Spatial Prewhitening Simulation

1. Introduction

2. Spatial Prewhitening Methodology

See main paper.

3. Simulation

To do: Add our methodology to the list of estimators.

This section examines the finite sample performance of inference based on data-dependent clusters in a series of simulation experiments roughly based upon the empirical illustration using data from Condra et al (as was done in Cao, Hansen, Kozbur, Villacorta). We present results for inference on a coefficient in a linear model with all exogenous variables and results for inference on the coefficient on an endogenous variable in a linear IV model. We refer to the first case as the "OLS simulation" and to the second case as the "IV simulation." We present the data generating process (DGP) for each case and outline the inferential procedures we consider in Section 3.1. We then present results in Section 3.2.

3.1. Data Generating Processes and Inferential Procedures

For both the OLS and IV simulation, we generate observations indexed by de for $d = 1, ..., 205$ and $e = 1, 2$ where d represents the cross-sectional unit and e the time period.[1] Each observation is thus associated with a spatial location given by $L_d = (\text{lat}_d, \text{long}_{d'})$, the latitude and longitude of the centroid of a district from the observed data in the empirical example, and a temporal index $e \in \{1, 2\}$.

[1] Additional results with 820 cross-sectional units and two time periods are provided in a supplementary appendix.

OLS Simulation DGP: For the OLS simulation, we generate observed data, $\mathscr{D} = \{Y_{de}, X_{de}, W_{de}\}_{d=1,\ldots,205, e=1,2}$ according to

$$Y_{de} = \theta_0 X_{de} + W_{de}'\gamma_0 + U_{de}, \tag{3.1}$$

where X_{de} is the variable of interest, W_{de} is a 10×1 vector of control variables, and U_{de} is unobservable. We set coefficients $\theta_0 = 0$ and $\gamma_0 = 0_{10} = (0, \ldots, 0)'$.

We condition on a single realization of $\{X_{de}, W_{de}\}_{d=1,\ldots,205, e=1,2}$ generated as follows: $\forall\ l \in \{1, \ldots, 10\}$, $\forall\ m \neq l$, $\forall\ (d, e) \in \{1, \ldots, 205\} \times \{1, 2\}$, $\forall\ (d', e') \neq (d, e)$,

$$\begin{aligned} &X_{de} \sim \mathrm{N}(0,1); W_{del} \sim \mathrm{N}(0,1), \\ &\mathrm{corr}(X_{de}, X_{d'e'}) = \mathrm{corr}(W_{del}, W_{d'e'l}) = f_\tau((L_d, e), (L_{d'}, e')), \\ &\mathrm{corr}(X_{de}, W_{del}) = \mathrm{corr}(W_{del}, W_{dem}) = 0.5, \\ &\mathrm{corr}(X_{de}, W_{d'e'l}) = \mathrm{corr}(W_{del}, W_{d'e'm}) = 0, \end{aligned} \tag{3.2}$$

for $f_\tau((L_d, e), (L_{d'}, e'))$ defined in (**??**) with $\tau = (0, 3, 1)'$. Unobservables U_{de} are drawn independently across simulation replications. We consider two settings for the distribution of U_{de}:

A. Homogeneous exponential covariance (BASELINE).

$$U_{de} \sim \mathrm{N}(0,1); \quad \mathrm{corr}(U_{de}, U_{d'e'}) = f_\tau((L_d, e), (L_{d'}, e')), \tag{3.3}$$

   for $f_\tau((L_d, e), (L_{d'}, e'))$ defined in (**??**) with $\tau = (0, 3, 1)'$.

B. Spatial auto-regression (SAR).

$$\begin{aligned} &U_{de} = 0.15 \textstyle\sum_{d' \neq d} U_{d'e} \mathbf{1}_{\{\|L_d - L_{d'}\|_2 < 0.3\}} + \varepsilon_{de}, \\ &\varepsilon_{de} \sim \mathrm{N}(0,1); \quad \mathrm{corr}(\varepsilon_{d1}, \varepsilon_{d2}) = \exp(-1),\ \mathrm{corr}(\varepsilon_{de}, \varepsilon_{d'e'}) = 0 \text{ for } d \neq d'. \end{aligned} \tag{3.4}$$

IV Simulation DGP: For the IV simulation, observed data are $\mathscr{D} = \{Y_{de}, X_{de}, W_{de}, Z_{de}\}_{d=1,\ldots,205, e=1,2}$ with outcomes Y_{de} and endogenous variable X_{de} generated from the system of equations

$$\begin{aligned} Y_{de} &= \theta_0 X_{de} + W_{de}'\gamma_0 + U_{de}, \\ X_{de} &= \pi_0 Z_{de} + W_{de}'\xi_0 + V_{de} \end{aligned} \tag{3.5}$$

where X_{de} is the endogenous variable of interest, W_{de} is a 10×1 vector of control variables, Z_{de} is a scalar instrumental variable, and U_{de} and V_{de} are structural unobservables. We set coefficients $\theta_0 = 0$, $\pi_0 = 2$, and $\gamma_0 = \xi_0 = 0_{10} = (0, \ldots, 0)'$.

We condition on a single realization of the exogenous variables, $\{Z_{de}, W_{de}\}_{d=1,\ldots,205, e=1,2}$, generated as follows: $\forall\ l \in \{1, \ldots, 10\}$, $\forall\ m \neq l$, $\forall\ (d, e) \in \{1, \ldots, 205\} \times \{1, 2\}$, $\forall\ (d', e') \neq (d, e)$,

$$\begin{aligned} &Z_{de} \sim \mathrm{N}(0,1); W_{del} \sim \mathrm{N}(0,1), \\ &\mathrm{corr}(Z_{de}, Z_{d'e'}) = \mathrm{corr}(W_{del}, W_{d'e'l}) = f_\tau((L_d, e), (L_{d'}, e')), \\ &\mathrm{corr}(Z_{de}, W_{del}) = \mathrm{corr}(W_{del}, W_{dem}) = 0.5, \\ &\mathrm{corr}(Z_{de}, W_{d'e'l}) = \mathrm{corr}(W_{del}, W_{d'e'm}) = 0, \end{aligned} \tag{3.6}$$

for $f_\tau((L_d, e), (L_{d'}, e'))$ defined in (**??**) with $\tau = (0, 3, 1)'$. We consider two scenarios for the unobservables (U_{de}, V_{de}) which are drawn independently across simulation replications:

A. Homogeneous exponential covariance (BASELINE).

$$\begin{aligned} &U_{de} \sim \mathrm{N}(0,1),\ V_{de} \sim \mathrm{N}(0,1), \\ &\mathrm{corr}(U_{de}, U_{d'e'}) = \mathrm{corr}(V_{de}, V_{d'e'}) = f_\tau((L_d, e), (L_{d'}, e')), \\ &\mathrm{corr}(U_{de}, V_{de}) = 0.8, \\ &\mathrm{corr}(U_{de}, V_{d'e'}) = 0 \text{ for } (d', e') \neq (d, e), \end{aligned} \tag{3.7}$$

for $f_\tau((L_d, e), (L_{d'}, e'))$ defined in (**??**) with $\tau = (0, 3, 1)'$.

B. Spatial auto-regression (SAR).

$$\begin{aligned} &U_{de} = 0.15 \textstyle\sum_{d' \neq d} U_{d'e} \mathbf{1}_{\{\|L_d - L_{d'}\|_2 < 0.3\}} + \varepsilon_{de}, \\ &V_{de} = 0.15 \textstyle\sum_{d' \neq d} V_{d'e} \mathbf{1}_{\{\|L_d - L_{d'}\|_2 < 0.3\}} + \eta_{de}, \\ &\varepsilon_{de} \sim \mathrm{N}(0,1),\ \eta_{de} \sim \mathrm{N}(0,1), \\ &\mathrm{corr}(\varepsilon_{d1}, \varepsilon_{d2}) = \mathrm{corr}(\eta_{d1}, \eta_{d2}) = \exp(-1), \\ &\mathrm{corr}(\varepsilon_{de}, \eta_{de}) = 0.8, \\ &\mathrm{corr}(\varepsilon_{de}, \varepsilon_{d'e'}) = \mathrm{corr}(\eta_{de}, \eta_{d'e'}) = 0 \text{ for } d' \neq d, \\ &\mathrm{corr}(\varepsilon_{de}, \eta_{d'e'}) = 0 \text{ for } (d', e') \neq (d, e). \end{aligned} \tag{3.8}$$

Inferential Procedures: Within each of the four simulation designs defined above, we generate 1000 simulation replications. We then report results for point estimation and for

inference, focusing on size and power of hypothesis tests, about the parameter θ_0 based on the following procedures:

1. SK. Inference based on the spatial HAC estimator from Sun and Kim (2015) with bandwidth selection adapted from the proposal of Lazarus et al. (2018).
2. UNIT-U. Inference based on the cluster covariance estimator with clusters defined as the cross-sectional unit of observation with a critical value from a t-distribution with $k - 1 = 204$ degrees of freedom.
3. UNIT. Inference based on the cluster covariance estimator with clusters defined as the cross-sectional unit of observation and a critical value obtained by solving (B.8) with cluster structure given by unit-level clustering.
4. CCE. Inference based on the cluster covariance estimator with clusters and rejection threshold obtained by solving (B.8) as described in Section B.
5. IM. Inference based on IM with clusters and rejection threshold obtained by solving (B.8) as described in Section B.
6. CRS. Inference based on CRS with clusters and rejection threshold obtained by solving (B.8) as described in Section B.

For UNIT, CCE, IM, and CRS, we obtain preliminary estimates of unobserved components U_{de} in the OLS setting and (U_{de}, V_{de}) in the IV setting. We then apply Gaussian QMLE with the covariance structure specified in BASELINE using these preliminary estimates as data to obtain the structure for simulating Type I and Type II error rates for use in (B.8). Thus, results for UNIT, CCE, IM, and CRS in the BASELINE OLS and BASELINE IV settings illustrate performance when tuning parameters result from solving (B.8) with a correctly specified model for the covariance structure with feasible estimates of the covariance parameters. In contrast, results from the SAR OLS and SAR IV setting illustrate performance when tuning parameters are obtained by solving (B.8) with a misspecified model. We provide detailed descriptions of the implementation of SK, UNIT, CCE, IM, and CRS in the Appendix.

TABLE 1
Simulation Results

Method	OLS			IV		
	Bias	RMSE	Size	Median Bias	MAD	Size
A. BASELINE						
SK			0.381			0.362
UNIT-U	0.015	0.337	0.577	0.002	0.114	0.568
UNIT			0.047			0.074
CCE			0.046			0.062
IM	0.014	0.213	0.044	-0.059	0.091	0.046
CRS	0.014	0.213	0.042	-0.061	0.095	0.040
B. SAR						
SK			0.447			0.324
UNIT-U	-0.013	0.866	0.639	0.002	0.280	0.502
UNIT			0.359			0.256
CCE			0.047			0.083
IM	-0.006	0.385	0.038	-0.046	0.158	0.025
CRS	-0.003	0.354	0.049	-0.095	0.213	0.041

Notes: Results from the OLS Simulation (in columns labeled "OLS") and IV Simulation (in columns labeled "IV") described in Section 3.1. Row labels indicate inferential method. Panel A corresponds to the BASELINE design where tuning parameters for UNIT, CCE, IM, and CRS are selected based on calculating size and power from feasible estimates of a correctly specific parametric model. Panel B corresponds to the SAR design where tuning parameters for UNIT, CCE, IM, and CRS are selected based on calculating size and power from feasible estimates of a misspecified parametric model. For the OLS designs, we report the bias and RMSE of the point estimator associated with each procedure along with size of 5% level tests. For the IV designs, we report the median bias and MAD of the point estimator associated with each procedure along with size of 5% level tests.

3.2. Simulation Results

We report size of 5% level tests as well as provide results on point estimation in Table 1. In the OLS simulation, we obtain point estimates by applying OLS to estimate the parameters of (3.1), and we obtain point estimates in the IV simulation by applying IV to estimate the parameters in (3.5). Recall that both IM and CRS rely on first obtaining within cluster estimates and have natural point estimator defined by $\frac{1}{\widehat{k}}\sum_{\mathsf{C}\in\widehat{\mathcal{C}}}\widehat{\theta}_{\mathsf{C}}$ where $\widehat{k}$ and $\widehat{\mathcal{C}}$ are the results from solving (B.8) for each procedure and $\widehat{\theta}_{\mathsf{C}}$ is a point estimator that uses only the observations in cluster C. For point estimation, we report bias and root mean square error (RMSE) for the OLS simulation and median bias and median absolute deviation (MAD) for the IV simulation.

