Inference for Dependent Data with Learned Clusters

Jianfei Cao

Northeastern University
Department of Economics
360 Huntington Ave., 301 Lake Hall
Boston, MA 02115, USA
e-mail: j.cao@northeastern.edu

Christian Hansen

The University of Chicago
Booth School of Business
5807 S. Woodlawn, Chicago, IL 60637, USA
e-mail: chansen1@chicagobooth.edu

Damian Kozbur

University of Zürich
Department of Economics
Schönberggasse 1, 8001 Zürich, Switzerland
e-mail: damian.kozbur@econ.uzh.ch

Lucciano Villacorta

Central Bank of Chile
Agustinas 1180, Santiago, Chile
e-mail: lvillacorta@bcentral.cl

Abstract: This paper presents and analyzes an approach to cluster-based inference for dependent data. The primary setting considered here is with spatially indexed data in which the dependence structure of observed random variables is characterized by a known, observed dissimilarity measure over spatial indices. Observations are partitioned into clusters with the use of an unsupervised clustering algorithm applied to the dissimilarity measure. Once the partition into clusters is learned, a cluster-based inference procedure is applied to a statistical hypothesis testing procedure. The procedure proposed in the paper allows the number of clusters to depend on the data, which gives researchers a principled method for choosing an appropriate clustering level. The paper gives conditions under which the proposed procedure asymptotically attains correct size. A simulation study shows that the proposed procedure attains near nominal size in finite samples in a variety of statistical testing problems with dependent data.

Key Words: Unsupervised Learning, Cluster-based Inference, HAR inference.
JEL Codes: C1.

1. Introduction

Conducting accurate statistical inference with data featuring serial or cross-sectional dependence requires carefully accounting for the underlying dependence structure. A variety of methods are available for researchers analyzing dependent data. An important class of inferential methods is the class of *cluster-based* methods. Cluster-based methods work with

*First version: October 2018. This version is of August 2, 2021. Christian Hansen would like to thank the National Science Foundation as well as The University of Chicago Booth School of Business for financial support of this research. Damian Kozbur would like to thank The University of Zürich for financial support of this research.

a partition of observations into clusters. Inference proceeds by treating as negligible any covariance between observations that fall in different clusters, followed by performing appropriate tests for a statistical hypothesis of interest. Cluster-based methods are known to deliver asymptotically valid inference, in the sense of controlling size of tests and coverage of interval estimates, in a variety of dependent-data settings; see, for example, Ibragimov and Müller (2010), Bester et al. (2011), Canay et al. (2017), and Hansen and Lee (2019).

Though the term can apply more generally, in empirical economics settings "clustering" is most commonly used in the special case of an $n \times T$ panel data setting, in which individual level observations $i = 1, ..., n$ are clustered together across the T dimension. This setting has been analyzed extensively, and one comprehensive reference is the textbook Wooldridge (2010). In the case of a panel data setting, there is natural, "known" way to partition data into clusters, which is justified from an assumption on the data, namely that observations i and j are independent whenever $i \neq j$.

However, there is not an immediate "natural" set of partitions available in many settings. Rather, researchers often use data where observations can be thought of as being located in some space and where a measure of distance or dissimilarity between observations is available. For example, researchers routinely use data where observations are indexed geographically over an irregular geographic region and spatial correlation is a concern. As another example, researchers may be working with data where observations are equipped with a notion of economic distance. Examples of such applications include Conley and Dupor (2003), which constructs a notion of economic distance based on input-output tables, and Conley and Topa (2002), which constructs a notion of economic distance based on demographic composition of zip codes. In such settings, researchers may wish to use cluster-based inference as a robust method to account for dependence in the data but desire some guidance about exactly which observations to cluster together and how many clusters to use.

In this paper, we consider settings in which natural partitions may not exist but a distance or dissimilarity measure that is informative about the underlying correlation structure is available. We assume that dependence between observable quantities is suitably small when the distance between them is suitably large. A main goal is then to verify that cluster-based methods remain asymptotically valid when clusters are not pre-specified but are constructed in a data-dependent fashion using the underlying notion of dissimilarity and 'unsupervised learning' methods. There are many such tools available for constructing groups from measures of dissimilarity including k-`medoids`, k-`means`, and `hierarchical clustering`; see Hastie et al. (2009) for a general review.

At a high level, the data-driven cluster-based inference approach we propose for testing a hypothesis H_0 proceeds in three main steps. In the first step, unsupervised learning methods are applied using the observed dissimilarity between observations to produce a collection of partitions of the observations into clusters, $\mathscr{C} = \{\mathcal{C}^{(k)}\}_{k=2}^{k_{\max}}$ for some integer $k_{\max}$. Each element $\mathcal{C}^{(k)}$ provides a candidate structure for use with a cluster-based method and may be based on a different number of clusters. Cluster-based methods for testing H_0 will involve decision thresholds for determining whether to reject H_0. For example, we may decide to reject H_0 when an empirical p-value is smaller than some threshold. In the second step, we select a partition $\widehat{\mathcal{C}}$ from $\mathscr{C}$ and a rejection rule for H_0 chosen to maximize exact weighted average power against a collection of alternatives subject to maintaining exact size control within completely specified data generating processes (DGPs) for the observed data. These DGPs may be data dependent; for example, they may involve parameters whose values are estimated in the data. While $\widehat{\mathcal{C}}$ and the rejection rule are selected based on size and power calculations in pre-specified DGPs, the selection is done in such a way that asymptotic size control is formally maintained regardless of whether these DGPs are correctly specified. In the final step, inference in the actual data is performed using the rejection rule and $\widehat{\mathcal{C}}$ determined in the second step. As $\mathscr{C}$ will typically contain candidate partitions with different numbers of groups, the procedure outlined above encompasses a data-driven method for choosing the appropriate level at which to cluster the data as well as a data-driven way to determine to which cluster each observation should be assigned.

The main theoretical results in this paper provide conditions under which inference based on the general procedure outlined in either Ibragimov and Müller (2010) or Canay et al. (2017) remains valid when data-dependent clusters produced by applying k-`medoids` are used in lieu of relying on a fixed number of pre-specified clusters. The regularity conditions involve moment and mixing restrictions, weak homogeneity assumptions on second moments of regressors and unobservables across groups, and restrictions on group boundaries. These moment and mixing conditions are implied by routine assumptions necessary for use of central limit approximations and the required homogeneity is less restrictive than covariance stationarity. The assumptions allow for relatively general heterogeneous dependent processes and are akin to conditions routinely imposed in other heteroskedasticity and autocorrelation robust (HAR) inference approaches; see, e.g. Andrews (1991) and Kelejian and Prucha (2007).

In the course of providing the main inferential results, we also develop novel insights about the finite-sample behavior of the k-`medoids` clustering algorithm that are important for cluster-based inference. Specifically, we verify that the partitions produced by k-`medoids` satisfy suitably defined balance and small boundary conditions without requiring a notion of consistency to a "true" partition. Establishing these properties for the groups produced by k-`medoids` is important for verifying their use as an input into cluster-based inference as existing results for such procedures in the setting where clusters are pre-specified rely on conditions that impose that clusters have small boundaries and have sizes that do not diverge relative to each other. We also note that these results on the behavior of k-`medoids` clustering are, to our knowledge, new to the unsupervised learning literature and may be of some independent interest.

As cluster-based methods provide a general approach to performing robust inference in the presence of dependent and heteroskedastic data, they belong to the general class of HAR inference procedures; see, e.g., Lazarus et al. (2018) for a recent review centered on time series analysis and Conley (1999) and Kelejian and Prucha (2007) for seminal references in the spatial context. Our results cover scenarios where the number of clusters used for inference is bounded as in Bester et al. (2011), Ibragimov and Müller (2010), and Canay et al. (2017) and thus are in the spirit of the "fixed-b" approach - e.g. Kiefer et al. (2000), Kiefer and Vogelsang (2002, 2005), Bester et al. (2016) - and the projection approach with fixed number of projections - e.g. Phillips (2005), Müller (2007), Sun (2013) - both of which are well-known to lead to improved size control relative to more traditional approaches under weak dependence. Within the general context of HAR inference, the choice of cluster structure is analogous to the choice of smoothing parameters in other HAR procedures, and our work thus complements the broad literature on data-driven tuning parameter choice in HAR inference. Our proposal in this paper is closely related to Sun and Kim (2015), Lazarus et al. (2018), Lazarus et al. (2021), and especially Müller and Watson (2020). Sun and Kim (2015) proposes a method to select the number of projections to use for HAR inference in a spatially dependent context by minimizing asymptotic mean-square error (MSE) of the variance estimator where terms that enter the expression for the asymptotic MSE are estimated using a parametric model. Lazarus et al. (2021) provide the size-power frontier based on the "fixed-b" approach in a time series context, and Lazarus et al. (2018) explicitly considers tuning parameter choice by trading off size and power along this frontier. Müller and Watson (2020) considers a novel standard error estimator for use with spatially dependent data based on spatial principal components and proposes a method to choose both the number of components to use in the method and the critical value to use in hypothesis testing based on minimizing confidence interval length subject to exactly controlling size within benchmark parametric models. Their procedure exactly controls size in the baseline model and can be constructed to asymptotically control size under general conditions. Relative to these approaches, we focus on data-dependent construction of clusters for use with cluster-based inference methods. Similar to Müller and Watson (2020), we choose both a partition and rejection rule for a hypothesis test to maximize weighted average power subject to exact size control within a parametric model in such a way that asymptotic size control is maintained. In a related paper, Bai et al. (2020) propose a new standard error estimator for panel data models robust to serial and spatial dependence when clusters are unknown, and dependence is weak. In their approach the time series dependence is controlled using a Newey-West HAC estimator, while the spatial dependence is controlled by assuming sparsity in the cross-section where only a small subgroup of (i, j) pairs have spatial dependence different from zero. The (i, j) pairs of this subgroup are unknown and are estimated by applying the thresholding approach in Bickel and Levina (2008). A similar procedure is proposed by Cai (2021).

The remainder of this paper is organized as follows. Section 2 presents a high-level overview of our proposed inferential procedure. We illustrate a concrete use of the method by applying it to the study of effect of insurgent attacks on voter turnout from Condra et al. (2018) in Section 3. Section 4 then illustrates the performance of our proposal in simulated data. Formal results are provided in Sections 5. Finally, the Appendix contains a detailed discussion of the implementation we use in the empirical and simulation examples. All proofs are contained in additional supplemental material.

2. 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. When added emphasis is helpful, X is written (X, d). The data $\mathscr{D}$ is distributed according to an unknown joint probability distribution $\mathscr{D} \sim \mathrm{P}_0$. The object X will be the main object used to characterize any dependence in the data $\mathscr{D}$ over i. Finally, P_0 is an element of a larger known class of distributions $\mathrm{P}_0 \in \mathbf{P}$.

