the Monte Carlo simulation of Table 3, a t-test for the population mean based on OLS standard errors of 5% nominal level has null rejection probability of 18.7%-26.8% when the time series correlation is ignored (what is called "Alternative Hypothesis" in Table 3). A switch to t^{IM}-inference as a function of the outcome of the 5% level test φ_f reduces these size distortions to 5.9%-15.3% (compared to 3.6%-7.9% of pure t^{IM} based inference). So while not perfect, a systematic use of φ_f as a pre-test does substantially reduce size distortions, at least in this simple set-up.

The test φ_f has a natural counterpart in the two-sample problem, with the variance of $\bar{\hat{\delta}}_1 - \bar{\hat{\delta}}_2$ then playing the role of the variance of $\bar{\hat{\beta}}$. In the implementation, the statistics S^2 and S_Y^2 are to be replaced by $U = S_1^2/q_1 + S_2^2/q_2$ and $U_Y = S_{Y,1}^2/q_1 + S_{Y,2}^2/q_2$, respectively, where $S_{Y,i}^2 = (q_i - 1)^{-1} \sum_{j=1}^{q_i} (Y_{i,j} - \bar{Y}_i)^2$, $\bar{Y}_i = q_i^{-1} \sum_{j=1}^{q_i} Y_i$ and $Y_{i,j} \sim \mathcal{N}(0, \hat{\omega}_{i,j}^2)$ conditional on $\{\hat{\omega}_{i,j}\}$, where $\hat{\omega}_{i,j}$ is the clustered standard error of the estimator $\hat{\delta}_{i,j}$, $j = 1, \ldots, q_i$, $i = 1, 2$ that assume that fine clustering is justified.

5 Illustrations

<< Table 4 about here>>

We now illustrate the cluster test and t-statistic based inference in four empirical applications. All reported t-tests are two-sided. The implementation of the various tests suggested here are summarized in Table 4.

5.1 Few Independent Clusters: Dal Bó and Fréchette (2011)

Dal Bó and Fréchette (2011) experimentally study the degree of cooperation in infinitely repeated games as a function of the probability of continuation p (δ in the notation of Dal Bó and Fréchette (2011)), and the pay-off of cooperation R. They consider two values of $p \in \{\frac{1}{2}, \frac{3}{4}\}$ and three values of the cooperation pay-off $R \in \{32, 40, 48\}$, leading to a total of 6 treatments. For each treatment, they conduct three sessions, where each session involves between 12-20 individuals that are randomly re-matched for a total of 50 minutes of play. The bottom right panel of their Table 3 provides the results of significance tests of equal propensity to cooperate in 7 pairs of treatment, using all games and all rounds of play (reproduced in Panel B of Table 5 of this paper). The comparisons are conducted by running a probit regression on a constant and a dummy for the treatment under consideration, with standard errors clustered at the session level. Since there are only 3 sessions per treatment, there is very substantial variability in these standard error estimates. This variability, however, is not appropriately taken into account in the assessment of significance using the default clustering approach.

<< Table 5 about here>>

An alternative mode of inference is to estimate the propensity to cooperate session by session. Under the assumption that there is enough independence within sessions for a central limit theorem to hold, the resulting 18 estimators are independent and normal, and each triple of sessions corresponding to the same treatment have the same mean. Given the heterogeneity in the number of individuals and games played across sessions, one would not want to assume that these estimators have the same variance. But given Theorem 1, valid pairwise comparisons between treatments may still be conducted by simply employing the two-sample t-statistic (11) with the 3 probit coefficients as observations from each treatment, using the critical value from a student-t distribution with 2 degrees of freedom. Table 5 reports the results. Compared to the original analysis in Dal Bó and Fréchette (2011), the significance of differences between treatments is lower. But even though the 10% and 5% two-sided critical values of a student-t statistic with 2 degrees of freedom are 1.92 and 4.30, respectively, four of the 7 tests are still significant at the 10% level, and one at the 5% level. The approach thus still yields at least somewhat informative inference.

One might argue that given the small number of sessions, it would be more appropriate to cluster at the level of individuals. But when we test the validity of clustering at the level of the individual, against the alternative of coarser clustering at the session level using the test φ_f described in Section 4 above, we reject at the 5% level for six out of the seven comparisons. (This might not be too surprising: individuals play against each other in each session, after all, which might well lead to non-trivial interaction effects). Thus, Dal Bó and Fréchette (2011) were right to be concerned about intra-session correlation, and inference based on the t-statistic (11) adequately accounts for the (substantive!) additional variability of the resulting inference.

5.2 Time Series Correlations: Keim (1983)

In a classic paper, Keim (1983) provides evidence that the size anomaly of stock returns is, to a substantial degree, due to very high excess returns in January. In his Table 2, he reports average differences between daily CRSP excess returns of portfolios constructed from firms in the top and bottom decile of equity market value, for each January of the 17 years 1963 to 1979, along with the OLS standard error estimate. The overall January average over these years is reported to equal 0.714 percent, with a t-statistic of 11.8.

The standard errors are not adjusted for potential serial correlation. Treating the average from each January as potentially heteroskedastic independent normal variates with common mean, one can apply the Ibragimov and Müller (2010)-method (that is, a t-test with 16 degrees of freedom using the 17 January estimators) to obtain valid inference that accounts for arbitrary serial correlation within each January.[6] Such an analysis still leads to a significant January size effect at the 0.1% level, confirming the result in Keim (1983). At

[6] The fact that the Ibragimov and Müller (2010)-method accommodates heterogenous variances is quite attractive here, given that stock returns display time dependent (stochastic) volatility.

the same time, using the 17 pairs of estimator and OLS standard error as inputs to the level of clustering test φ_f leads to a rejection at the 0.1% level, indicating that Keim's original standard errors are too small.

Theorem 1 of this paper also allows for the straightforward implementation of a Chow (1960)-type test of the stability of the January size effect: Consider the null hypothesis that the January size effect is time invariant, against the alternative that it is different in the 60s and the 70s. Again allowing for arbitrary serial correlation within each January, this can be tested by computing a two-sample t-test, with the 7 January averages from the years 1963 to 1969 in one group, and the 10 January averages from the year 1970 to 1979 in the other. The resulting test rejects at the 10% level, providing weak evidence against time invariance.