The main feature of the results presented in Table 1 is that both IM and CRS with data

dependent groups and rejection threshold determined by (B.8) control size across the four considered designs. CCE with data dependent groups and rejection threshold determined by (B.8) also does a reasonable job controlling size across the designs, though it has size distortions in the IV setting and is also not covered by our theoretical results. The performance of IM, CRS, and CCE is similar both in the BASELINE setting, where tuning parameters $\widehat{k}$ and $\widehat{\alpha}$ are obtained using a correctly specified covariance model, and in the SAR setting, where the problem solved to obtain tuning parameters makes use of a misspecified covariance model.

The robustness of IM, CRS, and, to a lesser extent, CCE is not exhibited by the remaining procedures. The poor behavior of UNIT-U, which ignores spatial dependence entirely, is unsurprising. More surprisingly, SK, which attempts to account for spatial dependence, also does poorly across all designs considered. Over-rejection of the spatial HAC estimator has been previously been documented in the literature. For example, Conley et al. (2021) present simulation results where the null rejection rate for a 5% level test reaches 0.600 and suggest a spatial dependence wild bootstrap approach to improve performance. We also suspect that improvement could substantially be improved by choosing tuning parameters and rejection rule for the spatial HAC procedure by solving a problem similar to (B.8) adapted to that estimator. We did not pursue that direction as we wished to compare to a benchmark from the existing literature.

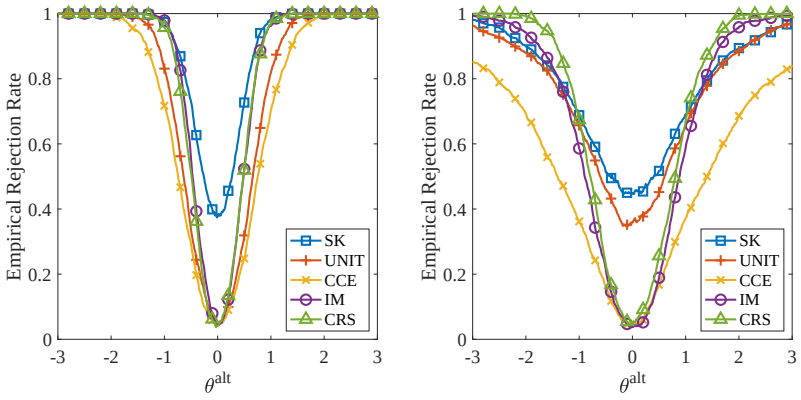
The performance of UNIT is interesting. This procedure treats cross-sectional units as spatially uncorrelated in constructing the standard error estimator but then uses a parametric model that has spatial correlation to adjust the decision threshold for rejecting a hypothesis according to (B.8). In this case, we see that using a correctly specified parametric model for this adjustment restores size control but that size is not controlled under the misspecified parametric structure. This behavior is in line with our theoretical results which rely on the use of a small number of clusters to maintain size control while allowing for relatively general dependence structures and not requiring correct specification of the parametric model used for tuning parameter choice. The simulation clearly demonstrates that robustness to misspecification is not maintained when large numbers of clusters are used.

It is also interesting to look at the point estimation results, though these mostly mirror

results already available in the literature; see, e.g., Ibragimov and Müller (2010), Bester et al. (2011), and Conley et al. (2018). Specifically, we see that both IM and CRS dominate the full sample OLS estimator in terms of both bias and RMSE in our simulation designs. That is, there appears to be a gain, in terms of both point estimation properties and size control, in using the IM or CRS procedure. The results are muddier in the IV simulation where the full sample estimator exhibits lower median bias at the cost of larger MAD and poorer size control. Our preference is still for IM or CRS as having both more accurate statistical inference and more precise (as measured by MAD) estimation seems worth the cost of larger bias. We also note again that these properties appear across simulation studies reported in the literature but are predicated on using a relatively small number of groups as the performance of IM and CRS becomes erratic when small numbers of observations are used to form the within-group estimators. See Conley et al. (2018) for further discussion.

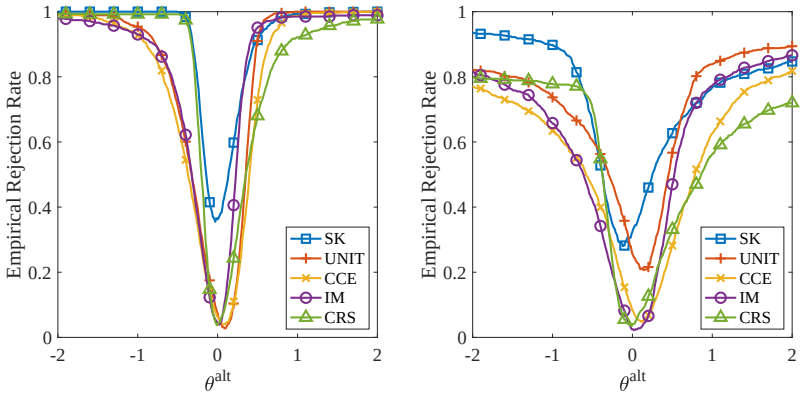
We report power curves for 5% level tests of the hypothesis $H_0 : \theta_0 = \theta^{\text{alt}}$ for alternatives θ^{alt} produced by the different procedures across the simulations in Figures 1 and 2. In these figures, the horizontal axis gives the hypothesized value, θ^{alt}, so size of the test is captured by the point $\theta^{\text{alt}} = 0$. Figure 1 presents the results from the OLS simulation. Here, we see that the power curves are symmetric and that the highest power among procedures that control size is obtained by IM and CRS, both of which perform similarly. Looking to the IV results in Figure 2, we see that power curves are asymmetric and slightly shifted due to the finite sample behavior of the IV estimator. We also see that there is no longer a clear picture about which of the procedures that controls size performs better in terms of power. Specifically, each of CCE, IM, and CRS exhibits higher power over different sets of alternative values. Exploring these tradeoffs more deeply may potentially be interesting but is beyond the scope of this paper. Overall, these figures reinforce the takeaways from the size and point estimation results presented in Table **??** which suggest that IM and CRS provide good default procedures which are robust, both in simulation and in the formal analysis, easy-to-compute, and not clearly dominated by other commonly used approaches.

We report properties of the data dependent number of clusters obtained from (B.8), $\widehat{k}$, across our simulation designs in Table 2. We first note that 5% level tests based on CRS have trivial power (power equal size) when based on fewer than six groups, so the number of groups selected for CRS is always chosen to be six or greater. We see that, with two



(a) BASELINE (b) SAR

Fig 1: OLS power curves.



(a) BASELINE (b) SAR

Fig 2: IV power curves.

exceptions, $\widehat{k} = 8$ is the most likely selection. The exceptions are for IM in the OLS and IV settings with SAR covariance structure where five groups are selected in 55.4% and 56.8% of the simulation replications, respectively. We also note that, while $\widehat{k} = 8$ is the most common solution, there is non-trivial weight on other numbers of groups for all procedures. Finally, we wish to reiterate that the grouping structure is not literally correct in any of our designs but is being used as a feasible way to downweight small covariances to allow for the construction of informative and robust inferential statements. Thus, the goal here is not to find the "correct" number of clusters.

Finally, we summarize the simulation distribution of the data dependent p-value threshold

TABLE 2
Distribution of $\widehat{k}$

	$\widehat{k}=2$	3	4	5	6	7	8
A. OLS - BASELINE							
CCE	0.000	0.000	0.001	0.098	0.311	0.260	0.330
IM	0.000	0.000	0.000	0.007	0.018	0.113	0.862
CRS	0.000	0.000	0.000	0.000	0.001	0.099	0.900
B. OLS - SAR							
CCE	0.000	0.001	0.022	0.051	0.113	0.268	0.545
IM	0.000	0.003	0.069	0.554	0.040	0.176	0.158
CRS	0.000	0.000	0.000	0.000	0.001	0.401	0.598
C. IV - BASELINE							
CCE	0.000	0.001	0.001	0.039	0.242	0.261	0.456
IM	0.000	0.000	0.005	0.051	0.062	0.066	0.816
CRS	0.000	0.000	0.000	0.000	0.000	0.157	0.843
D. IV - SAR							
CCE	0.006	0.027	0.048	0.047	0.071	0.257	0.544
IM	0.005	0.078	0.240	0.568	0.030	0.038	0.041
CRS	0.000	0.000	0.000	0.000	0.032	0.480	0.488

Notes: Simulation results for $\widehat{k}$, the data-dependent number of clusters obtained in (B.8). Panels correspond to the different designs described in Section 3.1. Each entry reports the simulation probability of $\widehat{k}$ being equal to the value given in the column label for the procedure indicated in the row label.

for rejecting a 5% level test obtained from (B.8), $\widehat{\alpha}$, in Table 3. Recall that using the asymptotic approximation underlying each procedure would correspond to rejecting with a threshold of .05. Here we see that for CCE, the distribution of the threshold that would be used to provide exact size control at the 5% level is substantially shifted away from .05 with most of the mass of the distribution over values that are much smaller than .05. This behavior suggests the asymptotic approximation does not provide a particularly reliable guide to the performance of inference based on CCE in our settings which is likely due to strong departures from homogeneity assumptions that are used to establish the asymptotic behavior of CCE with small numbers of groups. For both IM and CRS, we see substantially smaller shifts of $\widehat{\alpha}$ away from .05. Indeed, for both SAR cases, the distribution of $\widehat{\alpha}$ for IM and CRS has a substantial mass point at .05. In the BASELINE cases, solving (B.8) for both IM and CRS does lead to systematic use of p-value thresholds that are smaller than .05 to achieve 5% level size control. While still noticeable, these departures are significantly less extreme than when considering inference based on CCE. Again, this behavior is consistent with the

TABLE 3
Distribution of $\widehat{\alpha}$

	Quantile				
	0.1	0.25	0.5	0.75	0.9
A. OLS - BASELINE					
CCE	0.004	0.006	0.008	0.010	0.012
IM	0.039	0.042	0.046	0.050	0.050
CRS	0.039	0.043	0.047	0.050	0.050
B. OLS - SAR					
CCE	0.021	0.024	0.027	0.031	0.034
IM	0.047	0.049	0.050	0.050	0.050
CRS	0.047	0.047	0.050	0.050	0.050
C. IV - BASELINE					
CCE	0.004	0.005	0.007	0.008	0.010
IM	0.045	0.048	0.050	0.050	0.050
CRS	0.039	0.039	0.047	0.050	0.050
D. IV - SAR					
CCE	0.014	0.020	0.025	0.029	0.033
IM	0.032	0.050	0.050	0.050	0.050
CRS	0.008	0.043	0.050	0.050	0.050

Notes: Simulation results for $\widehat{\alpha}$, the data-dependent p-value threshold for rejecting a hypothesis at the 5% level obtained in (B.8). Panels correspond to the different designs described in Section 3.1. For each design and DGP we report the .1, .25, .5, .75, and .9 quantiles of the simulation distribution of $\widehat{\alpha}$.

more stable and robust performance of IM and CRS relative to CCE. It also highlights how having two tuning parameters is useful for maintaining finite sample properties as size control may not be achievable in finite samples if one may only choose the number of groups for clustering and is unable to adjust the decision threshold.

Overall, the simulation results suggest that IM and CRS with data dependent clusters constructed as outlined in Section B control size in the simulation designs we consider while maintaining non-trivial power. Next, we establish theoretical results that demonstrate that these procedures provide asymptotically valid inference under regularity conditions.

Appendix A: Implementation Details

This section gives complete implementation details for CCE, IM, and CRS in the empirical example and simulation sections as well as SK in the simulation section.

To describe the implementation, introducing vector and matrix notation is helpful. Let

N_{pan} and T_{pan} be the cross-sectional dimension and time dimensions (i.e., sample sizes) of the given panel dataset. Write n as the product $n = N_{\text{pan}}T_{\text{pan}}$. Let Y, X, W, and U be n-row matrices obtained by stacking Y_{de}, X_{de}, $(W'_{de}, 1)$, and U_{de}, respectively. In the IV model, let Z and V be n-row matrices obtained by stacking Z_{de} and V_{de}, respectively. Let $M_A = I_n - A(A'A)^{-1}A'$ for some matrix A. For some $B \in \mathbb{R}^{n\times n}$ with rank ℓ, let P_B be some bijective linear transformation from the subspace orthogonal to the one spanned by the columns of B (thus $(n-\ell)$-dimensional), to $\mathbb{R}^{n-\ell}$.

Step 1. Form candidate clusters. For $k = 2, ..., k_{max}$, apply k-`medoids` with dissimilarity matrix from the data to obtain $\mathcal{C}^{(k)}$, a collection of k partitions of the observations.

Step 2. Fit a parametric model to covariance structure. The implementation details in this step differ slightly in the cases of OLS and IV estimation.