Consider testing a null hypothesis, H_0, at level $\alpha \in (0,1)$. As a concrete example, we will focus on the problem of testing a hypothesis about a coefficient in a linear regression model (i.e. testing $H_0 : \theta_0 = 0$ where θ_0 is a parameter in a linear regression). In the case of this example, $\mathbf{P}$ consists of all possible data generating processes (DGPs) for the regression data, which typically include DGPs 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. In the following, we outline an approach for obtaining valid inference within this setting using cluster-based inferential procedures with clusters generated by an unsupervised clustering algorithm.

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

$$\texttt{Test} : (\mathscr{D}, \mathcal{C}) \mapsto T \in \big\{\text{Fail to Reject, Reject}\big\}. \tag{2.1}$$

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, ie. $T = \texttt{Test}(\mathscr{D}, \mathcal{C})$. The set containing the pair $(\mathscr{D}, \mathcal{C})$ remains unnamed to avoid additional notation.

We focus on 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).[1] For clarity, consider 't-statistic' based testing of the hypothesis $H_0 : \theta_0 = \theta^{\circledast}$ where θ_0 is some scalar parameter of interest associated to the data generating process P_0 and $\theta^{\circledast}$ is a hypothesized possible value for θ_0. 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 $S \in \mathbb{R}^{\mathcal{C}}$ and the t-statistic function $t : \mathbb{R}^{\mathcal{C}} \to \mathbb{R}$ such that

$$S = (S_{\mathsf{C}})_{\mathsf{C} \in \mathcal{C}}, \quad S_{\mathsf{C}} = (n/k)^{1/2}(\widehat{\theta}_{\mathsf{C}} - \theta^{\circledast}), \tag{2.2}$$

$$t(S) = \frac{k^{-1/2} \sum_{\mathsf{C} \in \mathcal{C}} S_{\mathsf{C}}}{\sqrt{(k-1)^{-1} \sum_{\mathsf{C} \in \mathcal{C}} \left(S_{\mathsf{C}} - k^{-1} \sum_{\mathsf{C}' \in \mathcal{C}} S_{\mathsf{C}'}\right)^2}}. \tag{2.3}$$

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

For a specified level $a \in (0,1)$ (which may distinct from α), there are the 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 $\mathcal{C}$:

$$T_{\mathrm{IM}(a),\mathcal{C}} = \text{Reject} \quad \text{if} \quad |t(S)| > t_{1-a/2,k-1}, \tag{2.4}$$

$$T_{\mathrm{CRS}(a),\mathcal{C}} = \text{Reject} \quad \text{if} \quad |t(S)| > \text{quantile}_{1-a}(\{|t(hS)|\}_{h\in\mathcal{H}_\mathcal{C}}), \tag{2.5}$$

$$T_{\mathrm{CCE}(a),\mathcal{C}} = \text{Reject} \quad \text{if} \quad \left|\frac{\widehat{\theta}_{\mathsf{X}} - \theta^{\circledast}}{\widehat{V}^{1/2}_{\mathrm{CCE},\mathcal{C}}}\right| > \sqrt{\frac{k}{k-1}} \times t_{1-a/2,k-1}, \tag{2.6}$$

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 $\{\pm1\}^{\mathcal{C}}$ 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 $\{\pm1\}$; and $\widehat{V}_{\mathrm{CCE},\mathcal{C}}$ is the standard cluster covariance matrix estimator. Similarly $\texttt{Test}_{\bullet(a)}$ and $T_{\bullet(a),\mathcal{C}}$ is used when the choice of IM, CRS, or CCE is unspecified. Note that explicit dependence on $\mathscr{D}$ is supressed in $T_{\bullet(a),\mathcal{C}}$ but that $T_{\bullet(a),\mathcal{C}}$ may be formally related back to $\mathscr{D}$ by $T_{\bullet(a),\mathcal{C}} = \texttt{Test}_{\bullet(a)}(\mathscr{D},\mathcal{C})$. With prespecified, 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 appropriate formal definition of negligible is method-specific.

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

$$\texttt{Cluster} : \mathsf{X} \mapsto \mathcal{C}. \tag{2.7}$$

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 observations belonging to different clusters exhibit negligible dependence. In the formal analysis in Section 5, 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 5.4. We provide a full description of the version of k-`medoids` we use in Appendix A. 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 final layer to our proposed testing procedure is a simple method for data-dependent choice of both 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 $\mathrm{Err}_{\text{Type-I}}(\texttt{Test}, \texttt{Cluster})$ denote type-I error for the testing outcome defined by $\texttt{Test}(\mathscr{D}, \texttt{Cluster}(\mathsf{X}))$. For a set of alternatives $\mathbf{P}_{\text{alt}}$, let $\text{Err}_{\text{Type-II},\mathbf{P}_{\text{alt}}}(\texttt{Test},\texttt{Cluster})$ denote a weighted average type-II error. Because $\text{Err}_{\text{Type-I}}(\texttt{Test},\texttt{Cluster})$ and $\text{Err}_{\text{Type-II},\mathbf{P}_{\text{alt}}}(\texttt{Test},\texttt{Cluster})$ will typically not be known, we consider a setting in which estimates $\widehat{\text{Err}}_{\text{Type-I}}(\texttt{Test},\texttt{Cluster})$ and $\widehat{\text{Err}}_{\text{Type-II},\mathbf{P}_{\text{alt}}}(\texttt{Test},\texttt{Cluster})$ are available.

To finish the final layer, let $\mathscr{T}$ be a collection of pairs of the form $(\texttt{Test},\texttt{Cluster})$. In this paper, interest will primarily be in $\mathscr{T}$ of the form

$$\mathscr{T}_{\bullet(\alpha),k_{\max}} = \Big\{(\texttt{Test}_{\bullet(a)}, k\text{-}\texttt{medoids}) : a \in [0,\alpha], k \in \{2,...,k_{\max}\}\Big\} \tag{2.8}$$

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.[2] We then choose $(\widehat{\texttt{Test}},\widehat{\texttt{Cluster}}) \in \mathscr{T}$ by solving

$$\begin{aligned}(\widehat{\texttt{Test}},\widehat{\texttt{Cluster}}) &\in \arg\min\ \widehat{\text{Err}}_{\text{Type-II},\mathbf{P}_{\text{alt}}}(\texttt{Test},\texttt{Cluster}) \\ \text{s.t. } (\texttt{Test},\texttt{Cluster}) &\in \mathscr{T}\ ,\ \widehat{\text{Err}}_{\text{Type-I}}(\texttt{Test},\texttt{Cluster}) \leqslant \alpha.\end{aligned} \tag{2.9}$$

Set the final testing outcome for H_0 to

$$\widehat{T} = \widehat{\texttt{Test}}(\mathscr{D},\widehat{\texttt{Cluster}}(\mathsf{X})). \tag{2.10}$$

In the case of interest, $\mathscr{T} = \mathscr{T}_{\bullet(\alpha),k_{\max}}$, the "parameter space" in (2.8) depends on two independent parameters a and k. It follows that (2.9) is a two-dimensional optimization problem with a single constraint. The resulting optimization 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 $\widehat{T} = \texttt{Test}_{\bullet(\widehat{\alpha})}(\mathscr{D},\widehat{\mathcal{C}})$, 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 when it is helpful to make the overall level α explicit, write also $\widehat{T} = \widehat{T}_{\bullet(\alpha)}$ 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 (2.8) guarantees that inference based on $\widehat{T}$ will maintain asymptotic size control under sensible 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 $\widehat{\text{Err}}_{\text{Type-II},\mathbf{P}_{\text{alt}}}(\texttt{Test},\texttt{Cluster})$ of Type-I and Type-II error rates are needed. In all results reported in the following sections, we obtain these estimates by forming preliminary estimates of the dependence structure for 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 $\widehat{\text{Err}}_{\text{Type-II},\mathbf{P}_{\text{alt}}}(\texttt{Test},\texttt{Cluster})$. Further details are provided in Appendix A.2. We note that

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

in principle any baseline model could be used in implementing (2.9) and one could consider uniform size control over classes of models, $\mathcal{M}$, by replacing the size constraint in (2.9) with $\max_{m\in\mathcal{M}} \widehat{\mathrm{Err}}_{\text{Type-I},m}(\mathtt{Test},\mathtt{Cluster}) \leqslant \alpha$. 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.

(2.9) differs from several recent 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` (ie for $\mathscr{T} = \mathscr{T}_{\bullet(\alpha),k_{\max}}$). To simplify notation at the cost of some abuse of notation, write $\widehat{\mathrm{Err}}_{\mathrm{Type-I}}(\mathtt{Test}_{\bullet(a)}, k\text{-}\mathtt{medoids}) = \widehat{\mathrm{Err}}_{\mathrm{Type-I}}(\bullet(a), k)$ and $\widehat{\mathrm{Err}}_{\mathrm{Type-II},\mathbf{P}_{\mathrm{alt}}}(\mathtt{Test}_{\bullet(a)}, k\text{-}\mathtt{medoids}) = \widehat{\mathrm{Err}}_{\mathrm{Type-II},\mathbf{P}_{\mathrm{alt}}}(\bullet(a), k)$. The concrete description in Algorithm 1 is the only version of the procedure described above implemented in the empirical study and simulation study that follow. The more general procedure (i.e. for arbitrary $\mathscr{T}$) could be stated analogously.

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

Procedure: Solve (2.9) to obtain $(\widehat{\alpha}, \widehat{k})$.

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

3. Empirical Application: Insurgent Electoral Violence

In this section, we illustrate the inferential method proposed in Section 2 by reexamining estimates of the effect of insurgent attacks on voter from Condra et al. (2018). We use the same data as Condra et al. (2018) which consist of district-level observations for 205 voting districts in Afghanistan in 2014. Each observation contains information on direct morning attacks, voter turnout, and district level control variables for two separate election rounds.

3.1. Model and Spatial Dependence

We focus on column (4) of Table 2 in Condra et al. (2018) which reports the estimated effect of nearby morning attacks on voter turnout obtained from instrumental variables (IV) estimation of the linear model

$$Y_{de} = \alpha_0 + \theta_0 Attacks_{de} + W_{de}'\gamma_0 + U_{de} \tag{3.1}$$

$$Attacks_{de} = \mu_0 + \pi_0 Z_{de} + W_{de}'\xi_0 + V_{de} \tag{3.2}$$

where (3.1) is the structural equation with parameter of interest θ_0 and (3.2) is the first stage representation. Here, Y_{de} is the turnout in district d and election round e, and $Attacks_{de}$ is the number of morning attacks in district d and election round e. U_{de} and V_{de} are unobservables where U_{de} and $Attacks_{de}$ are potentially correlated. Control variables W_{de} include election round dummy variables, voting hour wind conditions, population, measures of precipitation and ambient temperature, and the average of predawn and morning wind conditions during the pre-election period. The excluded instrument, Z_{de}, is a measure of early morning wind conditions.