5.3 Spatial Correlations: Obstfeld, Shambaugh and Taylor (2010)

Maurice Obstfeld (2010) study the determinants of central bank reserve holdings with a cross country regression involving an unbalanced panel of 26 years and 134 countries, for a total of 2671 observations. Their theoretical framework motivates a focus on four variables that relate to financial stability of a country: an index for "financial openness", dummies for a "Peg" or "Soft Peg", and the log of the ratio of M2 to GDP "ln(M2/GDP)". In regression (5) of their Table 1, they assess the significance of these four variables in a horserace against other factors, clustering standard errors by country to account for arbitrary serial correlation.

We reproduce these results in Panel B of Table 6 for convenience.

<< Table 6 about here >>

Give the close economic, political and historical ties between neighboring countries, one might worry about the presence of additional spatial correlation. To formally test this, we categorize the 134 countries into 6 regions: Western Europe/North America, Eastern Europe, Asia/Pacific, Middle East, South America and Africa, with 15 to 39 countries in each region. We re-estimated the horserace regression separately in each region, and used the 6 estimators and standard errors (clustered at the country level) to test for the validity of clustering at the country level. As reported in Table 5, the test φ_f of Section 4 is significant for two of the four coefficients of interest, indicating significant evidence of spatial correlation.[7] A test of significance of the coefficients based on the 6 estimates from each region using the Ibragimov and Müller (2010) method shows no evidence of the importance of Soft Peg, and only weak evidence for the importance of ln(M2/GDP), in contrast to the analysis in Maurice Obstfeld (2010).

An alternative interpretation of the results of the test φ_f is that there is regional heterogeneity in the parameter of interest, that is the effect β of, say, financial openness on central bank reserve holdings differs across the six regions. In this interpretation some of the observed differences between the $\hat{\beta}_j$'s in Panel A of Table 6 are due to heterogeneous

[7] The different results of φ_f for the 4 coefficients are not necessarily contradictory: the structure of intra-region correlation can be such that the standard error estimator, clustered by country, is correct for one coefficient, but too small for another.

means $\beta = \beta_j$, rather than just estimation error $\hat{\beta}_j \sim \mathcal{N}(\beta, \omega_j^2)$. But the homogeneity of β can only be tested with some knowledge of ω_j^2, so that empirically, one cannot distinguish between unspecific heterogeneity in the β_j's and the presence of intra-regional correlations that invalidate the standard error estimator $\hat{\omega}_j^2$. In any event, the analysis in Maurice Obstfeld (2010) implicitly assumes world homogeneity of β, and the test t^{IM} in Panel B of Table 3 provides inference about this parameter allowing for intra-regional spatial correlations.[8]

5.4 Difference-in-Difference: Bloom et al. (2013)

Bloom, Eifert, Mahajan, McKenzie and Roberts (2013) conducted a field experiment on randomly selected firms in the textile industry in India to determine the importance of management practices on productivity. Fourteen treatment plants received extensive management consulting over several months, while 6 control plants were only subject to an initial diagnostic consulting phase that lasted about one month.

Let $y_{i,j,t}$ be a weekly productivity measurement of plant j in week t in the treatment ($i = 1$) and control group ($i = 2$), respectively. Consider an arrangement of data such that the T_0 time periods $t < \tau$ are pre-treatment for all plants, the T_1 time periods $t \geq \tau$ are post-treatment for all plants. Then posit the model

$$y_{i,j,t} = \beta \text{treat}_{i,t} + \kappa_{i,j} + \alpha_t + u_{i,j,t} \tag{15}$$

[8] As noted in Section 3.3 of IM, if the heterogeneity in means arises due to $\beta_j = \beta + \nu_j$, where ν_j is independent across j with a distribution that be written as a scale mixture of mean zero normals, then t^{IM} still provides valid inference about β.

where $\text{treat}_{i,t} = \mathbf{1}[t \geq \tau \text{ and } i = 1]$ is an indicator for treatment, $\kappa_{i,j}$ is a full set of plant fixed effects, α_t is a full set of time fixed effects, and $u_{i,j,t}$ is a mean zero unobserved error term that is independent across firms. The parameter of interest is the coefficient β.

Now construct the difference in average productivity between post- and pre-treatment periods for each plant, $\hat{\delta}_{i,j} = T_1^{-1} \sum_{t \geq \tau} y_{i,j,t} - T_0^{-1} \sum_{t<\tau} y_{i,j,t}$, $j = 1, \ldots, q_i$, $i = 1, 2$. Note that in these differences, the plant fixed effect $\kappa_{i,j}$ cancels, and $E[\hat{\delta}_{i,j}] = \delta_i + m_T$ with $\delta_i = \beta \mathbf{1}[i = 1]$ and $m_T = T_1^{-1} \sum_{t \geq \tau} \alpha_t - T_0^{-1} \sum_{t<\tau} \alpha_t$. Furthermore, in the difference of the differences $\hat{\beta} = \bar{\hat{\delta}}_1 - \bar{\hat{\delta}}_2 = q_1^{-1} \sum_{j=1}^{q_1} \hat{\delta}_{1,j} - q_2^{-1} \sum_{j=1}^{q_2} \hat{\delta}_{2,j}$, as well as in the variance estimators $S_i^2 = (q_j - 1)^{-1} \sum_{j=1}^{q_j} (\hat{\delta}_{i,j} - \bar{\hat{\delta}}_i)^2$, $i = 1, 2$, also the average time fixed effects m_T cancel (cf. the discussion around (13) above).[9] Thus, if there are sufficiently many observations in time and $u_{i,j,t}$ is weakly dependent, then a central limit theorem yields approximate normality for $T_1^{-1} \sum_{t \geq \tau} u_{i,j,t} - T_0^{-1} \sum_{t<\tau} u_{i,j,t}$, (13) holds and inference based on the t-statistic t^{IM} (11) with 5 degrees of freedom is justified via Theorem 1.