- In the OLS model, let $\widehat{U}$ be the vector of residuals from the full-sample least-squares estimation. Then, under $U \sim \mathrm{N}(0, \Sigma)$,

$$P_{M_{[X,W]}}\widehat{U} \sim \mathrm{N}(0, \Sigma_{PMU}) \quad \text{where} \quad \Sigma_{PMU} = P_{M_{[X,W]}} M_{[X,W]} \Sigma M_{[X,W]} P'_{M_{[X,W]}}.$$

  Note that the covariance matrix Σ_{PMU} is made non-singular by applying the matrix $P_{M_{[X,W]}}$ to $\widehat{U}$. Σ is estimated by QMLE using an exponential covariance model with parameter $\tau = (\tau_1, \tau_2, \tau_3)'$,

$$\Sigma_{de,d'e'}(\tau) = cov[U_{de}, U_{d'e'}; \tau] = \exp(\tau_1)\exp(-\tau_2^{-1}\|L_d - L_{d'}\|_2 - \tau_3^{-1}|e - e'|),$$

  which, in the BASELINE simulation case, is the correct model. To implement, calculate

$$\widehat{\tau} = \arg\max_{\tau} \left\{ \frac{1}{2}\log\det(\Sigma_{PMU}(\tau)) + \frac{1}{2}\widehat{U}' P'_{M_{[X,W]}} (\Sigma_{PMU}(\tau))^{-1} P_{M_{[X,W]}} \widehat{U} \right\},$$

  where $\Sigma_{PMU}(\tau) = P_{M_{[X,W]}} M_{[X,W]} \Sigma(\tau) M_{[X,W]} P'_{M_{[X,W]}}$ and $\Sigma(\tau) = (\Sigma_{de,d'e'}(\tau))_{de,d'e'}$ is the implied covariance matrix of U under τ. The covariance matrix estimator is thus $\Sigma(\widehat{\tau})$.

- In the IV model, the covariance matrices for the structural and first-stage equations are estimated separately. Let $\widehat{U} = M_W Y - M_W X\widehat{\theta}$ and $\widehat{V} = M_W X - M_W Z\widehat{\pi}$, where $\widehat{\theta}$ is the 2SLS estimator for θ_0 and $\widehat{\pi}$ is the least-square estimator for π. Then, the covariance matrices for U and V are estimated by solving

$$\widehat{\tau}^{\varepsilon} = \arg\max_{\tau} \left\{ \frac{1}{2}\log\det(\Sigma_{PM\varepsilon}(\tau)) + \frac{1}{2}\widehat{\varepsilon}' P'_{M_W} (\Sigma_{PM\varepsilon}(\tau))^{-1} P_{M_W} \widehat{\varepsilon} \right\},$$

where $\Sigma_{PM\varepsilon} = P_{M_W} M_W \Sigma(\tau) M_W P'_{M_W}$, $\Sigma(\tau) = (\Sigma_{de,d'e'}(\tau))_{de,d'e'}$, $\Sigma_{de,d'e'}(\tau)$ is as previously defined, and ε is either U or V. Then, the covariance estimators for U and V are $\widehat{\Sigma}_U = \Sigma(\widehat{\tau}^U)$ and $\widehat{\Sigma}_V = \Sigma(\widehat{\tau}^V)$, respectively. Finally, estimate the correlation between first and second stage errors with $\widehat{\rho}$, the empirical correlation between $\widehat{\Sigma}_U^{-1/2}\widehat{U}$ and $\widehat{\Sigma}_V^{-1/2}\widehat{V}$.

Step 3. Simulate data. This step simulates size and power for all candidate partitions $\mathcal{C} \in \mathscr{C} = \{\mathcal{C}^{(2)}, ..., \mathcal{C}^{(k_{\max})}\}$. Given the covariance estimator(s) from *Step 2*, simulate independent copies of the observable data for each $b = 1, ..., B$ as follows for each $\theta \in \{-10/\sqrt{n}, -9/\sqrt{n}, \ldots, -1/\sqrt{n}, 0, 1/\sqrt{n}, 2/\sqrt{n}, \ldots, 10/\sqrt{n}\}$. (We use $B = 10000$ in the empirical example $B = 1000$ in the simulation experiments.)

- In the OLS model, draw U^b from the distribution $N(0, \Sigma(\widehat{\tau}))$. Reproduce data by $Y^b_{de} = \widehat{\alpha} + \theta X_{de} + W'_{de}\widehat{\gamma} + U^b_{de}$, where $\widehat{\alpha}$ and $\widehat{\gamma}$ are full-sample least-square estimators, and U^b_{de} is the de element on U^b.
- In the IV model, draw (U^b, V^b) such that

$$\begin{pmatrix} U^b \\ V^b \end{pmatrix} \sim \mathrm{N}\left(0, \begin{bmatrix} \widehat{\Sigma}_U & \widehat{\rho}\widehat{\Sigma}_U^{1/2}(\widehat{\Sigma}_V^{1/2})' \\ \widehat{\rho}\widehat{\Sigma}_V^{1/2}(\widehat{\Sigma}_U^{1/2})' & \widehat{\Sigma}_V \end{bmatrix}\right).$$

  Reproduce data by

$$\begin{cases} Y^b_{de} = \widehat{\alpha} + \theta X^b_{de} + W'_{de}\widehat{\gamma} + U^b_{de} \\ X^b_{de} = \widehat{\mu} + \widehat{\pi} Z_{de} + X'_{de}\widehat{\xi} + V^b_{de} \end{cases},$$

  where $\widehat{\alpha}$ and $\widehat{\gamma}$ are full-sample 2SLS estimators, $\widehat{\mu}$, $\widehat{\pi}$, $\widehat{\xi}$ are full-sample least-square estimators for the first-stage equation, and U^b_{de}, and V^b_{de} are respectively the de elements of U^b and V^b.

Step 4. Calculate Type I and Type II error rates. For each partition size $k = 2, ..., k_{\max}$ and for each $a \in [0, .05]$, compute simulated Type I error rate $\widehat{\mathrm{Err}}_{\mathrm{Type-I}}(\mathrm{IM}(a), k)$, $\widehat{\mathrm{Err}}_{\mathrm{Type-I}}(\mathrm{CRS}(a), k)$, $\widehat{\mathrm{Err}}_{\mathrm{Type-I}}(\mathrm{CCE}(a), k)$ by testing $H_0 : \theta_0 = 0$ on each simulated dataset with $\theta = 0$ from *Step 3*. Set $\widehat{\alpha}_{\mathrm{IM},k}$, $\widehat{\alpha}_{\mathrm{IM},k}$, $\widehat{\alpha}_{\mathrm{IM},k}$ to be the largest value $a \in [0, \alpha]$ such that $\widehat{\mathrm{Err}}_{\mathrm{Type-I}}(\mathrm{IM}(a), k) \leqslant \alpha$, $\widehat{\mathrm{Err}}_{\mathrm{Type-I}}(\mathrm{CRS}(a), k) \leqslant \alpha$, $\widehat{\mathrm{Err}}_{\mathrm{Type-I}}(\mathrm{CCE}(a), k) \leqslant \alpha$.

For each partition size $k = 2, \ldots, k_{\max}$, compute simulated average Type II error rate $\widehat{\mathrm{Err}}_{\mathrm{Type-II}}(\mathrm{IM}(\widehat{a}_{\mathrm{IM},k}), k)$, $\widehat{\mathrm{Err}}_{\mathrm{Type-II}}(\mathrm{CRS}(\widehat{a}_{\mathrm{CRS},k}), k)$,

$\widehat{\text{Err}}_{\text{Type-II}}(\text{CCE}(\widehat{a}_{\text{CCE},k}), k)$ by testing $H_0 : \theta_0 = 0$ for $\theta \in \{-10/\sqrt{n}, -9/\sqrt{n}, \ldots$ $\ldots, -1/\sqrt{n}, 1/\sqrt{n}, 2/\sqrt{n}, \ldots, 10/\sqrt{n}\}$ on each simulated dataset from *Step 3* and averaging the Type 2 error obtained at each θ value.

Step 5. Solution to (B.8). Set $(\widehat{\alpha}_{\text{CCE}}, \widehat{k}_{\text{CCE}})$, $(\widehat{\alpha}_{\text{IM}}, k_{\text{IM}})$, $(\widehat{\alpha}_{\text{CRS}}, \widehat{k}_{\text{CRS}})$ as the solution to (B.8).

The method SK is based on the spatial-HAC estimator proposed by Sun and Kim (2015). We apply the methods in Sun and Kim (2015) to transform an irregular lattice to a regular integer lattice and to deal with locations that do not form a rectangle. Sun and Kim (2015) require the input of smoothing parameters (K_1, K_2). We apply the method recommended by Lazarus et al. (2018); i.e., we let $K_1 = K_2 = \left[\sqrt{0.4 N_{\text{pan}}^{2/3}}\right]$ where N_{pan} is the number of locations.

References

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. (2017). When should you adjust standard errors for clustering? Technical report, National Bureau of Economic Research.

Andrews, D. W. K. (1991). Asymptotic normality of series estimators for nonparametric and semiparametric regression models. *Econometrica*, 59(2):307–345.

Assoud, P. (1977). *Espaces Métriques, Plongements, Facteurs*. Doctoral Dissertation, Université de Paris XI, 91405 Orsay France.

Bai, J., Choi, S. H., and Liao, Y. (2020). Standard errors for panel data models with unknown clusters. *Journal of Econometrics*.

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.

Bester, C. A., Conley, T. G., Hansen, C. B., and Vogelsang, T. J. (2016). Fixed-b asymptotics for spatially dependent robust nonparametric covariance matrix estimators. *Econometric Theory*, 32(1):154–186.

Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604.

Bolthausen, E. (1982). On the central limit theorem for stationary mixing random fields. *Ann. Probab.*, 10(4):1047–1050.

Cai, Y. (2021). Panel Data with Unknown Clusters. *ArXiv e-prints*.

Canay, I. A., Romano, J. P., and Shaikh, A. M. (2017). Randomization tests under an approximate symmetry assumption. *Econometrica*, 85(3):1013–1030.

Condra, L. N., Long, J. D., Shaver, A. C., and Wright, A. L. (2018). The logic of insurgent electoral violence. *American Economic Review*, 108(11):3199–3231.

Conley, T., Goncalves, S., Kim, M. S. K., and Perron, B. (2021). Bootstrap inference under cross sectional dependence. *Working Paper*.

Conley, T. G. (1999). GMM estimation with cross sectional dependence. *Journal of Econometrics*, 92:1–45.

Conley, T. G. and Dupor, B. (2003). A spatial analysis of sectoral complementarity. *Journal of Political Economy*, 111(2):311–352.

Conley, T. G., Gonçalves, S., and Hansen, C. (2018). Inference with dependent data in accounting and finance applications. *Journal of Accounting Research*, 56:1139–1203.

Conley, T. G. and Topa, G. (2002). Socio-economic distance and spatial patterns in unemployment. *Journal of Applied Econometrics*, 17(4):303–327.

Ferman, B. (2019). A simple way to assess inference methods. *ArXiv e-prints*.

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

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, NY.

Ibragimov, R. and Müller, U. K. (2010). t-statistic based correlation and heterogeneity robust inference. *Journal of Business & Economic Statistics*, 28(4):453–468.

Jenish, N. and Prucha, I. R. (2009). Central limit theorems and uniform laws of large numbers for arrays of random fields. *Journal of Econometrics*, 150(1):86–98.

Jirak, M. (2016). BerryEsseen theorems under weak dependence. *Ann. Probab.*, 44(3):2024–2063.

Kelejian, H. H. and Prucha, I. (2001). On the asymptotic distribution of the Moran I test statistic with applications. *Journal of Econometrics*, 104:219–257.

Kelejian, H. H. and Prucha, I. R. (2007). HAC estimation in a spatial framework. *Journal of Econometrics*, 140(1):131–154.

Kiefer, N. M. and Vogelsang, T. J. (2002). Heteroskedasticity-autocorrelation robust testing using bandwidth equal to sample size. *Econometric Theory*, 18:1350–1366.

Kiefer, N. M. and Vogelsang, T. J. (2005). A new asymptotic theory for heteroskedasticity-autocorrelation robust tests. *Econometric Theory*, 21:1130–1164.

Kiefer, N. M., Vogelsang, T. J., and Bunzel, H. (2000). Simple robust testing of regression hypotheses. *Econometrica*, 68:695–714.

Lazarus, E., Lewis, D. J., and Stock, J. H. (2021). The size-power tradeoff in HAR inference. *Econometrica*, pages 1–60. forthcoming.

Lazarus, E., Lewis, D. J., Stock, J. H., and Watson, M. W. (2018). HAR Inference: Recommendations for Practice. *Journal of Business & Economic Statistics*, 36(4):541–559.

Moran, P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, pages 17–23.

Moreira, H. and Moreira, M. J. (2013). Contributions to the theory of optimal tests.

Müller, U. and Watson, M. (2020). Spatial correlation robust inference. *Working paper*.

Müller, U. K. (2007). A theory of robust long-run variance estimation. *Journal of Econometrics*, 141:1331–1352.

Phillips, P. C. B. (2005). HAC estimation by automated regression. *Econometric Theory*, 21:116–142.