The analysis in Condra et al. (2018) reports inference for the parameter θ_0 based on CCE with clusters defined by the cross-sectional unit of observation, the district. For the sake of reference, from this point, the inference and corresponding method from Condra et al. (2018) will be named "UNIT". This approach will be asymptotically valid in the presence of dependence within district and heteroskedasticity as long as associations across districts are ignorable. To provide some evidence on the plausibility of this assumption, we provide Moran (1950) $\mathcal{I}$ tests; see also Kelejian and Prucha (2001). To construct the tests, we first estimate model (3.1) to obtain estimates of U_{de} which are then used to construct the observation specific score for $\widehat{\theta}$, the IV estimator of θ_0 based on (3.1)-(3.2). The scores are then used as the data input into the Moran $\mathcal{I}$ test. We consider two formulations for the weight matrix used in the Moran $\mathcal{I}$ test. In the first, we create the weight matrix by assigning a value of one for the weight corresponding to the two nearest neighbors for each observation on the basis of Euclidean distance in the latitude and longitude coordinates of the centroid of each district and set all other weights to zero. In the second, we create the weight matrix by assigning a value of one to the weight corresponding to observations from the same district and set all other weights to zero. Note that the first setup should have power against alternatives where dependence is related to geographic distance and the second against alternatives where dependence is intertemporal.

In Table 1, we report the value of the Moran $\mathcal{I}$ test statistic and the associated p-value for testing the null-hypothesis of zero correlation across observations based on the weight matrix formed from geographic distance and on the weight matrix for intertemporal dependence. For spatial weights, we consider each election round separately as well as report results

TABLE 1
Moran (1950) $\mathcal{I}$ Tests based on IV Scores

	Moran $\mathcal{I}$	p-value
Spatial weights - First election round	4.338	< 0.001
Spatial weights - Second election round	2.546	0.011
Spatial weights - Both election rounds	5.387	< 0.001
Intertemporal weights	2.755	0.006

Notes: This table provides Moran (1950) $\mathcal{I}$ test results for the Condra et al. (2018) example. The test statistic value and associated p-value are respectively reported in the columns labeled "Moran $\mathcal{I}$" and "p-value." The first three rows report tests with weights for the Moran test constructed on the basis of geographic distance as described in the text. The first row uses only observations from the first election round, the second row uses only observations from the second election round, and the third row uses all observations to construct the test statistic. The final row uses a weight matrix constructed solely on the basis of intertemporal distance as described in the text.

pooling both election rounds. In each case, the calculated test statistic is large, and the p-value for testing the null hypothesis of zero correlation is small. The results are highly suggestive of the presence of both cross-sectional dependence associated with geographic distance and intertemporal dependence. As there is no immediately obvious unique way to group districts on the basis of geographic proximity, it is potentially interesting to construct clusters in a data-dependent fashion.

3.2. Data-Dependent Cluster-Based Inference

We now apply the approach outlined in Section 2 to construct data-dependent clusters of observations and produce inferential statements robust to the presence of spatial and temporal correlation and heteroskedasticity. A key input into this process is a dissimilarity measure d capturing the "distance" between each pair of distinct districts, d, d'. For this application, we use dissimilarity defined by Euclidean distance in geography

$$\mathrm{d}(d, d') = \|L_d - L_{d'}\|_2 = \sqrt{(\mathrm{lat}_d - \mathrm{lat}_{d'})^2 + (\mathrm{long}_d - \mathrm{long}_{d'})^2} \tag{3.3}$$

where $L_d = (\mathrm{lat}_d, \mathrm{long}_{d'})$ consists of the latitude and longitude coordinates of the centroid of the district and $\|\cdot\|_2$ denotes Euclidean distance.

Because of the nonuniform locations of district centers (i.e. districts are not on a rectangular grid), it is unclear ex ante how to partition them or how many clusters to use. As one potential resolution to this ambiguity, we apply k-`medoids` to the data using dissimilarity d. Note that (3.3) assigns dissimilarity zero to observations from the same district which means that these observations will always be assigned to the same cluster. We thus allow general dependence across election rounds within district as well as capture dependence across election rounds and across districts that belong to the same cluster. Specifically, for each $k = 2, ..., k_{\max}$ where $k_{\max}$ is a user-specified upper bound on the number of clusters

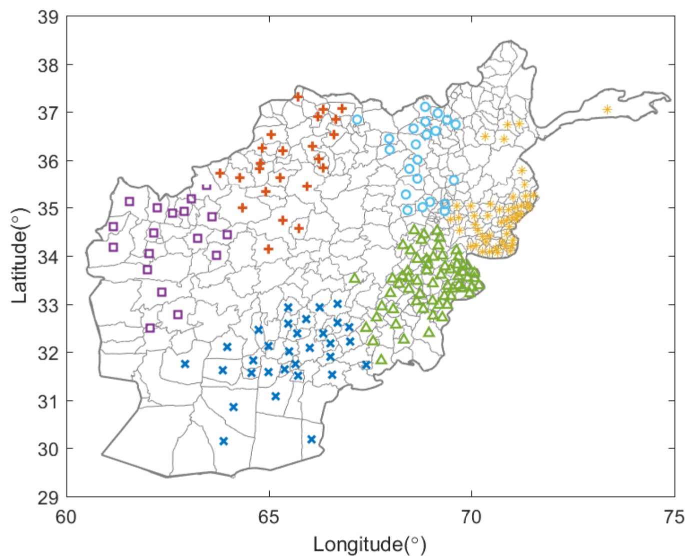


Fig 1: Display of partition of 205 districts in Afghanistan by k-`medoids` using final number of clusters given by $k = 6$. Distances are Euclidean distances based on latitude and longitude coordinates recorded at district centroids. Different marks correspond to different clusters in the partition. Marks are plotted at district centroids. Districts without symbols have insufficient data and were excluded in the analysis Condra et al. (2018) as well as in the current analysis.

to be considered, we apply k-`medoids`, yielding a set of partitions $\mathscr{C} = \{\mathcal{C}^{(2)}, ..., \mathcal{C}^{(k_{\max})}\}$. As an illustration, Figure 1 displays a partition of the data resulting from applying k-`medoids` with $k = 6$ with districts rendered with common markers belonging to common clusters $\mathsf{C} \in \mathcal{C}^{(6)}$.[3] In this empirical illustration, we set $k_{\max} = 8$.

In practice, we recommend choosing a small value for $k_{\max}$.This recommendation is based on two considerations. First, IM and CRS, our preferred procedures, rely on within-cluster estimates and the number of observations per cluster decreases quickly as the number of clusters considered increases, which results in unstable performance when large numbers of clusters are considered. Second, for dependence structures with non-trivial dependence, size distortions typically become pronounced with larger numbers of groups while improvements to power tend to rapidly diminish once a modest number of groups are considered.

With the set of potential partitions, $\mathscr{C}$, in hand, the final inputs we need to be able to select a partition and testing procedure using (2.9) are feasible guesses for Type I and average Type II error rates, $\widehat{\text{Err}}_{\text{Type-I}}(\bullet(a), k)$ and $\widehat{\text{Err}}_{\text{Type-II},\mathbf{P}_{\text{alt}}}(\bullet(a), k)$. Let U and V respectively denote the 410×1 vectors produced by stacking U_{de} and V_{de}. To obtain $\widehat{\text{Err}}_{\text{Type-I}}(\bullet(a), k)$

[3]We display the result with $k = 6$ as this is the solution obtained in (2.9) in this example for CRS.

TABLE 2
Impact of Early Morning Attacks on Voter Turnout during the 2014 Election

Cluster-based Inference								
	$\widehat{\theta}_0$	*s.e.*	*t*-stat	C.I.		C.I. (usual CV)		$\widehat{k}$
UNIT	-0.145	0.061	2.385	(-0.312,	0.022)	(-0.265,	-0.025)	205
CCE	-0.145	0.090	1.609	(-0.404,	0.114)	(-0.358,	0.068)	8
IM	-0.122	0.432	0.282	(-1.320,	1.077)	(-1.320,	1.077)	5
CRS	-0.242			(-1.497,	0.084)	(-1.497,	0.084)	6

Notes: This table presents inferential results based on selected clusters. Row labels indicate which procedure is used. The column labeled $\widehat{\theta}_0$ reports the IV estimate of θ_0 for the full sample in the rows labeled UNIT and CCE and the average of IV estimators of θ_0 of the five or six data-generated clusters in the rows labeled IM and CRS respectively. Column *s.e.* reports the estimated standard errors obtained in each procedure. Note that CRS does not rely on an explicit standard error estimate. Column *t*-stat reports the *t*-statistic for testing the null hypothesis that $\theta_0 = 0$ for each of the procedures. Column C.I reports confidence intervals for θ_0 obtained using $\widehat{\alpha}$ and $\widehat{k}$ from (2.9) for each procedure. Column C.I. (usual CV) reports confidence intervals of the IV estimate of θ_0 using fixed asymptotic thresholds. Column $\widehat{k}$ indicates the number of clusters selected in each procedure.

and $\widehat{\text{Err}}_{\text{Type-II},\mathbf{P}_{\text{alt}}}(\bullet(a), k)$, we use a parametric covariance model

$$\begin{aligned} \text{E}\begin{bmatrix} UU' & UV' \\ VU' & VV' \end{bmatrix} &= \begin{bmatrix} \Sigma_U(\tau^U) & \rho(\Sigma_U(\tau^U)^{1/2})(\Sigma_V(\tau^V)^{1/2})' \\ \rho(\Sigma_V(\tau^V))^{1/2}(\Sigma_U(\tau^U)^{1/2})' & \Sigma_V(\tau^V) \end{bmatrix} \\ &= \Sigma(\tau^U, \tau^V, \rho) \end{aligned} \tag{3.4}$$

with parameters (τ^U, τ^V, ρ) where $\Sigma_U(\tau)$ and $\Sigma_V(\tau)$ are covariance matrices with entries given by

$$[\Sigma_\varepsilon(\tau)]_{de,d'e'} = f_\tau((L_d, e), (L'_d, e')) = \exp(\tau_1)\exp(-\tau_2^{-1}\|L_d - L_{d'}\|_2 - \tau_3^{-1}|e - e'|) \tag{3.5}$$

for $\varepsilon \in \{U, V\}$. We obtain calibrated values for these parameters, $(\widehat{\tau}^U, \widehat{\tau}^V, \widehat{\rho})$ by forming initial estimates $\widehat{U}_{de}$ and $\widehat{V}_{de}$ of U_{de} and V_{de} from IV estimation of (3.1)-(3.2) and then using these residuals to estimate the covariance model via Gaussian QMLE. We can then approximate the exact Type I and average Type II error rates that would result from applying each inferential procedure with each potential partition within a Gaussian model with covariance matrix given by $\Sigma(\widehat{\tau}^U, \widehat{\tau}^V, \widehat{\rho})$ by simulation. Given these estimates of the Type I and Type II error rates, we then solve (2.9) to obtain the final partition and level for defining the cutoff in the decision rule used in the data. Further details are provided in Appendix A.2.