Bloom et al. (2013) implement the inference suggested here and find significant effects of the treatment on output, but not on quality defects, inventory and TFP on the 5% level. See their paper for details.

More generally, if within each cluster j of population i, one were to observe several firms $l = 1, \ldots, n_{i,j}$ with firm specific regressors $x_{i,j,t,l}$, one would set $\hat{\delta}_{i,j} = T_1^{-1} \sum_{t \geq \tau} \hat{f}_{i,j,t} -$

[9] For this cancellation to occur, it is necessary that the same time periods correspond to pre- and post-treatment for all observations. As the treatment in Bloom et al. (2013) was staggered in time, this requires omitting some productivity observations in the middle of their sample.

$T_0^{-1}\sum_{t<\tau}\hat{f}_{i,j,t}$, where $\hat{f}_{i,j,t}$ are the OLS estimators of the time fixed effects in a regression of the outcome $y_{i,j,t,l}$ on $x_{i,j,t,l}$ using data from cluster j of population i only, where the regression includes both time and firm fixed effects (with the time fixed effect for $t = 1$, say, normalized to zero). Since $E[\hat{f}_{i,j,t}] = \beta\text{treat}_{i,t} + \alpha_t$, $E[\hat{\delta}_{i,j}] = \delta_i + m_T$ as above.[10] For the approximate normality of $\hat{\delta}_{i,j}$, one could again resort to time series asymptotics under weak dependence, or argue that there are sufficiently many independent firms l in each cluster. Either way, (13) applies, and inference based on the t-test (11) would be asymptotically justified.

6 Conclusion

As the examples of the last section demonstrate, the approach developed in this paper is potentially useful in a variety of contexts, and entirely straightforward to implement. A key regularity assumption is the approximate Gaussianity of estimators from each group,[11] although in contrast to previously developed approaches, no additional homogeneity assumptions on second moments are required. The approximate Gaussianity follows from a central limit theorem if each group contains a reasonably large number of sufficiently independent

[10] If one is willing to assume that firms within a cluster have a common firm fixed effect $\kappa_{i,j,l} = \kappa_{i,j}$, then one could drop the firm fixed effects from the cluster specific regressions, so that now $E[\hat{f}_{i,j,t}] = \beta\text{treat}_{i,t}+\alpha_t+\kappa_{i,j}$, and $\kappa_{i,j}$ again cancels in $E[\hat{\delta}_{i,j}] = \delta_i + m_T$, so (13) remains applicable.

[11] As mentioned at the end of Section 3.2, many forms of heavy-tailed distributions do not actually pose a problem for our approach, but asymmetric distributions generally do.

observations, or if few observations in each group are already averages over sufficiently (observed or unobserved) independent quantities. The appropriateness of such an assumption can be hard to assess in practice. At the same time, some assumption seems necessary: The results of Bahadur and Savage (1956) show that without any constraint on the distribution, it is impossible to conduct inference about the population mean, and thus a fortiori, also about differences between population means. Nonparametric alternatives, such as the Mann-Whitney U test or permutation tests, require that under the null hypothesis, treated and control sample have *identical distributions*, and not just identical means, which can also be quite unappealing in many contexts. We consider the transparency and familiarity of t-statistic based inference an attractive feature of our proposal, and believe that approximate Gaussianity of estimators from each group may at least be a reasonable starting point in many applications.

References

Bahadur, Raghu R. and Leonard J. Savage, "The Non-Existence of Certain Statistical Procedures in Nonparametric Problems," *Annals of Mathematical Statistics*, 1956, *25*, 1115–1122.

Bakirov, Nail K., "The Extrema of the Distribution Function of Student's Ratio for Observations of Unequal Accuracy are Found," *Journal of Soviet Mathematics*, 1989, *44*, 433–440.

——— , "Extrema of the Distribution of Quadratic Forms of Gaussian Variables," *Theory of Probability and its Applications*, 1989, *34*, 207–215.

——— , "Comparison Theorems for Distribution Functions of Quadratic Forms of Gaussian Vectors," *Theory of Probability and its Applications*, 1995, *40*, 340–348.

——— , "Nonhomogenous Samples in the Behrens-Fisher Problem," *Journal of Mathematical Sciences*, 1998, *89*, 1460–1467.

——— **and G. J. Szêkely**, "Student's T-Test for Gaussian Scale Mixtures," *Zapiski Nauchnyh Seminarov POMI*, 2005, *328*, 5–19.

Benjamini, Yoav, "Is the T Test Really Conservative When the Parent Distribution is Long-Tailed?," *Journal of the American Statistical Association*, 1983, *78*, 645–654.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan, "How Much Should We Trust Differences-in-Differences Estimates?," *The Quarterly Journal of Economics*, 2004, *119*, 249–275.

Bester, C. Alan, Timothy G. Conley, and Christian B. Hansen, "Inference with Dependent Data Using Cluster Covariance Estimators," *Journal of Econometrics*, 2011, *165*, 137–151.

Bloom, Nicholas, Benn Eifert, Aprajit Mahajan, David McKenzie, and John Roberts, "Does Management Matter? Evidence from India," *Quarterly Journal of Economics*, 2013, *128*, 1–51.

Borkovec, Milan, "Asymptotic behavior of the sample autocovariance and autocorrelation function of the AR(1) process with ARCH(1) errors," *Bernoulli*, 2001, *7*, 847–872.

Brillinger, David R., "Estimation of the Mean of a Stationary Time Series by Sampling," *Journal of Applied Probability*, 1973, *10*, 419–431.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller, "Bootstrap-Based Improvements for Inference with Clustered Errors," *Review of Economics and Statistics*, 2008, *90*, 414–427.

Canay, Ivan A., Joseph P. Romano, and Azeem M. Shaikh, "Randomization Tests Under an Approximate Symmetry Assumption," *Technical Report No. 2014-13, Stanford University*, 2014.