Rothko, M. (1956). Orange and yellow. Oil on Canvas, 231.1 x 180.3 cm, Albright-Knox Art Gallery, Buffalo, NY, US.

Rudelson, M. and Vershynin, R. (2008). The Littlewood-Offord problem and invertibility of random matrices. *Advances in Mathematics*, 218(2):600 – 633.

Sun, Y. (2013). Heteroscedasticity and autocorrelation robust F test using orthonormal series variance estimator. *Econometrics Journal*, 16:1–26.

Sun, Y. and Kim, M. S. (2015). Asymptotic F-test in a GMM framework with cross-sectional dependence. *Review of Economics and Statistics*, 97(1):210–223.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, Massachusetts: The MIT Press, second edition.

A.1. Additional Simulation Results

This section provides additional simulation results to complement those in the main text.

The same settings as in Section 3 are considered; though here we provide results with number of locations given by $N_{\text{pan}} = 820$ in addition to $N_{\text{pan}} = 205$. In the $N_{\text{pan}} = 820$ settings, four copies of the locations from the empirical example are created by reflecting the original locations over the 29° latitude and 75° longitude lines. The data generating process follows Section 3. The maximal number of groups to be considered in CCE, IM, and CRS is chosen to be $k_{\max} = 12$. In all cases, we consider 1000 simulation replications.

The reported results in this section also display more detailed information about the simulation studies than given in the main text in both the $N_{\text{pan}} = 820$ and $N_{\text{pan}} = 205$ cases.

TABLE 4
Summary: OLS - BASELINE ($N_{\text{pan}} = 205$)

Method	Estim. Mean	Estim. RMSE	Size	Power			
				-1	-0.5	0.5	1
SK	0.015	0.337	0.381	0.974	0.730	0.727	0.981
UNIT-U	0.015	0.337	0.577	0.992	0.844	0.832	0.987
UNIT	0.015	0.337	0.047	0.831	0.335	0.319	0.818
CCE	0.015	0.337	0.046	0.717	0.292	0.248	0.704
IM	0.014	0.213	0.044	0.979	0.553	0.523	0.962
CRS	0.014	0.213	0.042	0.957	0.514	0.519	0.970

Notes: Simulation results for estimation in the design described in Section 3. The nominal size is 0.05. Estimates are presented for the estimators, SK, UNIT-U, UNIT, CCE, IM, CRS described in the text. Columns display method, estimated mean, estimated RMSE, size, and power against four alternatives (-1, -0.5, 0.5, 1).

TABLE 5
Summary: OLS- SAR ($N_{\text{pan}} = 205$)

Method	Estim. Mean	Estim. RMSE	Size	Power			
				-1	-0.5	0.5	1
SK	-0.013	0.866	0.447	0.688	0.514	0.517	0.691
UNIT-U	-0.013	0.866	0.639	0.761	0.683	0.681	0.766
UNIT	-0.013	0.866	0.359	0.657	0.461	0.452	0.656
CCE	-0.013	0.866	0.047	0.362	0.160	0.167	0.372
IM	-0.006	0.385	0.038	0.586	0.216	0.189	0.588
CRS	-0.003	0.354	0.049	0.674	0.231	0.256	0.670

Notes: Simulation results for estimation in the design described in Section 3. The nominal size is 0.05. Estimates are presented for the estimators, SK, UNIT-U, UNIT, CCE, IM, CRS described in the text. Columns display method, estimated mean, estimated RMSE, size, and power against four alternatives (-1, -0.5, 0.5, 1).

TABLE 6
Summary: IV - BASELINE ($N_{\text{pan}} = 205$)

Method	Estim. Median	Estim. MAD	Size	Power			
				-1	-0.5	0.5	1
SK	0.002	0.114	0.362	1.000	0.999	0.913	0.993
UNIT-U	0.002	0.114	0.568	0.996	0.950	1.000	1.000
UNIT	0.002	0.114	0.074	0.954	0.711	0.910	1.000
CCE	0.002	0.114	0.062	0.927	0.652	0.729	0.987
IM	-0.059	0.091	0.046	0.930	0.736	0.951	0.985
CRS	-0.061	0.095	0.040	0.992	0.992	0.680	0.921

Notes: Simulation results for estimation in the design described in Section 3. The nominal size is 0.05. Estimates are presented for the estimators, SK, UNIT-U, UNIT, CCE, IM, CRS described in the text. Columns display method, estimated median, estimated MAD, size, and power against four alternatives (-1, -0.5, 0.5, 1).

TABLE 7
Summary: IV - SAR ($N_{\text{pan}} = 205$)

Method	Estim. Median	Estim. MAD	Size	Power			
				-1	-0.5	0.5	1
SK	0.002	0.280	0.324	0.898	0.655	0.627	0.768
UNIT-U	0.002	0.280	0.502	0.797	0.675	0.798	0.942
UNIT	0.002	0.280	0.256	0.737	0.604	0.567	0.839
CCE	0.002	0.280	0.083	0.634	0.464	0.282	0.626
IM	-0.046	0.158	0.025	0.658	0.420	0.470	0.773
CRS	-0.095	0.213	0.041	0.778	0.703	0.331	0.563

Notes: Simulation results for estimation in the design described in Section 3. The nominal size is 0.05. Estimates are presented for the estimators, SK, UNIT-U, UNIT, CCE, IM, CRS described in the text. Columns display method, estimated median, estimated MAD, size, and power against four alternatives (-1, -0.5, 0.5, 1).

TABLE 8
Clustering: OLS - BASELINE ($N_{\text{pan}} = 205$)

		k							$\widehat{k}$
		2	3	4	5	6	7	8	
CCE	size (usual cv)	0.085	0.126	0.154	0.176	0.192	0.201	0.222	
	size (simulated cv)	0.056	0.049	0.048	0.046	0.050	0.042	0.044	0.046
	$\widehat{k}$ frequency	0.000	0.000	0.001	0.098	0.311	0.260	0.330	
	p(sim_size>.05)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	$\widehat{\alpha}$ quantile								
	q10	0.025	0.016	0.009	0.008	0.007	0.005	0.003	0.004
	q25	0.027	0.017	0.010	0.009	0.007	0.006	0.004	0.006
	q50	0.030	0.019	0.011	0.011	0.009	0.007	0.005	0.008
	q75	0.033	0.021	0.012	0.012	0.010	0.008	0.006	0.010
	q90	0.035	0.023	0.014	0.014	0.011	0.009	0.007	0.012
IM	size (usual cv)	0.041	0.058	0.063	0.062	0.054	0.050	0.054	
	size (simulated cv)	0.040	0.053	0.059	0.054	0.047	0.045	0.042	0.044
	$\widehat{k}$ frequency	0.000	0.000	0.000	0.007	0.018	0.113	0.862	
	p(sim_size>.05)	0.383	0.752	0.832	0.909	0.888	0.627	0.791	0.762
	$\widehat{\alpha}$ quantile								
	q10	0.044	0.038	0.037	0.035	0.035	0.040	0.037	0.039
	q25	0.048	0.041	0.040	0.038	0.039	0.044	0.040	0.042
	q50	0.050	0.046	0.044	0.042	0.043	0.048	0.044	0.046
	q75	0.050	0.050	0.048	0.046	0.047	0.050	0.049	0.050
	q90	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
CRS	size (usual cv)	0.000	0.000	0.000	0.000	0.031	0.040	0.048	
	size (simulated cv)	0.000	0.000	0.000	0.000	0.030	0.033	0.041	0.042
	$\widehat{k}$ frequency	0.000	0.000	0.000	0.000	0.001	0.099	0.900	
	p(sim_size>.05)	0.000	0.000	0.000	0.000	0.054	0.530	0.726	0.641
	$\widehat{\alpha}$ quantile								
	q10	0.050	0.050	0.050	0.050	0.050	0.047	0.039	0.039
	q25	0.050	0.050	0.050	0.050	0.050	0.047	0.039	0.043
	q50	0.050	0.050	0.050	0.050	0.050	0.047	0.047	0.047
	q75	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
	q90	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050

Notes: Simulation results for the design described in Section 3. Inferential properties are presented for the estimators, CCE, IM, CRS, described in the text. Columns under k report results with the number of groups fixed at a certain k. $\widehat{k}$ is the number of clusters chosen by the criterion on size-power tradeoff described in the text. Sizes under both usual critical value ("size (usual cv)") and adjusted critical value ("size (simulated cv)") are reported. The rows "$\widehat{k}$ frequency" is the frequency of a particular k achieving the highest simulated power among candidate k's in the setting. The rows "p(sim_size>.05)" report the empirical frequency of $\widehat{\text{Err}}_{\text{Type-I}}(\bullet(.05), k) > .05$. The row "$\widehat{\alpha}$ quantile" report the quantiles of the selected number of groups.

TABLE 9
Clustering: OLS - SAR ($N_{\text{pan}} = 205$)

		k							$\widehat{k}$
		2	3	4	5	6	7	8	
CCE	size (usual cv)	0.041	0.053	0.090	0.067	0.080	0.082	0.095	
	size (simulated cv)	0.034	0.035	0.037	0.037	0.049	0.040	0.043	0.047
	$\widehat{k}$ frequency	0.000	0.001	0.022	0.051	0.113	0.268	0.545	
	p(sim_size>.05)	0.940	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	$\widehat{\alpha}$ quantile								
	q10	0.034	0.026	0.020	0.024	0.021	0.021	0.018	0.021
	q25	0.037	0.029	0.022	0.027	0.023	0.024	0.021	0.024
	q50	0.041	0.032	0.025	0.030	0.026	0.027	0.024	0.027
	q75	0.045	0.035	0.028	0.033	0.030	0.031	0.027	0.031
	q90	0.048	0.038	0.031	0.037	0.033	0.034	0.031	0.034
IM	size (usual cv)	0.022	0.024	0.021	0.027	0.038	0.049	0.051	
	size (simulated cv)	0.021	0.024	0.020	0.026	0.038	0.044	0.047	0.038
	$\widehat{k}$ frequency	0.000	0.003	0.069	0.554	0.040	0.176	0.158	
	p(sim_size>.05)	0.389	0.419	0.385	0.426	0.385	0.536	0.521	0.336
	$\widehat{\alpha}$ quantile								
	q10	0.043	0.043	0.044	0.043	0.044	0.041	0.042	0.047
	q25	0.047	0.047	0.047	0.047	0.048	0.045	0.045	0.049
	q50	0.050	0.050	0.050	0.050	0.050	0.049	0.050	0.050
	q75	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
	q90	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
CRS	size (usual cv)	0.000	0.000	0.000	0.000	0.027	0.050	0.055	
	size (simulated cv)	0.000	0.000	0.000	0.000	0.027	0.041	0.046	0.049
	$\widehat{k}$ frequency	0.000	0.000	0.000	0.000	0.001	0.401	0.598	
	p(sim_size>.05)	0.000	0.000	0.000	0.000	0.000	0.434	0.408	0.268
	$\widehat{\alpha}$ quantile								
	q10	0.050	0.050	0.050	0.050	0.050	0.047	0.047	0.047
	q25	0.050	0.050	0.050	0.050	0.050	0.047	0.047	0.047
	q50	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
	q75	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
	q90	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050

Notes: Simulation results for the design described in Section 3. Inferential properties are presented for the estimators, CCE, IM, CRS, described in the text. Columns under k report results with the number of groups fixed at a certain k. $\widehat{k}$ is the number of clusters chosen by the criterion on size-power tradeoff described in the text. Sizes under both usual critical value ("size (usual cv)") and adjusted critical value ("size (simulated cv)") are reported. The rows "$\widehat{k}$ frequency" is the frequency of a particular k achieving the highest simulated power among candidate k's in the setting. The rows "p(sim_size>.05)" report the empirical frequency of $\widehat{\text{Err}}_{\text{Type-I}}(\bullet(.05), k) > .05$. The row "$\widehat{\alpha}$ quantile" report the quantiles of the selected number of groups.