Table 2 reports inferential results for clustering at the district level as in Condra et al. (2018) (UNIT) and for the data-dependent CCE, IM, and CRS cluster-based inferential procedures. We report results for all procedures for completeness and illustration, but note that the theory and simulation results lead us to recommend that one uses IM or CRS in practice. For each method, we report the corresponding estimate of θ_0, the standard error, t-statistic, 95% confidence interval using $\widehat{k}$ and $\widehat{\alpha}$ from (2.9), 95% confidence interval using $\alpha = .05$ and $\widehat{k}$, and $\widehat{k}$.

In this illustration, all 95% level confidence intervals include 0 with the exception of the interval produced by district level clustering with the asymptotic cutoff. That is, all of the considered procedures which allow for spatial correlation would lead to failure to reject the null hypothesis that insurgent attacks have no effect on voter turnout at the 5% level within the posited linear IV structure. There is substantive variation in the interval estimates provided by the various procedures more generally though.

Both procedures based on clustered standard errors (UNIT and CCE), while producing intervals that cover zero when spatial correlation is allowed for, are relatively narrow when compared to IM and CRS. For both UNIT and CCE, we can see a mild discrepancy between the intervals produced under the asymptotic approximation and those produced by allowing a data-dependent decision threshold based on the parametric model. Specifically, within the parametric model underlying the simulation, controlling size at the 5% level would correspond to rejecting the null only when the p-value is smaller than 0.006 or 0.022 for UNIT and CCE respectively (using the fixed number of groups approximation of Bester et al. (2011).) These differences suggest that the asymptotic approximation is potentially failing and suggest some caution should be made in relying on these results; see also Ferman (2019). We do note that the interval provided in column C.I. will exactly control size under the specified parametric model. However, we also find that both CCE and individual level clustering are relatively sensitive to misspecification of the parametric model underlying construction of the estimated Type I and Type II error rates when compared to IM and CRS in the simulation results reported in Section 4.

Based on the theoretical and simulation results reported in the next sections, our preferred procedures are IM and CRS. These procedures are provably valid asymptotically under relatively weak conditions and are also relatively robust to misspecification of the parametric model underlying the size and power calculations in our simulation experiments. Looking to IM and CRS, we see similar performance, in the sense of relatively wide 95% interval estimates, across both, though IM is substantially wider than CRS in this example. For both of these procedures, the asymptotic cutoffs coincide with those produced by solving (2.9) using the simulated Type I and Type II error rates, which is reassuring though not necessary for validity. We note that there is a danger of a researcher trying both IM and CRS (and perhaps CCE) and reporting the most optimistic result. For this reason, our recommendation is to either default to, and compute only, IM or CRSor to report results from all procedures to provide complete information to readers.

4. Simulation

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 of Section 3. 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 Subection 4.1. We then present results in Subection 4.2.

4.1. 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.[4] 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 $e \in \{1, 2\}$.

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

$$Y_{de} = \theta_0 X_{de} + W_{de}'\gamma_0 + U_{de}, \tag{4.1}$$

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 $\{X_{de}, W_{de}\}_{d=1,...,205,e=1,2}$ generated as follows: $\forall\, l \in \{1, \ldots, 10\}$, $\forall\, m \neq l$, $\forall\, (d, e) \in \{1, \ldots, 205\} \times \{1, 2\}$, $\forall\, (d', e') \neq (d, e)$,

$$\begin{aligned}
&X_{de} \sim \text{N}(0, 1); W_{del} \sim \text{N}(0, 1),\\
&\text{corr}(X_{de}, X_{d'e'}) = \text{corr}(W_{del}, W_{d'e'l}) = f_\tau((L_d, e), (L_{d'}, e')),\\
&\text{corr}(X_{de}, W_{del}) = \text{corr}(W_{del}, W_{dem}) = 0.5,\\
&\text{corr}(X_{de}, W_{d'e'l}) = \text{corr}(W_{del}, W_{d'e'm}) = 0,
\end{aligned} \tag{4.2}$$

for $f_\tau((L_d, e), (L_{d'}, e'))$ defined in (3.5) 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).

$$U_{de} \sim \text{N}(0, 1); \quad \text{corr}(U_{de}, U_{d'e'}) = f_\tau((L_d, e), (L_{d'}, e')), \tag{4.3}$$

for $f_\tau((L_d, e), (L_{d'}, e'))$ defined in (3.5) with $\tau = (0, 3, 1)'$.

B. Spatial auto-regression (SAR).

$$\begin{aligned}
&U_{de} = 0.15 \textstyle\sum_{d' \neq d} U_{d'e} \mathbf{1}_{\{\|L_d - L_{d'}\|_2 < 0.3\}} + \varepsilon_{de},\\
&\varepsilon_{de} \sim \text{N}(0, 1); \quad \text{corr}(\varepsilon_{d1}, \varepsilon_{d2}) = \exp(-1),\ \text{corr}(\varepsilon_{de}, \varepsilon_{d'e'}) = 0 \text{ for } d \neq d'.
\end{aligned} \tag{4.4}$$

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

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

$$\begin{aligned} Y_{de} &= \theta_0 X_{de} + W_{de}'\gamma_0 + U_{de}, \\ X_{de} &= \pi_0 Z_{de} + W_{de}'\xi_0 + V_{de} \end{aligned} \tag{4.5}$$

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: $\forall\ l \in \{1, \ldots, 10\}$, $\forall\ m \neq l$, $\forall\ (d, e) \in \{1, \ldots, 205\} \times \{1, 2\}$, $\forall\ (d', e') \neq (d, e)$,

$$\begin{aligned} &Z_{de} \sim \mathrm{N}(0,1); W_{del} \sim \mathrm{N}(0,1), \\ &\mathrm{corr}(Z_{de}, Z_{d'e'}) = \mathrm{corr}(W_{del}, W_{d'e'l}) = f_\tau((L_d, e), (L_{d'}, e')), \\ &\mathrm{corr}(Z_{de}, W_{del}) = \mathrm{corr}(W_{del}, W_{dem}) = 0.5, \\ &\mathrm{corr}(Z_{de}, W_{d'e'l}) = \mathrm{corr}(W_{del}, W_{d'e'm}) = 0, \end{aligned} \tag{4.6}$$

for $f_\tau((L_d, e), (L_{d'}, e'))$ defined in (3.5) 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).

$$\begin{aligned} &U_{de} \sim \mathrm{N}(0,1),\ V_{de} \sim \mathrm{N}(0,1), \\ &\mathrm{corr}(U_{de}, U_{d'e'}) = \mathrm{corr}(V_{de}, V_{d'e'}) = f_\tau((L_d, e), (L_{d'}, e')), \\ &\mathrm{corr}(U_{de}, V_{de}) = 0.8, \\ &\mathrm{corr}(U_{de}, V_{d'e'}) = 0 \text{ for } (d', e') \neq (d, e), \end{aligned} \tag{4.7}$$

for $f_\tau((L_d, e), (L_{d'}, e'))$ defined in (3.5) with $\tau = (0, 3, 1)'$.

B. Spatial auto-regression (SAR).

$$\begin{aligned} &U_{de} = 0.15 \textstyle\sum_{d' \neq d} U_{d'e} \mathbf{1}_{\{\|L_d - L_{d'}\|_2 < 0.3\}} + \varepsilon_{de}, \\ &V_{de} = 0.15 \textstyle\sum_{d' \neq d} V_{d'e} \mathbf{1}_{\{\|L_d - L_{d'}\|_2 < 0.3\}} + \eta_{de}, \\ &\varepsilon_{de} \sim \mathrm{N}(0,1),\ \eta_{de} \sim \mathrm{N}(0,1), \\ &\mathrm{corr}(\varepsilon_{d1}, \varepsilon_{d2}) = \mathrm{corr}(\eta_{d1}, \eta_{d2}) = \exp(-1), \\ &\mathrm{corr}(\varepsilon_{de}, \eta_{de}) = 0.8, \\ &\mathrm{corr}(\varepsilon_{de}, \varepsilon_{d'e'}) = \mathrm{corr}(\eta_{de}, \eta_{d'e'}) = 0 \text{ for } d' \neq d, \\ &\mathrm{corr}(\varepsilon_{de}, \eta_{d'e'}) = 0 \text{ for } (d', e') \neq (d, e). \end{aligned} \tag{4.8}$$

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 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 critical value obtained by solving (2.9) 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 (2.9) as described in Section 2.
5. IM. Inference based on IM with clusters and rejection threshold obtained by solving (2.9) as described in Section 2.
6. CRS. Inference based on CRS with clusters and rejection threshold obtained by solving (2.9) as described in Section 2.

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 with free parameters using these preliminary estimates as data to obtain the structure for use in simulating the Type I and Type II error rates to allow feasible solution of (2.9). Thus, results for UNIT, CCE, IM, and CRS in the BASELINE OLS and BASELINE IV setting illustrate performance when tuning parameters result from solving (2.9) 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 (2.9) with a misspecified model. We provide detailed descriptions of the implementation of SK, UNIT, CCE, IM, and CRS in Appendix A.2.

4.2. Simulation Results

We report size of 5% level tests as well as provide results on point estimation quality in Table 3. In the OLS simulation, we obtain point estimates by applying OLS to estimate the parameters of (4.1), and we obtain point estimates in the IV simulation by applying IV to estimate the parameters in (4.5). 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 (2.9) for each procedure and $\widehat{\theta}_{\mathsf{C}}$ is a point estimator that uses only the observations in cluster C. In terms of point estimation properties, we report bias and root mean square error (RMSE) for the OLS simulation. For the IV simulation, we report median bias and median absolute deviation (MAD) due to the IV estimator's lack of finite sample moments.

The main feature of the results presented in Table 3 related to the proposal in our paper is that both IM and CRS with data dependent groups and rejection threshold determined by

TABLE 3
Simulation Results

	OLS			IV		
Method	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 4.1. 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.

(2.9) control size relatively well across the four considered designs. CCE with data dependent groups and rejection threshold determined by (2.9) also does a reasonable job controlling size across the designs, though it does somewhat less well than IM and CRS in the IV setting and is also not covered by our theoretical results. The performance of these procedures is similar in both BASELINE settings, 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.

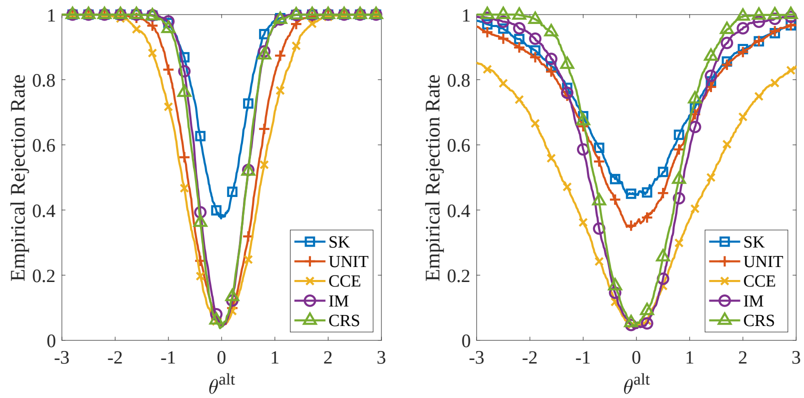
The robustness of IM, CRS, and, to a lesser extent, CCE is not exhibited by the remaining procedures. The poor behavior of UNIT-U, which ignores spatial dependence entirely, is unsurprising. More surprising is that SK, which attempts to account for spatial dependence, also does poorly across all designs considered. We suspect that adapting SK to choose tuning parameters on the basis of solving a problem similar to (2.9) would perform much better, but 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 (2.9).