Carter, Andrew V., Kevin T. Schnepel, and Douglas G. Steigerwald, "Asymptotic Behavior of a t Test Robust to Cluster Heterogeneity," *Working Paper, U.C. Santa Barbara*, 2013.

Chow, Gregory C., "Tests of Equality Between Sets of Coefficients in Two Linear Regressions," *Econometrica*, 1960, *28*, 591–605.

Cont, Rama, "Empirical properties of asset returns: Stylized facts and statistical issues," *Quantitative Finance*, 2001, *1*, 223–236.

Dal Bó, Pedro and Guillaume R. Fréchette, "The Evolution of Cooperation in Infinitely Repeated Games: Experimental Evidence," *American Econonomic Review*, 2011, *101*, 411–429.

Davis, Richard A. and Thomas Mikosch, "Limit theory for the sample ACF of stationary process with heavy tails with applications to ARCH," *Annals of Statistics*, 1998, *26*, 2049–2080.

Donald, Stephen G. and Kevin Lang, "Inference with Difference-in-Differences and Other Panel Data," *The Review of Economics and Statistics*, 2007, *89*, 221–233.

Dufour, Jean-Marie, "Nonuniform Bounds for Nonparametric t-Tests," *Econometric Theory*, 1991, *7*, 253–263.

___ **and Marc Hallin**, "Improved Eaton bounds for linear combinations of bounded random variables, with applications," *Journal of the American Statistical Association*, 1993, *88*, 1026–1033.

Efron, Bradley, "Student's t-Test under Symmetry Conditions," *Journal of the American Statistical Association*, 1969, *64*, 1278–1302.

Fama, Eugene F. and James D. MacBeth, "Risk, Return and Equilibrium: Empirical Tests," *Journal of Political Economy*, 1973, *81*, 607–636.

Hansen, Christian B., "Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data When T is Large," *Journal of Econometrics*, 2007, *141*, 597–620.

Hausman, Jerry A., "Specification Tests in Econometrics," *Econometrica*, 1978, *46*, 1251–1271.

Ibragimov, Rustam and Ulrich K. Müller, "T-Statistic Based Correlation and Heterogeneity Robust Inference," *Journal of Business and Economic Statistics*, 2010, *28*, 453–468.

Imbens, Guido W. and Michal Kolesár, "Robust Standard Errors in Small Samples: Some Practical Advice," *Working Paper, Harvard University*, 2012.

Keim, Donald B., "Size-Related Anaomalies and Stock Return Seasonality," *Journal of Financial Economics*, 1983, *12*, 13–32.

Kiefer, Nicholas M. and Timothy J. Vogelsang, "Heteroskedasticity-Autocorrelation Robust Testing Using Bandwidth Equal to Sample Size," *Econometric Theory*, 2002, *18*, 1350–1366.

____ **and** ____ , "A New Asymptotic Theory for Heteroskedasticity-Autocorrelation Robust Tests," *Econometric Theory*, 2005, *21*, 1130–1164.

MacKinnon, James G. and Matthew D. Webb, "Wild Bootstrap Inference for Wildly Different Cluster Sizes," *QueenŠs Economics Department Working Paper No. 1314*, 2014.

Mickey, M. Ray and Morton B. Brown, "Bounds on the Distribution Functions of the Behrens-Fisher Statistic," *The Annals of Mathematical Statistics*, 1966, *37*, 639–642.

Mikosch, Thomas and Cătălin Stărică, "Limit theory for the sample autocorrelations and extremes of a GARCH(1,1) process," *Annals of Statistics*, 2000, *24*, 1427–1451.

Müller, Ulrich K., "A Theory of Robust Long-Run Variance Estimation," *Journal of Econometrics*, 2007, *141*, 1331–1352.

—, "HAC Corrections for Strongly Autocorrelated Time Series," *Journal of Business and Economic Statistics*, 2014, *32*, 311–322.

Shambaugh, Alan M. Taylor Maurice Obstfeld Jay C., "Financial Stability, the Trilemma, and International Reserves," *American Economic Journal: Macroeconomics*, 2010, *2*, 57–94.

Stock, James H. and Mark W. Watson, "Heteroskedasticity-Robust Standard Errors for Fixed Effect Panel Data Regression," *Econometrica*, 2008, *76*, 155–174.

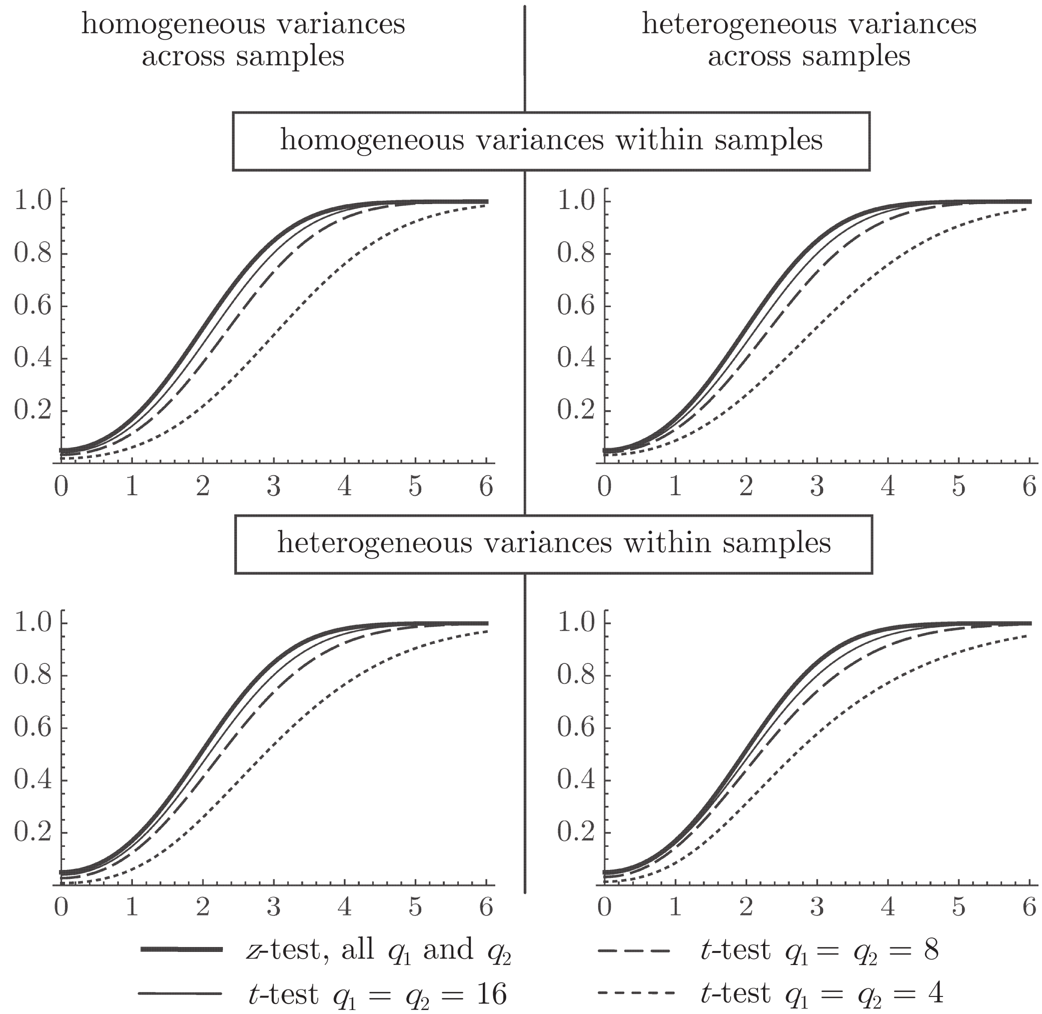
Sun, Yixiao, "Heteroscedasticity and Autocorrelation Robust F Test Using Orthonormal Series Variance Estimator," *The Econometrics Journal*, 2013, *16*, 1–26.