TABLE 10
Clustering: IV - BASELINE ($N_{\text{pan}} = 205$)

		k							$\widehat{k}$
		2	3	4	5	6	7	8	
CCE	size (usual cv)	0.079	0.123	0.157	0.162	0.176	0.186	0.213	
	size (simulated cv)	0.041	0.053	0.053	0.052	0.058	0.059	0.055	0.062
	$\widehat{k}$ frequency	0.000	0.001	0.001	0.039	0.242	0.261	0.456	
	p(sim_size>.05)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	$\widehat{\alpha}$ quantile								
	q10	0.025	0.016	0.009	0.008	0.006	0.004	0.003	0.004
	q25	0.028	0.017	0.010	0.009	0.007	0.005	0.004	0.005
	q50	0.030	0.019	0.011	0.010	0.008	0.006	0.004	0.007
	q75	0.033	0.021	0.012	0.012	0.010	0.008	0.005	0.008
	q90	0.036	0.023	0.014	0.014	0.011	0.010	0.007	0.010
IM	size (usual cv)	0.045	0.060	0.062	0.062	0.045	0.035	0.044	
	size (simulated cv)	0.045	0.055	0.060	0.054	0.042	0.033	0.042	0.046
	$\widehat{k}$ frequency	0.000	0.000	0.005	0.051	0.062	0.066	0.816	
	p(sim_size>.05)	0.300	0.683	0.702	0.786	0.640	0.280	0.414	0.375
	$\widehat{\alpha}$ quantile								
	q10	0.045	0.039	0.038	0.037	0.040	0.045	0.043	0.045
	q25	0.049	0.042	0.042	0.040	0.043	0.049	0.047	0.048
	q50	0.050	0.047	0.046	0.045	0.048	0.050	0.050	0.050
	q75	0.050	0.050	0.050	0.049	0.050	0.050	0.050	0.050
	q90	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
CRS	size (usual cv)	0.000	0.000	0.000	0.000	0.032	0.041	0.047	
	size (simulated cv)	0.000	0.000	0.000	0.000	0.029	0.032	0.036	0.040
	$\widehat{k}$ frequency	0.000	0.000	0.000	0.000	0.000	0.157	0.843	
	p(sim_size>.05)	0.000	0.000	0.000	0.000	0.038	0.618	0.820	0.682
	$\widehat{\alpha}$ quantile								
	q10	0.050	0.050	0.050	0.050	0.050	0.047	0.039	0.039
	q25	0.050	0.050	0.050	0.050	0.050	0.047	0.039	0.039
	q50	0.050	0.050	0.050	0.050	0.050	0.047	0.047	0.047
	q75	0.050	0.050	0.050	0.050	0.050	0.050	0.047	0.050
	q90	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050

Notes: Simulation results for the design described in Section 3. Inferential properties are presented for the estimators, CCE, IM, CRS, described in the text. Columns under k report results with the number of groups fixed at a certain k. $\widehat{k}$ is the number of clusters chosen by the criterion on size-power tradeoff described in the text. Sizes under both usual critical value ("size (usual cv)") and adjusted critical value ("size (simulated cv)") are reported. The rows "$\widehat{k}$ frequency" is the frequency of a particular k achieving the highest simulated power among candidate k's in the setting. The rows "p(sim_size>.05)" report the empirical frequency of $\widehat{\text{Err}}_{\text{Type-I}}(\bullet(.05), k) > .05$. The row "$\widehat{\alpha}$ quantile" report the quantiles of the selected number of groups.

TABLE 11
Clustering: IV - SAR ($N_{\text{pan}} = 205$)

		k							$\widehat{k}$
		2	3	4	5	6	7	8	
CCE	size (usual cv)	0.039	0.091	0.148	0.095	0.113	0.127	0.138	
	size (simulated cv)	0.030	0.044	0.062	0.058	0.069	0.079	0.082	0.083
	$\widehat{k}$ frequency	0.006	0.027	0.048	0.047	0.071	0.257	0.544	
	p(sim_size>.05)	0.929	0.998	1.000	0.998	0.998	0.998	1.000	0.999
	$\widehat{\alpha}$ quantile								
	q10	0.033	0.023	0.014	0.014	0.009	0.008	0.006	0.014
	q25	0.037	0.027	0.020	0.024	0.019	0.020	0.016	0.020
	q50	0.041	0.031	0.024	0.028	0.025	0.026	0.022	0.025
	q75	0.045	0.034	0.027	0.033	0.029	0.030	0.027	0.029
	q90	0.049	0.038	0.031	0.037	0.033	0.034	0.031	0.033
IM	size (usual cv)	0.025	0.022	0.027	0.022	0.031	0.031	0.034	
	size (simulated cv)	0.025	0.022	0.027	0.022	0.031	0.027	0.031	0.025
	$\widehat{k}$ frequency	0.005	0.078	0.240	0.568	0.030	0.038	0.041	
	p(sim_size>.05)	0.364	0.391	0.257	0.194	0.159	0.175	0.179	0.249
	$\widehat{\alpha}$ quantile								
	q10	0.043	0.041	0.042	0.043	0.031	0.020	0.013	0.032
	q25	0.047	0.046	0.050	0.050	0.050	0.050	0.050	0.050
	q50	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
	q75	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
	q90	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
CRS	size (usual cv)	0.000	0.000	0.000	0.000	0.027	0.045	0.049	
	size (simulated cv)	0.000	0.000	0.000	0.000	0.017	0.034	0.037	0.041
	$\widehat{k}$ frequency	0.000	0.000	0.000	0.000	0.032	0.480	0.488	
	p(sim_size>.05)	0.000	0.000	0.000	0.000	0.209	0.620	0.641	0.497
	$\widehat{\alpha}$ quantile								
	q10	0.050	0.050	0.050	0.050	0.031	0.016	0.008	0.008
	q25	0.050	0.050	0.050	0.050	0.050	0.047	0.031	0.043
	q50	0.050	0.050	0.050	0.050	0.050	0.047	0.047	0.050
	q75	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
	q90	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050

Notes: Simulation results for the design described in Section 3. Inferential properties are presented for the estimators, CCE, IM, CRS, described in the text. Columns under k report results with the number of groups fixed at a certain k. $\widehat{k}$ is the number of clusters chosen by the criterion on size-power tradeoff described in the text. Sizes under both usual critical value ("size (usual cv)") and adjusted critical value ("size (simulated cv)") are reported. The rows "$\widehat{k}$ frequency" is the frequency of a particular k achieving the highest simulated power among candidate k's in the setting. The rows "p(sim_size>.05)" report the empirical frequency of $\widehat{\text{Err}}_{\text{Type-I}}(\bullet(.05), k) > .05$. The row "$\widehat{\alpha}$ quantile" report the quantiles of the selected number of groups.

TABLE 12
Summary: OLS - BASELINE ($N_{\text{pan}} = 820$)

Method	Estim. Mean	Estim. RMSE	Size	Power			
				-1	-0.5	0.5	1
SK	0.002	0.259	0.395	0.998	0.856	0.847	0.998
UNIT-U	0.002	0.259	0.700	1.000	0.938	0.944	1.000
UNIT	0.002	0.259	0.039	0.962	0.517	0.496	0.960
CCE	0.002	0.259	0.052	0.917	0.449	0.449	0.899
IM	-0.001	0.146	0.058	1.000	0.877	0.881	1.000
CRS	-0.001	0.146	0.058	1.000	0.871	0.868	1.000

Notes: Simulation results for estimation in the design described in Section 3. The nominal size is 0.05. Estimates are presented for the estimators, SK, UNIT-U, UNIT, CCE, IM, CRS described in the text. Columns display method, estimated mean, estimated RMSE, size, and power against four alternatives (-1, -0.5, 0.5, 1).

TABLE 13
Summary: OLS- SAR ($N_{\text{pan}} = 820$)

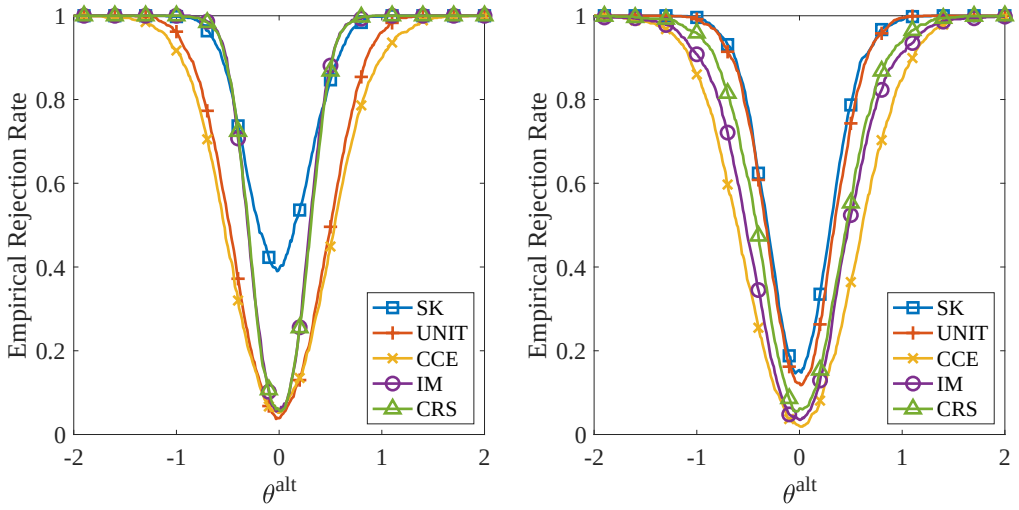
Method	Estim. Mean	Estim. RMSE	Size	Power			
				-1	-0.5	0.5	1
SK	0.003	0.229	0.154	0.996	0.780	0.787	0.990
UNIT-U	0.003	0.229	0.405	0.999	0.909	0.903	1.000
UNIT	0.003	0.229	0.123	0.992	0.756	0.743	0.995
CCE	0.003	0.229	0.020	0.860	0.363	0.364	0.848
IM	-0.013	0.249	0.035	0.908	0.465	0.524	0.911
CRS	-0.013	0.229	0.062	0.959	0.612	0.553	0.938

Notes: Simulation results for estimation in the design described in Section 3. The nominal size is 0.05. Estimates are presented for the estimators, SK, UNIT-U, UNIT, CCE, IM, CRS described in the text. Columns display method, estimated mean, estimated RMSE, size, and power against four alternatives (-1, -0.5, 0.5, 1).

TABLE 14
Summary: IV - BASELINE ($N_{\text{pan}} = 820$)

Method	Estim. Median	Estim. MAD	Size	Power			
				-1	-0.5	0.5	1
SK	0.004	0.089	0.390	1.000	1.000	0.973	1.000
UNIT-U	0.004	0.089	0.700	1.000	0.995	1.000	1.000
UNIT	0.004	0.089	0.048	0.998	0.870	1.000	1.000
CCE	0.004	0.089	0.051	0.989	0.837	0.963	1.000
IM	-0.041	0.061	0.055	0.994	0.940	0.998	0.999
CRS	-0.042	0.061	0.056	0.999	0.999	0.930	0.993

Notes: Simulation results for estimation in the design described in Section 3. The nominal size is 0.05. Estimates are presented for the estimators, SK, UNIT-U, UNIT, CCE, IM, CRS described in the text. Columns display method, estimated median, estimated MAD, size, and power against four alternatives (-1, -0.5, 0.5, 1).



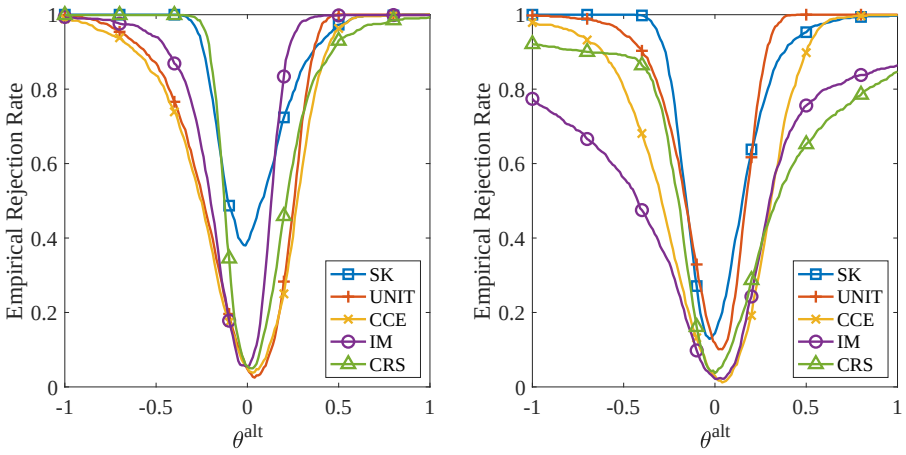
(a) BASELINE (b) SAR

Fig 3: OLS power curves ($N_{\text{pan}} = 820$).

TABLE 15
Summary: IV - SAR ($N_{\text{pan}} = 820$)

Method	Estim. Median	Estim. MAD	Size	Power			
				-1	-0.5	0.5	1
SK	0.003	0.076	0.144	1.000	1.000	0.953	0.997
UNIT-U	0.003	0.076	0.389	1.000	0.983	1.000	1.000
UNIT	0.003	0.076	0.118	0.998	0.948	1.000	1.000
CCE	0.003	0.076	0.030	0.979	0.804	0.898	0.999
IM	-0.052	0.130	0.022	0.774	0.561	0.756	0.864
CRS	-0.052	0.123	0.037	0.921	0.891	0.652	0.847

Notes: Simulation results for estimation in the design described in Section 3. The nominal size is 0.05. Estimates are presented for the estimators, SK, UNIT-U, UNIT, CCE, IM, CRS described in the text. Columns display method, estimated median, estimated MAD, size, and power against four alternatives (-1, -0.5, 0.5, 1).