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, to using the IM or CRS procedure. The results are more muddy 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 become erratic when small numbers of observations are used to form the within-group estimators. See Conley et al. (2018) for further discussion.

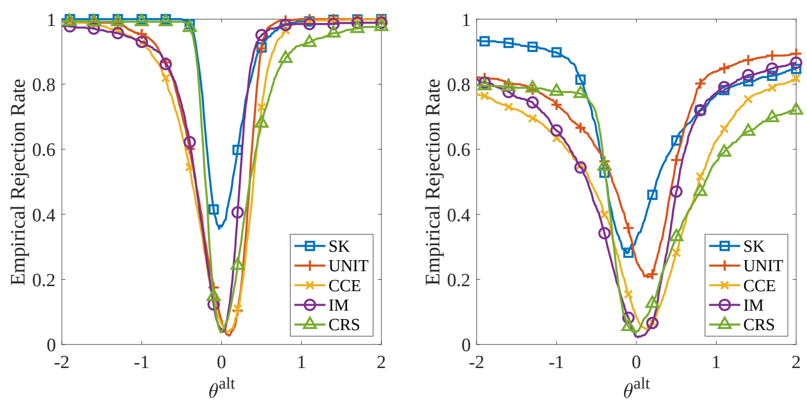
We report power curves for 5% level tests of the hypothesis $H_0 : \theta_0 = \theta^{\text{alt}}$ for alternatives θ^{alt} produced by the different procedures across the simulations in Figures 2 and 3. In these figures, the horizontal axis gives the hypothesized value, θ^{alt}, so size of the test is captured by the point $\theta^{\text{alt}} = 0$. Figure 2 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 3, 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 2 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 (2.9), $\widehat{k}$, across our simulation designs in Table 4. 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 also see that, with two 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



(a) BASELINE (b) SAR

Fig 2: OLS power curves.



(a) BASELINE (b) SAR

Fig 3: IV power curves.

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 for rejecting a 5% level test obtained from (2.9), $\widehat{\alpha}$, in Table 5. 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. We note

Table 4
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 (2.9). Panels correspond to the different designs described in Section 4.1. 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.

that this 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 (2.9) 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 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 2 control size in the simulation designs we consider while maintaining non-trivial power. In the remaining sections, we establish theoretical results that demonstrate that these procedures provide asymptotically valid inference under reasonable conditions.

TABLE 5
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 (2.9). Panels correspond to the different designs described in Section 4.1. For each design and DGP we report the .1, .25, .5, .75, and .9 quantiles of the simulation distribution of $\widehat{\alpha}$.

5. Formal Analysis of Inference with Learned Clusters

This section presents a formal analysis of the procedure defined by Algorithm 1.

The analysis of Algorithm 1 requires preliminary theoretical groundwork. This section begins by developing a notion of regularity for dissimilarity measures in Subsection 5.1. Subsections 5.2–5.6 then define an asymptotic theory which is helpful for the analysis of Algorithm 1. The regularity conditions in those subsections are sufficient for establishing a mixing central limit theorem to hold, as well as establishing relevant properties of k-`medoids`.

Subsections 5.1–5.6 are organized such that they build a ladder of increasingly higher-level regularity conditions. The penultimate rung of this ladder, Theorem 5, concludes with concrete properties for $\widehat{\theta}_{\mathsf{C}}$ for a sequence of linear models in which Algorithm 1 with $\mathsf{C} \in \mathscr{C}$ obtained by k-`medoids` has been applied. The conclusions of Theorem 5 are then used as a final rung of high-level conditions in Subections 5.8 and 5.9 to allow application of Theorems 6 and 7 which analyze size control for IM and CRS procedures with inputs more general than those arising from Algorithm 1. Versions of Theorems 6 and 7 were previously established in Ibragimov and Müller (2010) and Canay et al. (2017); and thus the primary theoretical contribution in this paper is the development of the theory in Sections 5.1–5.6. Theorems 6 and 7 do contain refinements relative to previous results which allow simultaneous analysis using multiple partitions. Sections 5.10 and 5.11 give discussion of properties of inference based on CCE as well as issues surrounding uniformity of the high-level conditions.

5.1. A notion of regularity for dissimilarity measures

This section defines formally a measure of regularity for a spatial indexing set X with dissimilarity d. The most basic condition on X is that the triangle inequality holds, in other words, each X is a genuine metric space. The next definition is used to restrict the growth rate of balls centered around any point of X. Let $|\mathsf{X}|$ be the cardinality of X. Define X to be (C, δ)-*finite-Ahlfors regular* if it is a metric space and $\min(|\mathsf{X}|, C^{-1}r^{\delta}) \leqslant |\mathsf{B}_{\mathsf{X},r}(i)| \leqslant \max(Cr^{\delta}, 1)$ for any $r > 0$, for any i in any X, where $\mathsf{B}_{\mathsf{X},r}(i)$ is the r-ball centered at i in the space X. Whenever X is both a metric space and satisfies the above definition with constants C, r, write $\mathsf{X} \in \mathbf{Ahlf}_{C,\delta}$.

(C, δ)-finite-Ahlfors regularity is a modification of the notion of Ahlfors regularity encountered in the theory of metric spaces equipped with a Borel measure μ, in which case $|\mathsf{B}_{\mathsf{X},r}(i)|$ is replaced by $\mu(\mathsf{B}_{\mathsf{X},r}(i))$ and the conditions $\min(|\mathsf{X}|, ...)$ and $\max(..., 1)$ are dropped. This notion has advantages relative to assuming that X be a subset of a Euclidean space. First, the definition refers only to intrinsic properties of the space. Second, this simple condition is sufficient both for realizing mixing central limit theorems and for analyzing clustering.[5]

Not every metric dissimilarity measure X satisfying (C, δ)-finite-Ahlfors regularity admits an isometric embedding into $\mathbb{R}^{\nu}$ for some ν. In fact, the condition is not even sufficient to guarantee that X admits a bi-Lipschitz embedding (defined so that the maximum distortion is bounded by a constant) into some $\mathbb{R}^{\nu}$. However, the condition is strong enough to ensure that X can be "regularized." The new space $(\mathsf{X}, \mathrm{d}^{1-\varepsilon})$, in which the exponent $1-\varepsilon$ is applied element-wise to d, is a metric space for all $\varepsilon \in (0, 1)$ and is called the ε-snowflake of X. The exponent ε serves to regularize the distance d so that it can be embedded into $\mathbb{R}^{\nu}$ with bounded distortion. This follows from the fact that the doubling constant of finite Ahlfors spaces can be effectively controlled, after which Assoud's Embedding Theorem (Assoud, 1977) applies. More details are given in the supplemental appendix.

5.2. Asymptotics A: Frames defined by an increasing sequence of dissimilarity measures

The analysis of Algorithm 1 in relies on asymptotic theory. In this paper, the task of indexing of data is delegated to a spatial indexing set. It is therefore convenient to build an asymptotic

[5]The theory considers X such that $(\mathrm{d}(i,j) = 0) \Rightarrow (i = j)$, as this is an immediate implication of the condition that $\mathsf{X} \in \mathbf{Ahlf}_{C,\delta}$. To rectify this convention with the data structure of the empirical example and simulation study in the previous sections, it is necessary either to consider observations $(Y_{d1}, X_{d1}, W_{d1}, Z_{d1}), (Y_{d2}, X_{d2}, W_{d2}, Z_{d2})$ as a common observation arising from a single index d, or to consider an expanded definition of $\mathbf{Ahlf}_{C,\delta}$ which allows finitely many spatial indexes to exist with mutual distance 0 to each other. Both possibilities are straightforward and can be carried without any essential changes to the arguments presented in this paper. For brevity, the details are omitted.

frame by first deliniating the classes of sequences of spatial indexing sets considered. Given the definition of (C, δ)-finite-Ahlfors regularity in the previous subsection, a natural class of sequences are of the form $\mathbb{N} \to \mathbf{Ahlf}_{C,\delta}$.

A sequence of finite dissimilarity measures will be denoted with $\mathsf{X}_{\to} = (\mathsf{X}_n)_{n\in\mathbb{N}}$. Note that in this section, unlike in the earlier sections, n appears as an explicit subscript in X_n precisely because X_n are elements of a newly introduced sequence. In the earlier sections, n was not needed to define quantities like $\widehat{T}$. From this point, any X_n appearing with a subscript always implicitly belongs to some sequence $\mathsf{X}_{\to}$.

Condition 1. *(Ahlfors Regularity) The sequence of dissimilarity measures $\mathsf{X}_{\to}$ satisfies $|\mathsf{X}_n| = n$ and is a sequence of the form $\mathbb{N} \to \mathbf{Ahlf}_{C,\delta}$ for some $0 < C < \infty, 1 \leqslant \delta \leqslant \infty$.*

Condition 1 defines a spatial asymptotic frame. Sequences satisfying Condition 1 will also be refered to as uniformly Ahlfors sequences. The utility of this notion is that it gives almost enough structure to allow analysis of k-`medoids` techniques analytically as well as derive dependent central limit theorems and laws of large numbers. The condition is also parsimonious and simple to express. Examples of metric spaces which belong to $\mathbf{Ahlf}_{C,\delta}$ for a common C, δ include the sequence with elements $\mathbb{Z}^m \cap \mathsf{Sq}_n$ where Sq_n is the m-dimensional cube of side length $2n$ centered at the origin and the sequence of annuli $\mathbb{Z}^m \cap \mathsf{A}_n$ where A_n is the m-dimensional annulus of outer radius length $2n$ and inner radius length n centered at the origin. An metric spaces which belong to $\mathbf{Ahlf}_{C,\delta}$ for a common C, δ for which no constituent metric space can be embedded into Euclidean space is the product $\{1, 2, 3, ..., n\} \times \mathsf{W}$ where $(\mathsf{W}, d_{\mathsf{W}})$ is any fixed finite metric space which cannot be embedded into Euclidean space, and the metric on the product is given by the sum of $d_{\mathbb{Z}}$ applied to the first component and d_{W} applied to the second component.

5.3. Asymptotics B: Mixing conditions and central limit theory

This section develops a central limit theorem for arrays of random variables on uniformly Ahlfors sequences of metric spaces.