—, "Let's fix it: Fixed-b asymptotics versus small-b asymptotics in heteroskedasticity and autocorrelation robust inference," *Journal of Econometrics*, 2014, *178*, 659–677.

—, Peter C. B. Phillips, and Sainan Jin, "Optimal Bandwidth Selection in Heteroskedasticity-Autocorrelation Robust Testing," *Econometrica*, 2008, *76*, 175–794.

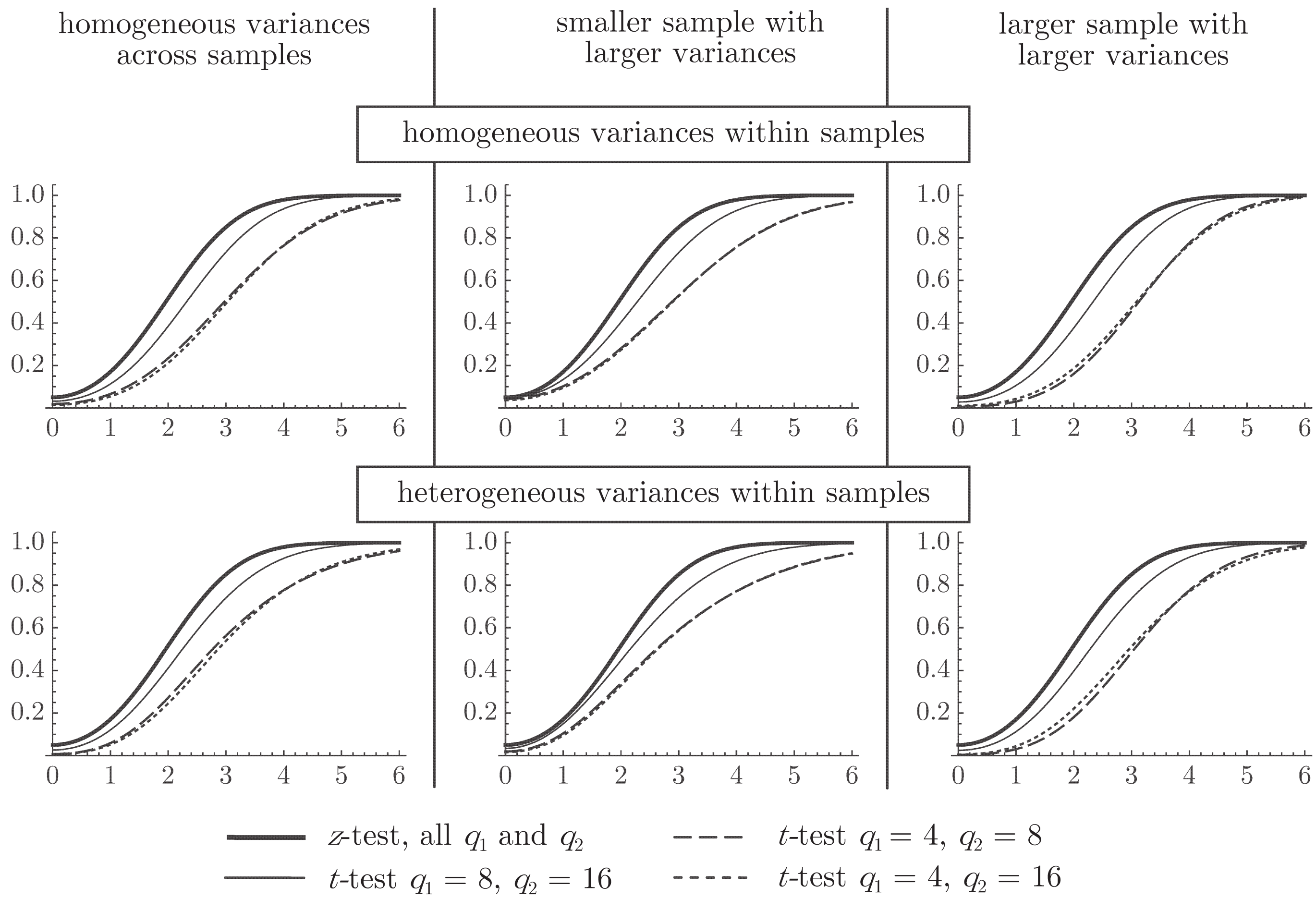
Webb, Matthew D., "Reworking Wild Bootstrap Based Inference for Clustered Errors," *QueenŠs Economics Department Working Paper No. 1315*, 2014.