(a) BASELINE (b) SAR

Fig 4: IV power curves ($N_{\text{pan}} = 820$).

TABLE 16
Clustering: OLS - BASELINE ($N_{\text{pan}} = 820$)

		k											$\widehat{k}$
		2	3	4	5	6	7	8	9	10	11	12	
CCE	size (usual cv)	0.056	0.093	0.117	0.146	0.180	0.178	0.177	0.188	0.209	0.209	0.218	
	size (simulated cv)	0.041	0.045	0.046	0.040	0.044	0.050	0.051	0.050	0.047	0.047	0.049	0.052
	$\widehat{k}$ frequency	0.000	0.003	0.006	0.045	0.086	0.068	0.119	0.098	0.189	0.150	0.236	
	p(sim_size>.05)	0.994	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	$\widehat{\alpha}$ quantile												
	q10	0.030	0.021	0.015	0.011	0.008	0.008	0.007	0.006	0.004	0.004	0.004	0.005
	q25	0.033	0.023	0.017	0.011	0.009	0.009	0.008	0.007	0.005	0.005	0.004	0.006
	q50	0.036	0.025	0.019	0.013	0.010	0.011	0.010	0.008	0.006	0.006	0.005	0.008
	q75	0.039	0.028	0.021	0.014	0.011	0.012	0.011	0.009	0.007	0.007	0.006	0.010
	q90	0.042	0.030	0.023	0.016	0.012	0.013	0.012	0.010	0.009	0.008	0.007	0.013
IM	size (usual cv)	0.046	0.060	0.059	0.064	0.064	0.075	0.085	0.091	0.096	0.090	0.076	
	size (simulated cv)	0.041	0.053	0.054	0.057	0.052	0.062	0.069	0.057	0.067	0.059	0.053	0.058
	$\widehat{k}$ frequency	0.000	0.000	0.000	0.000	0.000	0.002	0.013	0.013	0.005	0.195	0.772	
	p(sim_size>.05)	0.492	0.565	0.727	0.675	0.828	0.966	0.966	0.998	1.000	1.000	0.999	0.997
	$\widehat{\alpha}$ quantile												
	q10	0.042	0.041	0.039	0.039	0.037	0.032	0.031	0.025	0.021	0.024	0.024	0.026
	q25	0.046	0.045	0.042	0.043	0.039	0.036	0.035	0.028	0.024	0.027	0.027	0.029
	q50	0.050	0.049	0.046	0.047	0.044	0.039	0.038	0.032	0.027	0.030	0.031	0.032
	q75	0.050	0.050	0.050	0.050	0.048	0.043	0.042	0.035	0.031	0.034	0.035	0.036
	q90	0.050	0.050	0.050	0.050	0.050	0.047	0.046	0.039	0.035	0.038	0.038	0.040
CRS	size (usual cv)	0.000	0.000	0.000	0.000	0.034	0.074	0.082	0.088	0.100	0.093	0.077	
	size (simulated cv)	0.000	0.000	0.000	0.000	0.033	0.047	0.062	0.053	0.058	0.059	0.053	0.058
	$\widehat{k}$ frequency	0.000	0.000	0.000	0.000	0.000	0.001	0.004	0.009	0.004	0.170	0.812	
	p(sim_size>.05)	0.000	0.000	0.000	0.000	0.027	0.882	0.923	0.993	1.000	0.998	0.999	0.992
	$\widehat{\alpha}$ quantile												
	q10	0.050	0.050	0.050	0.050	0.050	0.047	0.031	0.027	0.021	0.023	0.023	0.024
	q25	0.050	0.050	0.050	0.050	0.050	0.047	0.039	0.029	0.023	0.026	0.026	0.027
	q50	0.050	0.050	0.050	0.050	0.050	0.047	0.039	0.033	0.027	0.030	0.030	0.031
	q75	0.050	0.050	0.050	0.050	0.050	0.047	0.047	0.035	0.031	0.034	0.034	0.035
	q90	0.050	0.050	0.050	0.050	0.050	0.050	0.047	0.039	0.035	0.038	0.038	0.039

Notes: Simulation results for the design described in Section 3. Inferential properties are presented for the estimators, CCE, IM, CRS, described in the text. Columns under k report results with the number of groups fixed at a certain k. $\widehat{k}$ is the number of clusters chosen by the criterion on size-power tradeoff described in the text. Sizes under both usual critical value ("size (usual cv)") and adjusted critical value ("size (simulated cv)") are reported. The rows "$\widehat{k}$ frequency" is the frequency of a particular k achieving the highest simulated power among candidate k's in the setting. The rows "p(sim_size>.05)" report the empirical frequency of $\widehat{\text{Err}}_{\text{Type-I}}(\bullet(.05), k) > .05$. The row "$\widehat{\alpha}$ quantile" report the quantiles of the selected number of groups.

TABLE 17
Clustering: OLS - SAR ($N_{\text{pan}} = 820$)

		k											$\widehat{k}$
		2	3	4	5	6	7	8	9	10	11	12	
CCE	size (usual cv)	0.040	0.043	0.037	0.033	0.033	0.035	0.035	0.045	0.047	0.048	0.052	
	size (simulated cv)	0.036	0.037	0.026	0.020	0.021	0.017	0.012	0.017	0.019	0.020	0.020	0.020
	$\widehat{k}$ frequency	0.000	0.000	0.004	0.008	0.028	0.052	0.105	0.113	0.167	0.220	0.303	
	p(sim_size>.05)	0.767	0.987	0.992	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	$\widehat{\alpha}$ quantile												
	q10	0.038	0.030	0.028	0.024	0.021	0.023	0.022	0.020	0.019	0.018	0.016	0.020
	q25	0.041	0.033	0.031	0.027	0.024	0.025	0.024	0.022	0.021	0.020	0.019	0.022
	q50	0.045	0.036	0.035	0.030	0.027	0.028	0.027	0.025	0.024	0.023	0.021	0.026
	q75	0.050	0.040	0.038	0.033	0.030	0.031	0.030	0.028	0.027	0.026	0.024	0.029
	q90	0.050	0.044	0.041	0.036	0.033	0.034	0.033	0.031	0.029	0.028	0.026	0.033
IM	size (usual cv)	0.051	0.058	0.027	0.024	0.027	0.037	0.040	0.039	0.034	0.034	0.042	
	size (simulated cv)	0.047	0.055	0.026	0.020	0.024	0.034	0.037	0.036	0.033	0.032	0.040	0.035
	$\widehat{k}$ frequency	0.000	0.000	0.002	0.006	0.073	0.044	0.409	0.223	0.061	0.067	0.115	
	p(sim_size>.05)	0.495	0.366	0.515	0.383	0.471	0.522	0.535	0.491	0.453	0.482	0.483	0.354
	$\widehat{\alpha}$ quantile												
	q10	0.042	0.043	0.041	0.044	0.042	0.041	0.041	0.042	0.043	0.042	0.042	0.047
	q25	0.045	0.047	0.045	0.048	0.046	0.045	0.045	0.046	0.046	0.046	0.046	0.049
	q50	0.050	0.050	0.050	0.050	0.050	0.049	0.049	0.050	0.050	0.050	0.050	0.050
	q75	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
	q90	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
CRS	size (usual cv)	0.000	0.000	0.000	0.000	0.022	0.053	0.050	0.054	0.048	0.047	0.058	
	size (simulated cv)	0.000	0.000	0.000	0.000	0.022	0.049	0.048	0.054	0.046	0.040	0.053	0.062
	$\widehat{k}$ frequency	0.000	0.000	0.000	0.000	0.000	0.019	0.199	0.143	0.127	0.188	0.324	
	p(sim_size>.05)	0.000	0.000	0.000	0.000	0.001	0.355	0.342	0.383	0.487	0.515	0.522	0.277
	$\widehat{\alpha}$ quantile												
	q10	0.050	0.050	0.050	0.050	0.050	0.047	0.047	0.043	0.042	0.042	0.041	0.047
	q25	0.050	0.050	0.050	0.050	0.050	0.047	0.047	0.047	0.046	0.045	0.045	0.050
	q50	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
	q75	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
	q90	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050

Notes: Simulation results for the design described in Section 3. Inferential properties are presented for the estimators, CCE, IM, CRS, described in the text. Columns under k report results with the number of groups fixed at a certain k. $\widehat{k}$ is the number of clusters chosen by the criterion on size-power tradeoff described in the text. Sizes under both usual critical value ("size (usual cv)") and adjusted critical value ("size (simulated cv)") are reported. The rows "$\widehat{k}$ frequency" is the frequency of a particular k achieving the highest simulated power among candidate k's in the setting. The rows "p(sim_size>.05)" report the empirical frequency of $\widehat{\text{Err}}_{\text{Type-I}}(\bullet(.05), k) > .05$. The row "$\widehat{\alpha}$ quantile" report the quantiles of the selected number of groups.

TABLE 18
Clustering: IV - BASELINE ($N_{\text{pan}} = 820$)

		k											$\widehat{k}$
		2	3	4	5	6	7	8	9	10	11	12	
CCE	size (usual cv)	0.059	0.091	0.113	0.150	0.171	0.156	0.169	0.186	0.201	0.197	0.210	
	size (simulated cv)	0.044	0.047	0.048	0.043	0.046	0.047	0.054	0.052	0.047	0.040	0.041	0.051
	$\widehat{k}$ frequency	0.000	0.002	0.008	0.009	0.053	0.038	0.069	0.096	0.185	0.207	0.333	
	p(sim_size>.05)	0.994	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	$\widehat{\alpha}$ quantile												
	q10	0.030	0.021	0.015	0.010	0.007	0.008	0.007	0.005	0.004	0.004	0.003	0.004
	q25	0.033	0.023	0.017	0.011	0.008	0.009	0.008	0.006	0.005	0.005	0.004	0.005
	q50	0.036	0.025	0.019	0.012	0.009	0.010	0.009	0.008	0.006	0.006	0.005	0.007
	q75	0.039	0.028	0.021	0.014	0.011	0.012	0.011	0.009	0.007	0.007	0.006	0.009
	q90	0.042	0.030	0.023	0.016	0.012	0.014	0.012	0.010	0.009	0.009	0.007	0.011
IM	size (usual cv)	0.050	0.054	0.058	0.064	0.062	0.069	0.081	0.087	0.093	0.079	0.074	
	size (simulated cv)	0.048	0.051	0.057	0.059	0.054	0.056	0.073	0.063	0.065	0.057	0.054	0.055
	$\widehat{k}$ frequency	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.009	0.000	0.128	0.862	
	p(sim_size>.05)	0.483	0.544	0.636	0.528	0.694	0.890	0.903	0.988	0.998	0.993	0.998	0.993
	$\widehat{\alpha}$ quantile												
	q10	0.042	0.042	0.039	0.041	0.039	0.035	0.035	0.028	0.025	0.028	0.028	0.028
	q25	0.046	0.045	0.043	0.045	0.043	0.039	0.038	0.032	0.028	0.031	0.031	0.031
	q50	0.050	0.049	0.048	0.049	0.047	0.043	0.042	0.035	0.032	0.034	0.034	0.035
	q75	0.050	0.050	0.050	0.050	0.050	0.046	0.046	0.039	0.035	0.038	0.038	0.039
	q90	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.043	0.039	0.042	0.041	0.042
CRS	size (usual cv)	0.000	0.000	0.000	0.000	0.034	0.072	0.084	0.087	0.107	0.098	0.089	
	size (simulated cv)	0.000	0.000	0.000	0.000	0.033	0.043	0.060	0.058	0.060	0.057	0.053	0.056
	$\widehat{k}$ frequency	0.000	0.000	0.000	0.000	0.000	0.001	0.001	0.008	0.001	0.125	0.864	
	p(sim_size>.05)	0.000	0.000	0.000	0.000	0.030	0.889	0.941	0.996	1.000	1.000	1.000	0.996
	$\widehat{\alpha}$ quantile												
	q10	0.050	0.050	0.050	0.050	0.050	0.047	0.031	0.027	0.021	0.022	0.022	0.022
	q25	0.050	0.050	0.050	0.050	0.050	0.047	0.039	0.029	0.023	0.025	0.024	0.025
	q50	0.050	0.050	0.050	0.050	0.050	0.047	0.039	0.031	0.026	0.028	0.028	0.028
	q75	0.050	0.050	0.050	0.050	0.050	0.047	0.047	0.035	0.029	0.031	0.031	0.032
	q90	0.050	0.050	0.050	0.050	0.050	0.050	0.047	0.039	0.033	0.035	0.034	0.036

Notes: Simulation results for the design described in Section 3. Inferential properties are presented for the estimators, CCE, IM, CRS, described in the text. Columns under k report results with the number of groups fixed at a certain k. $\widehat{k}$ is the number of clusters chosen by the criterion on size-power tradeoff described in the text. Sizes under both usual critical value ("size (usual cv)") and adjusted critical value ("size (simulated cv)") are reported. The rows "$\widehat{k}$ frequency" is the frequency of a particular k achieving the highest simulated power among candidate k's in the setting. The rows "p(sim_size>.05)" report the empirical frequency of $\widehat{\text{Err}}_{\text{Type-I}}(\bullet(.05), k) > .05$. The row "$\widehat{\alpha}$ quantile" report the quantiles of the selected number of groups.