Let $\mathscr{D}_{\to} = (\mathscr{D}_n)_{n\in\mathbb{N}}$ and $\mathscr{D}_n = \{\zeta_i\}_{i\in\mathsf{X}_n}$ be an array of real scalar random variables on a common probability space $(\Omega, \mathscr{F}, \mathrm{P})$ taking values in some measurable space $(V, \mathscr{V})$ with spatial indices given by X_n. In this paper, V will always be a finite dimensional real vector space $\mathbb{R}^p$ for some p. Let $\mathscr{A}$ and $\mathscr{B}$ be two (sub-)σ-algebras of $\mathscr{F}$. Define the following *mixing coefficients*.

$$\alpha^{\mathrm{mix}}(\mathscr{A}, \mathscr{B}) = \sup\{|\mathrm{P}(A \cap B) - \mathrm{P}(A)\mathrm{P}(B)| : A \in \mathscr{A}, B \in \mathscr{B}\}. \tag{5.1}$$

For $\mathsf{U}, \mathsf{V} \subseteq \mathsf{X}_n$ and integers g, l, r, define additionally

$$\begin{aligned} \alpha_n^{\mathrm{mix}}(\mathsf{U}, \mathsf{V}) &= \alpha^{\mathrm{mix}}(\sigma(\zeta_i : i \in \mathsf{U}), \sigma(\zeta_i : i \in \mathsf{V})), \\ \alpha_{g,l,n}^{\mathrm{mix}}(r) &= \sup_{\mathsf{U},\mathsf{V}\subseteq\mathsf{X}_n} \left\{\alpha_n^{\mathrm{mix}}(\mathsf{U}, \mathsf{V}), |\mathsf{U}| \leqslant g, |\mathsf{V}| \leqslant l, \mathrm{d}(\mathsf{U}, \mathsf{V}) \geqslant r\right\}, \\ \bar{\alpha}_{g,l}^{\mathrm{mix}}(r) &= \sup_{n\in\mathbb{N}} \alpha_{g,l,n}^{\mathrm{mix}}(r). \end{aligned} \tag{5.2}$$

The above definitions are the same definition as used in Jenish and Prucha (2009). The next condition, Condition 2, relies on the above definition as well as the following fact. Let $(\mathsf{X}_n, \mathrm{d}_n) \in \mathbf{Ahlf}_{C,\delta}$ for every n. Then each $(\mathsf{X}_n, \mathrm{d}_n^{3/4})$ has an L-bi-Lipschitz map into $\mathbb{R}^\nu$ where ν and the Lipschitz constant L depend only on C, δ.

Condition 2 (Mixing for an Array of Real Scalar Random Variables)**.** *The array $\mathscr{D}_\rightarrow$ is an array of real random vectors ζ_i on a common probability space $(\Omega, \mathscr{F}, \mathrm{P})$ taking values in a common real space $\mathbb{R}^p$, with spatial indices given by $\mathsf{X}_\rightarrow$ satisfying Condition 1 with Ahlfors constants $C < \infty, \delta \geqslant 1$. Let ν correspond to C, δ as defined in the preceding paragraph above. For any $\mathsf{C} \subseteq \mathsf{X}$, where $\mathsf{X} = \mathsf{X}_n$ for some n, let $\sigma^2(\mathsf{C}) = \mathrm{var}(\sum_{i\in\mathsf{C}} \zeta_i)$. There exists an array of positive constants $\{\{c_i\}_{i\in\mathsf{X}_n}\}_{n\in\mathbb{N}}$ and a positive $\mu > 0$ such that,*

(i) $\mathrm{E}[\zeta_i] = 0$.
(ii) $\lim_{k\to\infty} \sup_{n\in\mathbb{N}} \sup_{i\in\mathsf{X}_n} \mathrm{E}[\|\zeta_i/c_i\|_2^{2+\mu} \mathbf{1}_{\|\zeta_i/c_i\|_2 > k}] = 0$.
(iii) $\sum_{m=1}^\infty \bar{\alpha}_{1,1}^{\mathrm{mix}}(m) m^{\nu \times \frac{\mu+2}{\mu} - 1} < \infty$.
(iv) $\sum_{m=1}^\infty m^{\nu-1} \bar{\alpha}_{g,l}^{\mathrm{mix}}(m) < \infty$ *for* $g + l \leqslant 4$.
(v) $\bar{\alpha}_{1,\infty}^{\mathrm{mix}}(m) = O(m^{-\nu-\frac{4}{3}\mu})$.
(vi) $\inf_{n\in\mathbb{N}} \inf_{\mathsf{C}\subseteq\mathsf{X}_n} |\mathsf{C}|^{-1} (\max_{i\in\mathsf{C}} c_i^{-2}) \lambda_{\min} \sigma(\mathsf{C})^2 > 0$.

The conditions are similar to those given for the mixing central limit theorem in Jenish and Prucha (2009) which requires that each X_n be a possibly uneven lattice in a finite dimensional Euclidean space with a minimum separation between all points. Note that Condition 1 automatically implies a minimum separation between points. Note also that there are only two differences between Condition 2 and the conditions for Corollary 1 of Jenish and Prucha (2009) in the case that X_n embeds isometrically into $\mathbb{R}^\nu$ without the need to apply the snowflake regularization construction. First, here $\bar{\alpha}_{1,\infty}^{\mathrm{mix}}(m) = O(m^{-\nu-\frac{4}{3}\mu})$ rather than the slightly weaker $\bar{\alpha}_{1,\infty}^{\mathrm{mix}}(m) = O(m^{-\nu-\mu})$. Second, Condition 2$(vi)$ now entails an infimum over n and over all subsets $\mathsf{C} \subseteq \mathsf{X}_n$ while Jenish and Prucha (2009) only requires the condition hold for an infimum over n with the particular choice $\mathsf{C} = \mathsf{X}_n$. In terms of notation, here, ζ_i is used rather than the more explicit $\zeta_{i,n}$, because formally the indexes i belong to distinct sets, X_n, for each n. As a result, Condition 2 is a condition on arrays of random variables.

In the next proposition, let Φ be the Gaussian cumulative distribution function. Also, as in the statement of Condition 2, for $\mathsf{C} \subseteq \mathsf{X}$ for any $\mathsf{X} = \mathsf{X}_n$ for some n, let $\sigma(\mathsf{C}) = \mathrm{var}(\sum_{i\in\mathsf{C}} \zeta_i)$

Theorem 1. *Suppose that $\mathsf{X}_\rightarrow$ satisfies Condition 1 and $\mathscr{D}_\rightarrow$ satisfies Condition 2 with scalar ζ_i. Then for every $z \in \mathbb{R}$, $\lim_{n\to\infty} \mathrm{P}(\sigma(\mathsf{X}_n)^{-1} \sum_{i\in\mathsf{X}_n} \zeta_i \leqslant z) = \Phi(z)$.*

Theorem 1 proves convergence in distribution to a Gaussian measure for spatially indexed arrays $\mathscr{D}_n = \{\zeta_i\}_{i\in\mathsf{X}_n}$ which satisfy Conditions 1 and 2. The proof is carried out in detail in the appendix. The proposition can be proven by regularizing X_n using the snowflake construction described above with $\varepsilon = 1/4$, on which the Jenish and Prucha (2009) central limit theorem can then be applied. The same argument can be carried out more generally for other $\varepsilon \in (0, 1/2)$, however the single $\varepsilon = 1/4$ snowflake is referenced for added concreteness.

5.4. Asymptotics C: Balance and small common boundary conditions

A key observation in Bester et al. (2011) is that if a fixed sequence of partitions of X_n have balanced cluster sizes and small boundaries, then under mixing and additional regularity conditions, the CCE procedure as described in Bester et al. (2011) achieves asymptotically correct size. We show that this phenomenon also holds in the settings considered in this paper. This section formalizes notions of balanced cluster sizes and small boundaries appropriate for sequences $\mathsf{X}_\rightarrow$ satisfying Condition 1. This section then gives a proposition which shows that under the appropriate regularity conditions, the endpoint of the k-`medoids` algorithm satisfies these two desired properties. Let $\mathcal{C}_\rightarrow = (\mathcal{C}_n)_{n\in\mathbb{N}}$ be a sequence such that $\mathcal{C}_n$ is a partition of X_n for each n.

Condition 3 (Balance and Small Boundaries)**.** *The sequence of partitions $\mathcal{C}_\rightarrow$ satisfies $|\mathcal{C}_n| \leqslant k_{\max}$ for some $k_{\max}$ which does not depend on n, and is asymptotically balanced with small boundaries in the sense of the following conditions.*

(i) The cluster sizes satisfy $\liminf_{n\to\infty} \frac{\min_{\mathsf{C}\in\mathcal{C}_n} |\mathsf{C}|}{\max_{\mathsf{D}\in\mathcal{C}_n} |\mathsf{D}|} > 0$,

(ii) There is a sequence $\bar{r}_\rightarrow = (\bar{r}_n)_{n\in\mathbb{N}} \nearrow \infty$ *with* $\lim_{n\to\infty} \frac{\max_{\mathsf{C}\in\mathcal{C}_n} |\{i\in\mathsf{C}: \mathrm{d}(i,\mathsf{X}_n\setminus\mathsf{C})\leqslant\bar{r}_n\}|}{\min_{\mathsf{D}\in\mathcal{C}_n} |\mathsf{D}|} = 0.$

The definition of small boundary provided in Condition 3(ii) differs slightly from the definition given in Bester et al. (2011). In particular, Bester et al. (2011) leverage the fact their spatial domain is a subset of the integer lattice to define neighbor orders for pairs of locations. Their definition of small boundaries entails a bound on the number of first order neighbors from C to $\mathsf{X}_n \setminus \mathsf{C}$. They assume that their given spatial clusters are contiguous and use that fact to bound the number of higher order neighbors from C to $\mathsf{X}_n \setminus \mathsf{C}$. In this context, there is no available definition of first order neighbor since X_n can be irregular (even non-Euclidean). As a result, this paper works instead with an asymptotic notion of boundary which entails a sequence $\bar{r}_\rightarrow$ which allows boundaries to widen as $n \to \infty$. An implication of having asymptotically balanced with small boundaries clusterings is the following proposition.

Theorem 2. *Suppose that $\mathsf{X}_\rightarrow$ satisfies Condition 1, $\mathscr{D}_\rightarrow$ satisfies Condition 2, and $\mathcal{C}_\rightarrow$ satisfies Condition 3. Then* $\lim_{n\to\infty} \sup_{\mathsf{C}\neq\mathsf{D}\in\mathcal{C}_n} \mathrm{cov}_{\mathrm{P}} \left(\sigma(\mathsf{C})^{-1}\sum_{i\in\mathsf{C}}\zeta_i,\ \sigma(\mathsf{D})^{-1}\sum_{i\in\mathsf{D}}\zeta_i\right) = 0.$

The argument establishing the above fact is related to but not identical to arguments in Bester et al. (2011), which were previously also given in Jenish and Prucha (2009) and Bolthausen (1982). Instead of counting points in "shells" around the boundaries of clusters, the proof of Theorem 2 instead relies on the doubling structure implied by the fact that X_n are Ahlfors regular. Both arguments leverage a bound on covariances $\mathrm{cov}(\zeta_i, \zeta_j)$ for sufficiently distant spatial locations as implied by the mixing conditions stated in Condition 2.