Figure 1: Rejection Probabilties of Two-Sample t-test and z-test when $q_1 = q_2$



Notes: Rejection probabilities of nominal 5% level t-test (8) and z-test (9), with the alternative for Δ normalized so that $z \sim \mathcal{N}(b, 1)$ throughout, with the value of b reported on the x-axes. Under variance heterogeneity within sample, the variance of the first $q_i/2$ observations is 9 times as as large as of the last $q_i/2$ observations, in both samples. Under variance heterogeneity across samples the variances in one sample is 9 times as large as the variances of the other sample.

Figure 2: Rejection Probabilties of Two-Sample t-test and z-test when $q_1 < q_2$



Notes: See Figure 1.

Table 1: Lower Bounds on Size of Tests in Generic Normal Linear Regression with Few Clusters and Heterogeneous $X_j'X_j$

	CGM					BCH				
	min	Q_1	med	Q_3	max	min	Q_1	med	Q_3	max
					$q = 4$					
$k = 1$	0.0	0.0	0.0	0.0	0.0	3.9	8.6	14.0	17.4	100.0
$k = 2$	3.7	7.2	9.1	12.4	63.6	4.7	8.6	15.3	20.0	100.0
$k = 3$	6.3	11.2	14.5	20.6	68.8	6.2	11.8	17.5	20.8	100.0
					$q = 8$					
$k = 1$	4.3	4.4	4.5	4.8	4.9	6.6	9.7	11.8	16.7	27.3
$k = 2$	4.8	9.1	11.1	13.9	33.2	6.3	9.3	11.6	14.9	27.4
$k = 3$	6.3	10.8	14.0	17.8	35.8	6.0	9.3	11.6	14.7	26.2
					$q = 12$					
$k = 1$	4.5	4.6	4.7	5.0	5.1	7.2	9.8	12.0	15.9	24.2
$k = 2$	5.3	6.8	8.3	10.2	19.9	6.0	8.4	10.0	12.9	28.2
$k = 3$	6.3	8.5	9.9	13.0	22.2	6.3	8.5	10.6	12.8	22.2

Notes: Entries are lower bounds on size in percent of nominal 5 percent tests using the Cameron et al. (2008) (CGM) and Bester et al. (2011) (BCH) methods for inference about a scalar coefficient with q clusters in a linear regression with k non-cluster specific regressors. The columns report the minimum, first quartile, median, third quartile and maximum of the lower bound over 100 draws of $\{X_j'X_j\}_{j=1}^q$ from an i.i.d. Wishart distribution with $2k$

degrees of freedom and scale matrix I_k. Based on 10,000 Monte Carlo draws.

Table 2: Lower Bounds on Size of Tests on Dummy Coefficient in Two-Sample Design of a Normal Linear Regression with Few Clusters and Heterogeneous $X_j'X_j$

	CGM					BCH				
	min	Q_1	med	Q_3	max	min	Q_1	med	Q_3	max
					$q_1+q_2=4$					
$q_1=2$	5.1	11.0	13.1	16.2	32.9	18.3	22.6	35.6	100.0	100.0
					$q_1+q_2=8$					
$q_1=2$	19.6	38.9	42.3	46.1	100.0	20.9	28.9	32.0	100.0	100.0
$q_1=3$	10.8	17.0	20.3	25.9	47.8	11.7	21.7	27.5	32.4	100.0
$q_1=4$	7.5	13.0	15.8	19.4	38.9	13.4	22.8	27.3	30.6	100.0
					$q_1+q_2=12$					
$q_1=2$	36.1	44.2	46.0	47.8	100.0	21.0	32.8	34.5	100.0	100.0
$q_1=3$	16.9	20.6	22.6	28.5	100.0	11.3	19.3	24.2	31.8	100.0
$q_1=4$	10.2	13.1	16.2	21.0	41.3	10.1	16.6	22.2	26.8	100.0
$q_1=5$	8.2	10.8	12.3	16.1	100.0	11.2	16.4	21.7	25.2	100.0
$q_1=6$	7.7	12.1	14.9	18.4	36.1	13.4	20.2	24.1	27.8	100.0

Notes: Entries are lower bounds on size in percent of nominal 5 percent tests using the Cameron et al. (2008) (CGM) and Bester et al. (2011) (BCH) methods for inference about the difference between a scalar regression coefficient between two populations, with q_1 clusters from the first population and q_2 clusters from the second population. The columns report the minimum, first quartile, median, third quartile and maximum of the lower bound over 100 draws of $X_j'X_j$ that are proportional to i.i.d. Chi-squared random variables with 2 degrees of freedom. Based on 10,000 Monte Carlo draws.

Table 3: Small Sample Rejection Probabilities of Level of Clustering Test φ_f

	Null hypothesis						Alternative hypothesis					
	homogenous σ_j			heterogeneous σ_j			homogenous σ_j			heterogeneous σ_j		
$q\backslash T$	5	10	20	5	10	20	5	10	20	5	10	20
Normal innovations $\varepsilon_{j,t} \sim \mathcal{N}(0,1)$												
4	7.2	6.6	5.2	8.3	7.2	5.5	45.7	47.1	46.2	42.8	42.7	41.6
8	6.0	5.4	5.0	7.0	5.9	5.5	65.7	69.1	69.7	58.6	61.8	60.8
16	5.1	5.4	5.0	5.9	5.6	5.1	87.2	89.7	90.0	79.6	82.7	83.1
Chi-squared innovations $\varepsilon_{j,t} \sim \chi_1^2 - 1$												
4	3.9	3.1	3.1	7.9	6.0	5.0	43.5	46.4	45.1	45.0	44.3	42.3
8	2.8	2.4	3.0	4.3	3.1	3.5	58.9	63.9	67.3	57.1	58.4	59.7
16	1.7	1.8	2.5	2.4	2.2	2.5	79.0	86.2	88.9	72.9	78.7	81.5