TABLE 19
Clustering: IV - SAR ($N_{\text{pan}} = 820$)

		k											$\widehat{k}$
		2	3	4	5	6	7	8	9	10	11	12	
CCE	size (usual cv)	0.034	0.042	0.033	0.035	0.041	0.030	0.040	0.041	0.041	0.043	0.047	
	size (simulated cv)	0.029	0.035	0.021	0.021	0.019	0.020	0.022	0.025	0.027	0.029	0.030	0.030
	$\widehat{k}$ frequency	0.000	0.000	0.000	0.000	0.007	0.016	0.046	0.076	0.151	0.268	0.436	
	p(sim_size>.05)	0.800	0.992	0.994	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000	1.000
	$\widehat{\alpha}$ quantile												
	q10	0.037	0.030	0.028	0.024	0.022	0.022	0.022	0.020	0.019	0.018	0.016	0.019
	q25	0.041	0.033	0.031	0.026	0.024	0.025	0.024	0.022	0.021	0.020	0.019	0.021
	q50	0.044	0.036	0.035	0.029	0.027	0.028	0.027	0.025	0.024	0.023	0.021	0.024
	q75	0.049	0.040	0.039	0.033	0.030	0.031	0.031	0.028	0.027	0.026	0.024	0.027
	q90	0.050	0.043	0.042	0.036	0.033	0.034	0.034	0.031	0.030	0.029	0.027	0.031
IM	size (usual cv)	0.041	0.056	0.029	0.021	0.020	0.037	0.030	0.028	0.026	0.019	0.026	
	size (simulated cv)	0.039	0.055	0.025	0.019	0.019	0.035	0.029	0.028	0.026	0.019	0.026	0.022
	$\widehat{k}$ frequency	0.000	0.004	0.008	0.008	0.153	0.003	0.777	0.046	0.001	0.000	0.000	
	p(sim_size>.05)	0.481	0.321	0.399	0.234	0.256	0.224	0.240	0.176	0.115	0.104	0.139	0.160
	$\widehat{\alpha}$ quantile												
	q10	0.042	0.045	0.043	0.046	0.046	0.047	0.046	0.048	0.049	0.050	0.049	0.049
	q25	0.046	0.048	0.047	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
	q50	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
	q75	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
	q90	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
CRS	size (usual cv)	0.000	0.000	0.000	0.000	0.025	0.054	0.051	0.045	0.042	0.039	0.052	
	size (simulated cv)	0.000	0.000	0.000	0.000	0.025	0.044	0.041	0.042	0.040	0.032	0.038	0.037
	$\widehat{k}$ frequency	0.000	0.000	0.000	0.000	0.000	0.031	0.469	0.320	0.090	0.048	0.042	
	p(sim_size>.05)	0.000	0.000	0.000	0.000	0.002	0.389	0.446	0.607	0.780	0.899	0.957	0.352
	$\widehat{\alpha}$ quantile												
	q10	0.050	0.050	0.050	0.050	0.050	0.047	0.047	0.039	0.038	0.036	0.034	0.045
	q25	0.050	0.050	0.050	0.050	0.050	0.047	0.047	0.043	0.041	0.039	0.036	0.047
	q50	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.047	0.045	0.042	0.040	0.050
	q75	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.046	0.043	0.050
	q90	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.047	0.050

Notes: Simulation results for the design described in Section 3. Inferential properties are presented for the estimators, CCE, IM, CRS, described in the text. Columns under k report results with the number of groups fixed at a certain k. $\widehat{k}$ is the number of clusters chosen by the criterion on size-power tradeoff described in the text. Sizes under both usual critical value ("size (usual cv)") and adjusted critical value ("size (simulated cv)") are reported. The rows "$\widehat{k}$ frequency" is the frequency of a particular k achieving the highest simulated power among candidate k's in the setting. The rows "p(sim_size>.05)" report the empirical frequency of $\widehat{\text{Err}}_{\text{Type-I}}(\bullet(.05), k) > .05$. The row "$\widehat{\alpha}$ quantile" report the quantiles of the selected number of groups.

A.2. Computation times

The procedure defined in Algorithm 1 in the main text is computationally intensive. This section reports various computation times associated with running Algorithm 1 as well as the alternative techniques described in the main text.

We present two tables documenting computation time. Table 20 records the computation time for all inferential methods in all simulation settings. Table 21 records the number of iterations needed to reach convergence in k-medoids using the metric space defined by the Afghanistan district centroids as described in the main text. Note that we apply k-medoids to 100 sets of randomly chosen centroids and pick the one with the lowest objective function value at convergence.

The experiment is implemented in MATLAB 9.6.0.1072779 (R2019a). Two models of CPUs are used: Intel Xeon E5-2690 v3 2.60GHz and Intel Xeon E5-2680 v4 2.40GHz. We assigned 10 cores and 10 GB memory for each task. The operating system is Linux 3.10.0-1160.25.1.el7.x86_64.

TABLE 20
Runtime in seconds

Method	$N = 205$				$N = 820$			
	OLS		IV		OLS		IV	
	BASELINE	SAR	BASELINE	SAR	BASELINE	SAR	BASELINE	SAR
SK	52	38	44	19	19	38	102	166
UNIT-U	51	85	48	87	116	147	45	20
UNIT	49	24	58	23	223	210	279	231
CCE	71	67	62	36	137	206	427	293
IM	72	66	61	60	282	214	308	358
CRS	119	118	107	95	409	395	504	496

Notes: This table records computation runtime (in seconds) of implementing a specific method in the corresponding setting for once.

Appendix B: Comparison Methodology: Inference with Unsupervised Cluster Learning

Consider data given by $\mathscr{D} = \{\zeta_i\}_{i \in \mathsf{X}}$. Here, ζ_i are observable random variables or vectors and X is a (spatial) indexing set of cardinality n. This paper assumes that X is equipped with a known dissimilarity measure d, which is an $n \times n$ array of nonnegative real dissimilarities.

TABLE 21
Number of iterations in k-medoids

G	$N = 205$		$N = 820$	
	best	average	best	average
2	3	2.71	3	2.59
3	3	3.11	3	2.89
4	3	3.31	2	2.58
5	3	3.67	3	9.95
6	4	3.79	3	22.67
7	5	3.90	4	13.05
8	3	3.93	4	3.50
9	-	-	3	3.50
10	-	-	6	3.70
11	-	-	4	4.14
12	-	-	6	4.60

Notes: Iterations until convergence for k-medoids using Afghanistan voting districts for $N = 205$ and a space derived from spatially displaced copies of Afghanistan voting districts for $N = 820$ as described in the main text. The "best" column shows the number of iterations until numeric convergence for the initial centroids that achieve the lowest objective function value at convergence among the 100 sets of initial centroids. The "average" column shows the average number of iterations until numeric convergence across all 100 sets of initial centroids.

When added emphasis is helpful, X is written (X, d). The data $\mathscr{D}$ is distributed according to an unknown (joint) data generating process (DGP) – $\mathscr{D} \sim \mathrm{P}_0$. The object X will be the main object used to characterize any dependence in the data $\mathscr{D}$ over i.

Consider testing a scalar null hypothesis, $H_0 : \theta_0 = \theta^{\circledast}$, at level $\alpha \in (0, 1)$. Here $\theta^{\circledast}$ is a hypothesized value of a parameter that reflects the data generating process P_0. Our focus will be on the problem of testing a hypothesis about a coefficient in a linear regression model (e.g. testing $H_0 : \theta_0 = 0$ where θ_0 is a parameter in a linear regression, with $\zeta_i = (y_i, x_i)$ or $\zeta_i = (y_i, x_i, z_i)$ representing observations of an outcome variable Y, possibly endogenous regressor D, and exogenous instruments Z.) In the case of this example, P_0 will be allowed to be a DGP in which observations are correlated with each other. Failure to account for dependence in $\mathscr{D}$ across $i \in \mathsf{X}$ may lead to substantial size distortion when testing H_0.

Let $\mathcal{C} = \{\mathsf{C}_1, ..., \mathsf{C}_k\}$ be a partition of X of cardinality $k \geqslant 2$. The elements $\mathsf{C}_1, ..., \mathsf{C}_k$ are referred to as clusters. A *cluster-based* inferential procedure for testing H_0 is a (possibly

random) assignment

$$\texttt{Test} : (\mathscr{D}, \mathcal{C}) \mapsto T \in \{\text{Fail to Reject, Reject}\}. \tag{B.1}$$

Here, the decision rule itself is called `Test` and will generally depend on the level $\alpha \in (0,1)$ of the test. The outcome of the test `Test` is referred to as T; i.e. $T = \texttt{Test}(\mathscr{D}, \mathcal{C})$. The set containing the pair $(\mathscr{D}, \mathcal{C})$ remains unnamed to avoid additional notation.

We consider the following three cluster-based inferential procedures for testing a scalar hypothesis in this paper: the procedure of Ibragimov and Müller (2010) (IM), the procedure of Canay et al. (2017) (CRS), and inference based on the cluster covariance estimator as described in Bester et al. (2011) (CCE).[2] For clarity, consider 't-statistic' based testing of the hypothesis $H_0 : \theta_0 = \theta^{\circledast}$. Let $\mathsf{C} \subseteq \mathsf{X}$ and $\widehat{\theta}_{\mathsf{C}}$ be an estimator of θ_0 using only data corresponding to observations in C. Now define $S \in \mathbb{R}^{\mathcal{C}}$ (note: $\mathbb{R}^{\mathcal{C}}$ denotes functions $\mathcal{C} \to \mathbb{R}$) and the t-statistic function $t : \mathbb{R}^{\mathcal{C}} \to \mathbb{R}$ such that

$$S = (S_{\mathsf{C}})_{\mathsf{C} \in \mathcal{C}}, \quad S_{\mathsf{C}} = (n/k)^{1/2}(\widehat{\theta}_{\mathsf{C}} - \theta^{\circledast}), \tag{B.2}$$

$$t(S) = \frac{k^{-1/2} \sum_{\mathsf{C} \in \mathcal{C}} S_{\mathsf{C}}}{\sqrt{(k-1)^{-1} \sum_{\mathsf{C} \in \mathcal{C}} \left(S_{\mathsf{C}} - k^{-1} \sum_{\mathsf{C}' \in \mathcal{C}} S_{\mathsf{C}'}\right)^2}}. \tag{B.3}$$

For a specified level $a \in (0,1)$ (which may differ from α), there are IM, CRS, and CCE tests, denoted by $\texttt{Test}_{\mathrm{IM}(a)}$, $\texttt{Test}_{\mathrm{CRS}(a)}$, $\texttt{Test}_{\mathrm{CCE}(a)}$. These tests are defined by their outcomes given data $\mathscr{D}$ and a partition $\mathcal{C}$:

$$T_{\mathrm{IM}(a)} = \text{Reject} \quad \text{if} \quad |t(S)| > t_{1-a/2,k-1}, \tag{B.4}$$

$$T_{\mathrm{CRS}(a)} = \text{Reject} \quad \text{if} \quad |t(S)| > \text{quantile}_{1-a}(\{|t(hS)|\}_{h \in \mathcal{H}_{\mathcal{C}}}), \tag{B.5}$$

$$T_{\mathrm{CCE}(a)} = \text{Reject} \quad \text{if} \quad \left| \frac{\widehat{\theta}_{\mathsf{X}} - \theta^{\circledast}}{\widehat{V}^{1/2}_{\mathrm{CCE},\mathcal{C}}} \right| > \sqrt{\frac{k}{k-1}} \times t_{1-a/2,k-1}, \tag{B.6}$$

where $t_{1-a/2,k-1}$ is the $(1-a/2)$-quantile of a t-distribution with $k-1$ degrees of freedom; the set $\{hS\}_{h \in \mathcal{H}_{\mathcal{C}}}$ is the orbit of the action of $\{\pm 1\}^{\mathcal{C}}$ on S, so that for each h, $hS \in \mathbb{R}^{\mathcal{C}}$ has C^{th} component $\pm(n/k)^{1/2}(\widehat{\theta}_{\mathsf{C}} - \theta^{\circledast})$ for some sign in $\{\pm 1\}$; and $\widehat{V}_{\mathrm{CCE},\mathcal{C}}$ is the standard cluster covariance matrix estimator. $\texttt{Test}_{\bullet(a)}$ and $T_{\bullet(a)}$ are used when the choice of IM, CRS, or CCE is unspecified. When we wish to be clear about explicit dependence on $\mathscr{D}$ and $\mathcal{C}$ we use the more cumbersome notation $T_{\bullet(a),\mathcal{C}} = \texttt{Test}_{\bullet(a)}(\mathscr{D}, \mathcal{C})$. With prespecified,

[2] Extension of formal results for testing joint hypotheses using CRS is straightforward.

non-data-dependent clusters, each of the IM, CRS, and CCE procedures has the favorable property of asymptotic nominal size control under respective regularity conditions whenever dependence in observations ζ_i, ζ_j with i, j in different clusters is suitably negligible.