Condition 3 is a taken as a high-level assumption on clustering structures in Theorem 2. Next, we verify that clusters produced by k-`medoids` satisfy Condition 3 under an additional convexity assumption. k-`medoids` is a popular clustering technique which is related to k-`means`, and both produce similar clustering results in many settings. However, k-`means` centroids are not necessarily defined in the space X_n when dealing with a dissimilarity which does not necessarily arise from a Euclidean space. k-`medoids` differs by requiring that clusters are defined around elements of X_n, called medoids, which must themselves also be elements of X_n. For a detailed description of k-`medoids` see Hastie et al. (2009).

The next proposition derives relevant asymptotic properties for a k-`medoids` algorithm by studying the sequence of partitions

$$\mathcal{C}_{\rightarrow} = (\mathcal{C}_n)_{n\in\mathbb{N}} \text{ defined by } \mathcal{C}_n = k\text{-}\mathtt{medoids}(\mathsf{X}_n). \tag{5.3}$$

Theorem 3. *Suppose that $\mathsf{X}_{\rightarrow}$ satisfies Condition 1. Assume the following additional convexity condition holds. There is a constant K independent of n such that for each n, (1) X_n is K-coursely isometric[6] to a subset of a Euclidean space with dimension ν independent of n, and (2) for any two point $i, j \in \mathsf{X}_n$ and any $a \in [0, 1]$ there is an interpolant $g \in \mathsf{X}_n$ such that $|\mathrm{d}_n(i, g) - a\mathrm{d}_n(j, g)| \leqslant K$ and $|\mathrm{d}_n(i, g) - (1-a)\mathrm{d}_n(j, g)| \leqslant K$. Let $\mathcal{C}_n = k$-`medoids`(X_n) and let $\mathcal{C}_{\rightarrow} = (\mathcal{C}_n)_{n\in\mathbb{N}}$. Then $\mathcal{C}_{\rightarrow}$ satisfies Condition 3.*

Theorem 3 does not require that the sequence of partitions $\mathcal{C}_n$ of X_n resulting from applying k-`medoids` to converge in any way. It also does not require that there be any notion of true clusters or a true partition associated to any of the X_n.

The convexity assumption is needed to exclude cases such as the following would-be counterexample:

$$\begin{aligned}\mathsf{Rothko}_n = \{(i^1, i^2) \in \mathbb{Z}^2 : |i^1| \leqslant n, |i^2| \leqslant 2n\} \\ \setminus \{(i^1, i^2) \in \mathbb{Z}^2 : |i^1| \leqslant n-2, 1 \leqslant |i^2 + \frac{1}{2}| \leqslant 2n-2\}.\end{aligned} \tag{5.4}$$

The naming for the above example is based on the fact that when plotted, Rothko_n resembles the background orange region in the oil painting Rothko (1956), titled *Orange and Yellow.* A possible sequence of endpoints of the 2-`medoids` algorithm for the above example is with medoids at $(0, 0)$ and $(0, -1)$ and boundary (with $\bar{r} = 1$) at $\{0, 1\} \times \{-n, -n + 1, ..., n\}$. Note that these medoid endpoints are not unique, and other medoid sequences, like $(-n, 2n), (n, -2n)$ do lead to small boundaries.

5.5. Asymptotics D: Almost sure representation

The next preliminary for the analysis of Algorithm 1 is an almost sure representation theorem. Suppose that $\mathscr{C}_{\rightarrow}$ is a sequence of collections of partitions of X_n. That is $\mathscr{C}_{\rightarrow} = (\mathscr{C}_n)_{n\in\mathbb{N}}$

[6] $f : (\mathsf{Y}, \mathrm{d}_\mathsf{Y}) \rightarrow (\mathsf{Z}, \mathrm{d}_\mathsf{Z})$ is a K-course isometry if $\mathrm{d}_\mathsf{Z}(f(i), f(j)) - K \leqslant \mathrm{d}_\mathsf{Y}(i, j) \leqslant \mathrm{d}_\mathsf{Z}(f(i), f(j)) + K$.

and each $\mathscr{C}_n$ is a set containing elements $\mathcal{C}$ which are partitions of X_n. The following condition concerns corresponding collections $S_{\mathcal{C}}$ of random vectors taking values in $\mathbb{R}^{\mathcal{C}}$ where $\mathcal{C} \in \mathscr{C}_n$ for some n. Let $\{S_{\mathcal{C}}\}$ be a collection of random vectors in $\mathbb{R}^{\mathcal{C}}$ where $\mathcal{C}$ range in $\cup_{n\in\mathbb{N}}\mathscr{C}_n$.

Condition 4 (Almost Sure Asymptotic Gaussian Representation)**.** *Associated to all* $S_{\mathcal{C}} \in \mathbb{R}^{\mathcal{C}}$ *for all* $\mathcal{C} \in \cup_{n\in\mathbb{N}}\mathscr{C}_n$ *is a common probability space* $(\tilde{\Omega}, \tilde{\mathscr{F}}, \tilde{\mathrm{P}})$ *with random variables* $\tilde{S}_{\mathcal{C}}, \tilde{S}^*_{\mathcal{C}} \in \mathbb{R}^{\mathcal{C}}$. *Furthermore,* $\tilde{S}_{\mathcal{C}}$ *and* $S_{\mathcal{C}}$ *have the same distribution,* $\tilde{S}^*_{\mathcal{C}}$ *is zero-mean Gaussian with independent components,* $\liminf_{n\to\infty} \min_{\mathsf{C}\in\mathcal{C}\in\mathscr{C}_n} \mathrm{var}(\tilde{S}^*_{\mathcal{C},\mathsf{C}}) > 0$, $\limsup_{n\to\infty} \max_{\mathsf{C}\in\mathcal{C}\in\mathscr{C}_n} \mathrm{var}(\tilde{S}^*_{\mathcal{C}_n,\mathsf{C}}) < \infty$ *and* $\lim_{n\to\infty} \sup_{\mathcal{C}\in\mathscr{C}_n} \|\tilde{S}_{\mathcal{C}} - \tilde{S}^*_{\mathcal{C}}\|_2 = 0$ $\tilde{\mathrm{P}}$*-almost surely.*

If the above condition holds, then $S_{\mathcal{C}}$ are said to be uniformly asymptotically almost surely representable by non degenerate independent Gaussians. When needed, the representors relative to $S_{\mathcal{C}}$ will always be referred to by $\tilde{S}^*_{\mathcal{C}}$ and $\tilde{S}_{\mathcal{C}}$ just as in the definition. It is a well known result that convergence in distribution of a sequence of random variables implies almost sure convergence for an auxiliary sequence of random variables with the same distribution as the original sequence. Specifically, Theorem 2.19 in van der Vaart (1998) states: "*Suppose that the sequence of random vectors* X_n *converges in distribution to a random vector* X_0. *Then there exists a probability space* $(\tilde{\Omega}, \tilde{\mathcal{U}}, \tilde{P})$ *and random vectors* $\tilde{X}_n$ *defined on it such that* $\tilde{X}_n$ *is equal in distribution to* X_n *for every* $n \geqslant 0$ *and* $\tilde{X}_n \to \tilde{X}_0$ *almost surely.*"

A subtle problem that occurs in the current setting is that the random vectors $S_{\mathcal{C}}$ take values in $\mathbb{R}^{\mathcal{C}}$, which are technically different spaces for different $\mathcal{C}$. One collection of $S_{\mathcal{C}}$ of interest are constructed from $\mathscr{D}_\rightarrow$ and a sequence of partitions $\mathcal{C}_\rightarrow$ of $\mathsf{X}_\rightarrow$ by

$$S_{\mathcal{C},\mathsf{C}} = (n/k)^{-1/2} \sum_{i\in\mathsf{C}} \zeta_i, \text{ for } \mathcal{C} \text{ partitioning } \mathsf{X}_n \text{ and } |\mathcal{C}| = k. \tag{5.5}$$

The issue of $S_{\mathcal{C}}$ belonging to different spaces does not yet exist in Theorem 2 because the covariance is relative to the measure P, which is common for all n. The issue is also circumvented in Theorem 1 because there, the terms $\sigma(\mathsf{X}_n)^{-1} \sum_{i\in\mathsf{X}_n} \zeta_i$ may be taken to belong to a common set, namely $\mathbb{R}$, as each ζ_i belongs to $\mathbb{R}$. Nevertheless, the following holds.

Theorem 4. *Suppose that* $\mathsf{X}_\rightarrow$ *satisfies Condition 1,* $\mathscr{D}_\rightarrow$ *satisfies Condition 2, and* $\mathcal{C}_\rightarrow$ *satisfies Condition 3 with the additional condition that* $|\mathcal{C}_n| \leqslant k_{\max}$ *with* $k_{\max}$ *independent of* n. *For* $\mathcal{C}$ *with* $|\mathcal{C}| = k$, *let* $S_{\mathcal{C},\mathsf{C}} = (n/k)^{-1/2} \sum_{i\in\mathsf{C}} \zeta_i$. *Let* $\mathscr{C}_n$ *be the singleton* $\mathscr{C}_n = \{\mathcal{C}_n\}$. *Then* $\{S_{\mathcal{C}}\}_{\mathcal{C}\in\cup_{n\in\mathbb{N}}\mathscr{C}_n}$ *satisfies Condition 4.*

5.6. Asymptotics E: Central Limit theory in the cases of OLS and IV

The previous theorems establish asymptotic normality and diminishing across-group correlation for means. This section gives a similar limiting result in the context of ordinary least squares (OLS) and instrumental variables (IV) estimators. These are the precise central limit theorems which are relevant to the subsequent analysis of cluster-based inferential procedures.

Consider data $\mathscr{D}_n = \{\zeta_i\}_{i \in \mathsf{X}_n}$ defined on $(\Omega, \mathscr{F}, \mathrm{P})$ with $\zeta_i = (Y_i, X_i, W_i, Z_i)$ in which for each $i \in \mathsf{X}_n$ (and for each n) a linear model holds as follows:

$$Y_i = \theta_0 X_i + W_i' \gamma_0 + U_i, \tag{5.6}$$

where Y_i is a scalar outcome variable, X_i is a scalar variable of interest, U_i is a scalar idiosyncratic disturbance term, W_i is a p-dimensional vector of control variables, and Z_i is a $(p+1)$-dimensional vector of instruments. The following regularity condition is helpful.