Notes: Rejection probabilities in percent of nominal 5% level test φ_f. The data generating process is $y_{j,t} = \beta + \sigma_j u_{j,t}$, $t = 1, \cdots, T$, $j = 1, \ldots, q$, where $u_{j,t} = \rho u_{j,t} + \varepsilon_{j,t}$, $u_{j,0} = 0$, and $\varepsilon_{j,t}$ is i.i.d. across j and t. Standard errors $\hat{\omega}_j$ for $\hat{\beta}_j = T^{-1} \sum_{t=1}^{T} y_{j,t}$ are computed via the usual OLS formula $\hat{\omega}_j^2 = (T(T-1))^{-1} \sum_{t=1}^{T} (y_{j,t} - \hat{\beta}_j)^2$, that is "fine clustering" treats all observations as independent. The autocorrelation ρ is zero under the null hypothesis, and $\rho = 0.5$ under the alternative. Under homogeneity, σ_j is a positive constant, and under heterogeneity, half of the groups $j = 1, \ldots, q/2$ have σ_j twice as large as the remaining groups, $\sigma_j = 2\sigma_{q/2+j}$. Based on 10,000 replications.

Table 4: Summary of Empirical Strategy

Single population characterized by β	Two populations characterized by δ_1 and δ_2, interest in $\beta = \delta_1 - \delta_2$
Common computations	
★ Partition sample into q clusters that provide approximately independent and Gaussian information about β.	★ Partition samples from population i into q_i clusters that provide approximately independent and Gaussian information about δ_i, $i = 1, 2$.
★ Estimate the model (including nuisance parameters) using cluster j data only to obtain $\hat{\beta}_j$, $j = 1, \ldots, q$.	★ Estimate the model (including nuisance parameters) using cluster j of population i data only to obtain $\hat{\delta}_{i,j}$, $j = 1, \ldots, q_i$, $i = 1, 2$.
★ Compute $\bar{\hat{\beta}} = q^{-1} \sum_{j=1}^{q} \hat{\beta}_j$ and $S^2 = (q-1)^{-1} \sum_{j=1}^{q} (\hat{\beta}_j - \bar{\hat{\beta}})^2$.	★ Compute $\bar{\hat{\delta}}_i = q_i^{-1} \sum_{j=1}^{q_i} \hat{\delta}_{i,j}$ and $S_i^2 = (q_i - 1)^{-1} \sum_{j=1}^{q_i} (\hat{\delta}_{i,j} - \bar{\hat{\delta}}_i)^2$, $i = 1, 2$.
Inference about β	
★ Reject $H_0 : \beta = \beta_0$ at level α if $\lvert t^{IM} \rvert > \mathrm{cv}(\alpha, q-1)$, where $t^{IM} = \sqrt{q}(\bar{\hat{\beta}} - \beta_0)/S$ and $\mathrm{cv}(\alpha, q-1)$ is the two-sided critical value of the student-t distribution with $q-1$ degrees of freedom of level α. Valid for $\alpha \leq 8.3\%$ for any $q \geq 2$, and for $\alpha \leq 10\%$ for $q \leq 14$.	★ Reject $H_0 : \beta = \beta_0$ at level α if $\lvert t^{IM} \rvert > \mathrm{cv}(\alpha, \min(q_1, q_2) - 1)$, where $t^{IM2} = (\bar{\hat{\delta}}_1 - \bar{\hat{\delta}}_2 - \beta_0)/\sqrt{S_1^2/q_1 + S_2^2/q_2}$ and $\mathrm{cv}(\alpha, m)$ is the two-sided critical value of the student-t distribution with m degrees of freedom of level α. Valid for α an integer multiple of 0.1% for $\alpha \leq 8.3\%$ and any $2 \leq q_1, q_2 \leq 50$, and also for $8.4\% \leq \alpha \leq 10\%$ if $2 \leq q_1, q_2 \leq 14$.
★ 95% confidence set for β has endpoints $\bar{\hat{\beta}} \pm \mathrm{cv}(0.05, q-1)S/\sqrt{q}$.	★ 95% confidence set for β has endpoints $\bar{\hat{\delta}}_1 - \bar{\hat{\delta}}_2 \pm \mathrm{cv}(0.05, \min(q_1, q_2) - 1)\sqrt{S_1^2/q_1 + S_2^2/q_2}$.

★ Student-t p-value (with $q-1$ degrees of freedom) valid for $\lvert t^{IM}\rvert$ if $\lvert t^{IM}\rvert > \text{cv}(0.083, q-1)$ for any $q \geq 2$, and for $2 \leq q \leq 14$ if $\lvert t^{IM}\rvert > \text{cv}(0.1, q-1)$.	★ Student-t p-value (with $\min(q_1, q_2) - 1$ degrees of freedom) rounded up to multiples of 0.1% valid for $\lvert t^{IM2}\rvert$ if $\lvert t^{IM2}\rvert > \text{cv}(0.083, \min(q_1, q_2) - 1)$ for $2 \leq q_1, q_2 \leq 50$, and for $2 \leq q_1, q_2 \leq 14$ if $\lvert t^{IM2}\rvert > \text{cv}(0.1, \min(q_1, q_2) - 1)$.
Test of validity of fine clustering	
★ In estimation of $\hat{\beta}_j$, also estimate its standard error $\hat{\omega}_j$ assuming fine level of clusters is appropriate, $j = 1, \ldots, q$.	★ In estimation of $\hat{\delta}_{i,j}$, also estimate its standard error $\hat{\omega}_{i,j}$ assuming fine level of clusters is appropriate, $j = 1, \ldots, q_i$, $i = 1, 2$.
★ Draw $Z_j \sim iid\mathcal{N}(0,1)$, $j = 1, \ldots, q$, and compute $Y_j = \hat{\omega}_j Z_j$, $\bar{Y} = q^{-1}\sum_{j=1}^{q} Y_j$ and $S_Y^2 = (q-1)^{-1}\sum_{j=1}^{q}(Y_j - \bar{Y})^2$. Repeat 10,000 times.	★ For $i = 1, 2$, draw $Z_{i,j} \sim iid\mathcal{N}(0,1)$, $j = 1, \ldots, q_i$ and compute $Y_{i,j} = \hat{\omega}_{i,j} Z_{i,j}$, $\bar{Y}_i = q_i^{-1}\sum_{j=1}^{q_i} Y_{i,j}$ and $S_{Y,i}^2 = (q_i - 1)^{-1}\sum_{j=1}^{q_i}(Y_{i,j} - \bar{Y}_i)^2$. Compute $U_Y = q_1^{-1} S_{Y,1}^2 + q_2^{-1} S_{Y,2}^2$. Repeat 10,000 times.
★ Reject validity of fine clustering at 5% level if S^2 is larger than 95% quantile of the 10,000 draws of S_Y^2.	★ Reject validity of fine clustering at 5% level if $U = q_1^{-1} S_1^2 + q_2^{-1} S_2^2$ is larger than 95% quantile of the 10,000 draws of U_Y.
★ p-value of test of clustering equals fraction of S_Y^2 larger than S^2.	★ p-value of test of validity of fine clustering equals fraction of U_Y larger than U.