The second important definition is that of an *unsupervised clustering algorithm*, which is an assignment that returns, to every $\mathsf{X} = (\mathsf{X}, \mathrm{d})$, a partition of X given by the mapping

$$\texttt{Cluster} : \mathsf{X} \mapsto \mathcal{C}. \tag{B.7}$$

The idea behind using an unsupervised clustering algorithm is that if the dissimilarity d appropriately reflects the dependence in ζ_i, then the resulting partition $\mathcal{C}$ may have the desired property that averages of observations belonging to different clusters exhibit negligible dependence. In the formal analysis in Section **??**, the imposed mixing conditions imply that dependence between ζ_i and ζ_j vanishes as $\mathrm{d}(i, j)$ becomes large. Then, if $\mathcal{C}$ places distant observations (as defined by d) in different clusters, favorable properties of the test T may be anticipated.

Though there are many commonly used unsupervised clustering algorithms and we expect most to be usable as methods for forming data-dependent clusters, we consider only k-`medoids` in this paper for technical reasons discussed in Section **??**. Note that by composition of specific `Test` and `Cluster` procedures, it is already possible to define an outcome T for H_0 given $\mathscr{D}, \mathsf{X}$ by constructing $T = \texttt{Test}(\mathscr{D}, \texttt{Cluster}(\mathsf{X}))$.

The k-`medoids` algorithm we use for establishing our results is as follows. For finite (X, d) and medoids $\mathscr{I} \subseteq \mathsf{X}$ define $\mathrm{cost}(\mathscr{I}) = \sum_{j \in \mathsf{X}} \min_{i \in \mathscr{I}} \mathrm{d}(i, j)^2$.

k-`medoids`: *Input.* (X, d), k.

Initialize a set of medoids $\mathscr{I} \subseteq \mathsf{X}$ with cardinality $|\mathscr{I}| = k$.

While $\mathrm{cost}((\mathscr{I} \cup \{j\}) \setminus \{i\}) < \mathrm{cost}(\mathscr{I})$ for some $i \in \mathscr{I}$ and $j \in \mathsf{X} \setminus \mathscr{I}$,

Replace $\mathscr{I}$ with $\mathscr{I} \cup \{\widehat{j}\} \setminus \{\widehat{i}\}$ where $(\widehat{i}, \widehat{j}) \in \arg\min_{(i,j) \in \mathscr{I} \times (\mathsf{X} \setminus \mathscr{I})} \mathrm{cost}((\mathscr{I} \cup \{j\}) \setminus \{i\})$.

Output. $\mathcal{C} = \{\mathsf{C}_i\}_{i \in \mathscr{I}}$ with $j \in \mathsf{C}_i$ if $\mathrm{d}(i, j) = \min(\{\mathrm{d}(i', j)\}_{i' \in \mathscr{I}})$.

This implementation has run time $O(k(n-k)^2)$ per iteration. Additional details about computational run times in our application and simulations are presented in supplemental material.

The final layer to our proposed testing procedure is a method for data-dependent

choice of the cluster-based inferential procedure by considering a collection of candidate testing and clustering procedures of the form `Test` and `Cluster`. We propose making this choice on the basis of simultaneously controlling Type-I and Type-II error rates. Let $\text{Err}_{\text{Type-I}}(\texttt{Test}, \texttt{Cluster})$ denote type-I error for the testing outcome defined by $\texttt{Test}(\mathscr{D}, \texttt{Cluster}(\mathsf{X}))$. Next, consider a set of alternatives $\boldsymbol{\Theta}_{\text{alt}}$. Let $\text{Err}_{\text{Type-II}}(\texttt{Test}, \texttt{Cluster})$ denote a weighted average type-II error against the alternatives in $\boldsymbol{\Theta}_{\text{alt}}$. The choice of the alternative set and weighting function will be application specific and should depend on details of the problem. In the empirical illustration in Section **??**, we chose simple average power over an equally spaced grid of values that we believe encompass all remotely plausible values for the parameter of interest. We believe this practice provides a simple default. Because $\text{Err}_{\text{Type-I}}(\texttt{Test}, \texttt{Cluster})$ and $\text{Err}_{\text{Type-II}}(\texttt{Test}, \texttt{Cluster})$ will typically not be known, we consider a setting in which estimates $\widehat{\text{Err}}_{\text{Type-I}}(\texttt{Test}, \texttt{Cluster})$ and $\widehat{\text{Err}}_{\text{Type-II}}(\texttt{Test}, \texttt{Cluster})$ are available.

To finish the final layer, let $\mathscr{T}$ be a collection of pairs of the form $(\texttt{Test}, \texttt{Cluster})$. Note that the components `Test` in $\mathscr{T}$ are assumed to control Type I error asymptotically given suitable partitions of the data. This assumption is formalized by Condition 6 in Section **??**. We then choose $(\widehat{\texttt{Test}}, \widehat{\texttt{Cluster}}) \in \mathscr{T}$ by solving

$$\begin{aligned} &(\widehat{\texttt{Test}}, \widehat{\texttt{Cluster}}) \in \arg\min \ \widehat{\text{Err}}_{\text{Type-II}}(\texttt{Test}, \texttt{Cluster}) \\ &\text{s.t. } (\texttt{Test}, \texttt{Cluster}) \in \mathscr{T} \ , \ \widehat{\text{Err}}_{\text{Type-I}}(\texttt{Test}, \texttt{Cluster}) \leqslant \alpha. \end{aligned} \tag{B.8}$$

The final testing outcome for H_0 is then denoted

$$\widehat{T} = \widehat{\texttt{Test}}(\mathscr{D}, \widehat{\texttt{Cluster}}(\mathsf{X})). \tag{B.9}$$

In this paper, interest will be in $\mathscr{T}$ of the form

$$\mathscr{T}_{\bullet(\alpha),k_{\max}} = \left\{ (\texttt{Test}_{\bullet(a)}, k\text{-}\texttt{medoids}) : a \in [0, \alpha], k \in \{2, ..., k_{\max}\} \right\} \tag{B.10}$$

where α is the nominal testing level for H_0, $k_{\max}$ is a researcher-chosen upper bound on the number of clusters, and $\texttt{Test}_{\bullet(a)}$ is defined above.[3] With $\mathscr{T} = \mathscr{T}_{\bullet(\alpha),k_{\max}}$, the "parameter space" in (B.10) depends on two independent parameters a and k. It follows that (B.8) is a two-dimensional optimization problem with a single constraint. The resulting optimization

[3] One could also include pairs involving pre-specified partitions in $\mathscr{T}$. We ignore this possibility for notational convenience.

problem is therefore non-degenerate. The solution is then determined by two parameters $\widehat{\alpha}$ and $\widehat{k}$. Furthermore, the testing outcome can be expressed as $\widehat{T} = \texttt{Test}_{\bullet(\widehat{\alpha})}(\mathscr{D}, \widehat{\mathcal{C}})$, with $\widehat{\mathcal{C}} = \mathcal{C}^{(\widehat{k})} = \widehat{k}\text{-}\texttt{medoids}(\mathsf{X})$. When tests are based on $\bullet$ being IM, CRS, or CCE and it is helpful to make the overall level α explicit, we write $\widehat{T} = \widehat{T}_{\bullet(\alpha)}$ for added emphasis.

$\widehat{k}$ provides a data-dependent answer to how many clusters to use. Optimization over a allows the parameter entering the data-dependent decision rule, $\widehat{\alpha}$, to be smaller than the nominal level of the test, α. That is, the data-dependent decision rule may be more conservative than would be implied by conventional, fixed rules that asymptotically control size but may fail to do so in finite samples. Finally, the constraint on a in the definition of $\mathscr{T}$ in (B.10) guarantees that inference based on $\widehat{T}$ will maintain asymptotic size control under conditions that do not require the estimator $\widehat{\text{Err}}_{\text{Type-I}}(\texttt{Test}, \texttt{Cluster})$ to agree with the true Type-I error rate in finite-samples or asymptotically.

In practice, estimates $\widehat{\text{Err}}_{\text{Type-I}}(\texttt{Test}, \texttt{Cluster})$ and $\widehat{\text{Err}}_{\text{Type-II}}(\texttt{Test}, \texttt{Cluster})$ of Type-I and Type-II error rates are needed. In all results reported in the following sections, we obtain these estimates from auxiliary estimation of the dependence structure in the data based on Gaussian Quasi Maximum Likelihood Estimation (QMLE) using a simple exponential covariance function. We then use the estimated dependence structure within a Gaussian model to obtain $\widehat{\text{Err}}_{\text{Type-I}}(\texttt{Test}, \texttt{Cluster})$ and $\widehat{\text{Err}}_{\text{Type-II}}(\texttt{Test}, \texttt{Cluster})$. Further details are provided in Appendix A. We note that in principle any baseline model could be used in implementing (B.8) and one could consider uniform size control over classes of models, $\mathcal{M}$, by replacing the size constraint in (B.8) with $\max_{m \in \mathcal{M}} \widehat{\text{Err}}_{\text{Type-I},m}(\texttt{Test}, \texttt{Cluster}) \leqslant \alpha$. We wish to reemphasize that our theoretical results do not require the consistency of the estimated dependence structure of P_0 in order to control size. As a result, misspecification in the model for dependence asymptotically leads only to potential loss of power.

(B.8) differs from most methods in the literature for choosing data-dependent tuning parameters for use in conducting inference with dependent data. Much of the existing literature suggests choosing a single tuning parameter to optimize a weighted combination of size distortion and power; see, for instance, Lazarus et al. (2021), Sun and Kim (2015), and references therein. Instead, our proposal leverages the fact that most commonly used inferential procedures for dependent data depend on two parameters - nominal size and a smoothing parameter - and focuses on maximizing power within procedures that control

size. Our proposal is closely related to Müller and Watson (2020) who consider an inference approach for spatially dependent data that makes use of a tuning parameter and a critical value which are chosen by minimizing confidence interval length subject to exactly controlling size. Relative to the existing literature, both Müller and Watson (2020) and our approach offer additional flexibility by explicitly considering two choice variables and make use of criteria, minimizing interval length or maximizing power subject to maintaining size control, that we believe will be appealing to many researchers.

For convenience of reference in the following sections, the above described procedure is stated under Algorithm 1 below. For concreteness, Algorithm 1 references the procedures specialized to IM, CRS, CCE in conjunction with k-`medoids`. A more general procedure (i.e. for arbitrary $\mathscr{T}$) could be stated analogously. To simplify notation at the cost of some abuse of notation, write $\widehat{\text{Err}}_{\text{Type}-\text{I}}(\texttt{Test}_{\bullet(a)}, k\text{-}\texttt{medoids}) = \widehat{\text{Err}}_{\text{Type}-\text{I}}(\bullet(a), k)$ and $\widehat{\text{Err}}_{\text{Type}-\text{II}}(\texttt{Test}_{\bullet(a)}, k\text{-}\texttt{medoids}) = \widehat{\text{Err}}_{\text{Type}-\text{II}}(\bullet(a), k)$. We employ Algorithm 1 in the empirical example and simulation study in Sections **??** and 3. The Appendix contains full implementation details in these examples.

Algorithm 1. (*Inference with Cluster Learning with* k-`medoids` *and* IM, CRS *or* CCE). Testing H_0 at level $0 < \alpha < 1$.

Data: $\mathscr{D}, \mathsf{X}$.

Inputs: $k_{\max}$; $\boldsymbol{\Theta}_{\text{alt}}$; $\bullet$ = IM, CRS, or CCE; Estimates $\widehat{\text{Err}}_{\text{Type-I}}(\bullet(a), k), \widehat{\text{Err}}_{\text{Type-II}}(\bullet(a), k)$

Procedure: Solve (B.8) to obtain $(\widehat{\alpha}, \widehat{k})$.

Output: Set $\widehat{T}_{\bullet(\alpha)} = \texttt{Test}_{\bullet(\widehat{\alpha})}(\mathscr{D}, \widehat{k}\text{-}\texttt{medoids}(\mathsf{X}))$.

Note, Algorithm 1 can be inverted into confidence sets. For $\{H_0 : \theta_0 = \theta^{\circledast} : \theta^{\circledast} \in \Theta\}$ a family of hypotheses, C.I. $= \{\theta^{\circledast} \in \Theta :$ Algorithm 1 at level α returns Fail to Reject for$H_0 : \theta_0 = \theta^{\circledast}\}$ calculates a $(1 - \alpha)$ confidence set.