Table 5: Empirical Results in Dal Bó and Fréchette (2011)

Panel A: Probit coefficients $\hat{\delta}_{i,j}$ and estimated standard errors $\hat{\omega}_{i,j}$ (clustered by individual) in session j of treatment i

i	(p, R)	$\hat{\delta}_{i,1}$	$\hat{\omega}_{i,1}$	$\hat{\delta}_{i,2}$	$\hat{\omega}_{i,2}$	$\hat{\delta}_{i,3}$	$\hat{\omega}_{i,3}$
1	$(\frac{1}{2}, 32)$	-1.538	0.163	-0.963	0.183	-1.698	0.216
2	$(\frac{1}{2}, 40)$	-1.052	0.147	-0.813	0.146	-0.878	0.148
3	$(\frac{1}{2}, 48)$	-0.262	0.185	-0.261	0.221	-0.684	0.179
4	$(\frac{3}{4}, 32)$	-0.833	0.142	-0.698	0.167	-0.974	0.198
5	$(\frac{3}{4}, 40)$	0.176	0.153	0.905	0.099	-0.200	0.205
6	$(\frac{3}{4}, 48)$	0.458	0.118	1.037	0.132	0.674	0.113

Panel B: Significance Tests

H_0 is equal cooperation under (p_1, R_1) and (p_2, R_2)							
(p_1, R_1)	$(\frac{1}{2}, 32)$	$(\frac{1}{2}, 40)$	$(\frac{1}{2}, 32)$	$(\frac{1}{2}, 40)$	$(\frac{1}{2}, 48)$	$(\frac{3}{4}, 32)$	$(\frac{3}{4}, 40)$
(p_2, R_2)	$(\frac{1}{2}, 40)$	$(\frac{1}{2}, 48)$	$(\frac{3}{4}, 32)$	$(\frac{3}{4}, 40)$	$(\frac{3}{4}, 48)$	$(\frac{3}{4}, 40)$	$(\frac{3}{4}, 48)$
p-value of Test of $H_0 : \delta_1 = \delta_2$							
Dal Bó and Fréchette	8.6%	0.0%	3.9%	0.0%	0.0%	0.0%	11.5%
t^{IM2} with df $= 2$	$> 10\%$	8.4%	$> 10\%$	6.8%	3.7%	7.8%	$> 10\%$
p-value Test of validity of clustering at level of individuals							
φ_f	2.5%	28.5%	3.6%	0.0%	3.7%	0.0%	0.0%

Notes: For all considered tests, rejections are in the direction of (p_1, R_1) yielding a lower level of cooperation than (p_2, R_2). The row labelled "Dal Bó and Fréchette" reports the original results of Dal Bó and Fréchette (2011) based on probit regressions, clustered by session. The row "t^{IM2} with df=2" implements the two-sample t-test (11) using a critical value with two degrees of freedom, based on the probit coefficient estimates $\hat{\delta}_{i,j}$ of panel A.

Table 6: Empirical Results in Obstfeld, Shambaugh and Taylor (2010)

Panel A: Estimators $\hat{\beta}_j$ and estimated standard errors $\hat{\omega}_j$ (clustered by country)

	fin. openness		Peg		Soft Peg		ln(M2/GDP)	
Region	$\hat{\beta}_j$	$\hat{\omega}_j$	$\hat{\beta}_j$	$\hat{\omega}_j$	$\hat{\beta}_j$	$\hat{\omega}_j$	$\hat{\beta}_j$	$\hat{\omega}_j$
Asia/Pacific	1.110	0.221	0.035	0.113	-0.060	0.119	0.627	0.164
W Europe/N America	0.805	0.430	0.089	0.179	0.069	0.147	1.041	0.319
Eastern Europe	0.423	0.353	0.317	0.168	0.281	0.111	0.633	0.144
Africa	0.508	0.433	0.413	0.151	0.318	0.101	-0.019	0.179
Middle East	1.665	0.438	-0.236	0.193	-0.056	0.153	0.511	0.152
South America	0.770	0.309	-0.279	0.165	-0.067	0.146	-0.201	0.196

Panel B: Significance Tests

variable	fin. openness	Peg	Soft Peg	ln(M2/GDP)
Tests of H_0 is that coefficient of variable is zero				
Obstfeld et al.	0.1%	24.6%	0.8%	0.1%
t^{IM} with df = 5	0.51%	>10%	>10%	7.0%
Tests of validity of clustering at level of countries				
φ_f	19.3%	1.4%	10.8%	0.1%

Notes: For all considered tests, rejections are in the direction of a positive coefficient. The row labelled "Obstfeld et al." reports the original results of Obstfeld et al. (2010) based on a single linear regression, clustered by countries. The row "t^{IM} with df=5" implements the Ibragimov and Müller (2010) t-test using the 6 coefficient estimates $\hat{\beta}_j$ of panel A.