Condition 5. *$\mathscr{D}_{\rightarrow}$ consists of $\mathscr{D}_n = \{(Y_i, X_i, W_i, Z_i)\}_{i \in \mathsf{X}_n}$ satisfying the linear model $Y_i = \theta_0 X_i + W_i' \gamma_0 + U_i$ for some θ_0 and γ_0. $\mathrm{E}[Z_i U_i] = 0$, $\mathrm{E}[Z_i X_i']$ exists and equals some M which does not depend on i and which has positive eigenvalues. $(\{Z_i U_i, Z_i X_i' - \mathrm{E}[Z_i X_i']\}_{i \in \mathsf{X}_n})_{n \in \mathbb{N}}$ satisfies Condition 2.*

Consider the IV estimator for θ_0 using only observations $\mathsf{C} \subseteq \mathsf{X}_n$, and, as earlier, denote this by $\widehat{\theta}_{\mathsf{C}}$. Note, the OLS estimate is a special case where $Z_i = X_i$. Let $\mathcal{C}$ be a partition of X_n for some n with k clusters. Define $S_{\mathcal{C}} \in \mathbb{R}^{\mathcal{C}}$ by

$$S_{\mathcal{C},\mathsf{C}} = (n/k)^{1/2} (\widehat{\theta}_{\mathsf{C}} - \theta_0) \text{ for } \mathsf{C} \in \mathcal{C}. \tag{5.7}$$

Let $\mathscr{C}_n$ again be the singleton $\mathscr{C}_n = \{\mathcal{C}_n\}$. The next theorem is the particular asymptotic normality result about $\widehat{\theta}_{\mathsf{C}}$, $\mathsf{C} \in \mathcal{C}_n$ which is useful for input into the analysis of Algorithm 1 in the context of testing hypotheses about coefficients in a linear model.

Theorem 5. *Suppose that $\mathsf{X}_{\rightarrow}$ satisfies Condition 1. Suppose linear regression data $\mathscr{D}_{\rightarrow}$ satisfies Condition 5. Suppose that $\mathcal{C}_{\rightarrow}$ satisfies Condition 3. Then for $S_{\mathcal{C}}$ defined above in this subsection and for $\mathscr{C}_n = \{\mathcal{C}_n\}$, the collection $\{S_{\mathcal{C}}\}_{\mathcal{C} \in \cup_{n \in \mathbb{N}} \mathscr{C}_n}$ satisfies Condition 4.*

5.7. Analysis of Cluster-Based Inference with Learned Clusters

Consider an array $\mathscr{D}_{\rightarrow}$ of spatially indexed datasets $\mathscr{D}_n = \{\zeta_i\}_{i \in \mathsf{X}_n}$ all distributed according to one common P_0. Consider a sequence of statistical hypotheses $H_{0\rightarrow} = \{H_{0n}\}_{n \in \mathbb{N}}$ where each H_{0n} is a restriction on only the distribution of the nth sample of $\mathscr{D}_{\rightarrow}$ (i.e., on $\mathscr{D}_n$). The coming sections discuss formal properties of sequences of tests $\widehat{T}_{\rightarrow} = (\widehat{T}_n)_{n \in \mathbb{N}}$ with primary interest in $\widehat{T}_{\rightarrow} = \widehat{T}_{\bullet(\alpha)\rightarrow}$ arising from $\widehat{T}_{\bullet(\alpha)}$ defined by applying Algorithm 1 to data $\mathscr{D}_n, \mathsf{X}_n$ with $\bullet =$ IM, CRS, or CCE as in Section 2.

For any n, let $\mathscr{C}_n$ be its associated collection of partitions

$$\mathscr{C}_n = \{\mathcal{C} : \mathcal{C} = \texttt{Cluster}(\mathsf{X}_n) \text{ and } (\texttt{Test}, \texttt{Cluster}) \in \mathscr{T}_n \text{ for some } \texttt{Test}\}. \tag{5.8}$$

For each n, there is an element $\widehat{\texttt{Cluster}}(\mathsf{X}_n) \in \mathscr{C}_n$ which is defined by the solution to the optimization problem in equation (2.9).

An important property used in the analysis of Algorithm 1 in Sections 5.8 and 5.9 is that the act of selecting partitions $\widehat{\mathcal{C}}_n$ has negligible effect on $\widehat{T}_n$. This property is formalized in the next condition.

Condition 6. *Corresponding to $\widehat{T}_{\bullet(\alpha)\rightarrow}$, there are sets $\mathscr{C}_{1,n} \subseteq \mathscr{C}_n$ for all n such that the following are satisfied:*

(i) $\lim_{n\to\infty} \mathrm{P}_0(\widehat{\texttt{Cluster}}(\mathsf{X}_n) \notin \mathscr{C}_{1,n}) = 0,$

(ii) $\lim_{n\to\infty} \sup_{\mathcal{C}\in\mathscr{C}_{1,n}} \left| \mathrm{P}_0(T_{\bullet(\alpha),\mathcal{C}} = \mathrm{Reject} | \widehat{\texttt{Cluster}}(\mathsf{X}_n) = \mathcal{C}) - \mathrm{P}_0(T_{\bullet(\alpha),\mathcal{C}} = \mathrm{Reject}) \right| = 0.$

When specializing to the context of Algorithm 1 in Section 2, Condition 6 is equivalently a high-level condition on the properties $\widehat{\mathrm{Err}}_{\text{Type-I}}$ and $\widehat{\mathrm{Err}}_{\text{Type-II},\mathbf{P}_{\text{alt}}}$. Recall that the method for selecting $\widehat{\mathcal{C}}$ described in Section 2 is through a combination of QMLE, simulation, and constrained minimization of a loss function (defined as estimated weighted average power). Under sufficient regularity conditions, the QMLE procedure defined in the appendix is consistent for a fixed object which may depend on n. If the loss function is sufficiently regular, than $\mathscr{C}_{1,n}$ satisfying Condition 6 can be taken to be a singleton. Alternatively, if a single fixed k is used in conjunction with k-`medoids` clustering, Condition 6 is automatically satisfied.

5.8. General analysis for IM with learned clusters

The IM procedure at level a was defined in Section 2 for a scalar hypothesis about a regression coefficient in a linear model of the form $H_0 : \theta_0 = \theta^{\circledast}$, depending on a partition $\mathcal{C}$ of the data's spatial indexing set. In particular, recall that in Section 2, the IM test was defined for $a \in (0,1)$ and given by $T_{\mathrm{IM}(a),\mathcal{C}} = \mathrm{Reject} \quad \text{if} \quad |t(S)| > t_{1-a/2,k-1}$, where $t_{1-a/2,k-1}$ is the $(1-a/2)$-quantile of t-distribution with $k-1$ degrees of freedom and with $S = (S_{\mathsf{C}})_{\mathsf{C}\in\mathcal{C}} \in \mathbb{R}^{\mathcal{C}}$, $|\mathcal{C}| = k$, $S_{\mathsf{C}} = (n/k)^{1/2}(\widehat{\theta}_{\mathsf{C}} - \theta^{\circledast})$.

Theorem 6. *Suppose that $\{S_{\mathcal{C}}\}_{\mathcal{C}\in\cup_{n\in\mathbb{N}}\mathscr{C}_n}$ satisfy Condition 4 under* P_0. *Suppose $\widehat{T}_{\mathrm{IM}(\alpha)\rightarrow}$ satisfies Condition 6. Suppose further that $\alpha \leq 2\Phi(-\sqrt{3}) = 0.083...$, where Φ the standard Gaussian cumulative distribution function. Then* $\limsup_{n\to\infty} \mathrm{P}_0(\widehat{T}_{\mathrm{IM}(\alpha),n} = \mathrm{Reject}) \leqslant \alpha$.

The condition that $\alpha \leq 2\Phi(-\sqrt{3})$ is also needed in Ibragimov and Müller (2010). Note that the conditions above were verified for the OLS and IV models under regularity conditions given in the previous section with partitions generated by k-`medoids` with fixed k. The properties of $\widehat{\mathrm{Err}}_{\text{Type-I}}$ and $\widehat{\mathrm{Err}}_{\text{Type-I},\mathbf{P}_{\text{alt}}}$ do not appear explicitly in the conditions of Theorem 6. However, their properties are reflected through Condition 6. The proof of Theorem 6 builds largely on the arguments in Ibragimov and Müller (2010). Relative to Ibragimov and Müller (2010), Theorem 6 requires establishing uniform limit results over tests corresponding to partitions in $\mathscr{C}_{\rightarrow}$. Additional details can be found in the discussions provided within the proof in the supplemental appendix.

Theorem 6 is helpful in understanding the statistical performance of Algorithm 1 with IM in the context of the hypothesis tests considered in the empirical application in Section

3 and simulation study in Section 4. In particular, consider a sequence of regression datasets $\mathscr{D}_\rightarrow$ with $\mathscr{D}_n = \{\zeta_i\}_{i \in \mathsf{X}_n}$ with $\zeta_i = (Y_i, X_i, W_i, Z_i)$ as in Section 5.6 satisfying $Y_i = \theta_0 X_i + W_i'\gamma_0 + U_i$.

Corollary 1. *(IM and k-medoids in the linear model). Suppose that $\mathsf{X}_\rightarrow$ satisfies Condition 1. Suppose that $\mathscr{D}_\rightarrow$ satisfies Condition 5. Let $H_{0\rightarrow} = (H_{0,n})_{n\in\mathbb{N}}$ be the sequence of hypotheses $H_{0,n} : \theta_0 = \theta^{\circledast}$. Let $\widehat{T}_{\text{IM}(\alpha)\rightarrow} = (\widehat{T}_{\text{IM}(\alpha),n})_{n\in\mathbb{N}}$ be defined by Algorithm 1 applied to $(\mathscr{D}_n, \mathsf{X}_n)$ for hypotheses $H_{0,n}$ for each n with k-medoids and $k_{\max}$ fixed independent of n. Suppose $\widehat{\text{Err}}_{\text{Type-I}}$ and $\widehat{\text{Err}}_{\text{Type-II},\mathbf{P}_{\text{alt}}}$ are such that $\widehat{T}_{\text{IM}(\alpha)\rightarrow}$ satisfies Condition 6. Then $\lim_{n\to\infty} \mathrm{P}_0(\widehat{T}_{\text{IM}(\alpha),n} = \text{ Reject }) \leqslant \alpha$.*

5.9. General analysis for CRS with learned clusters

The procedure defining $T_{\text{CRS}(\alpha)}$ can be stated in slightly more generality. In this more general setup, CRS relative to a partition $\mathcal{C}$ depends on a real-valued function $w : \mathbb{R}^{\mathcal{C}} \to \mathbb{R}$, and a statistic $S \in \mathbb{R}^{\mathcal{C}}$, which is chosen such that large values of w provide evidence against H_0. As was described in Section 2, in all implementations in this paper, w is chosen to be $w(S) = |t(S)|$ and S has components $S_{\mathsf{C}} = (n/k)^{1/2}(\widehat{\theta}_{\mathsf{C}} - \theta^{\circledast})$ with $k = |\mathcal{C}|$. Note, the CRS test was developed in Canay et al. (2017) for a single fixed partition and relies on a notion of approximate symmetry. Specifically, given a partition $\mathcal{C}$ of X_n, the CRS procedure also depends on the availability of a finite group $\mathcal{H}_{\mathcal{C}}$ of symmetries which act within component C of $\mathcal{C}$. These symmetries are formalized by a group action $\mathscr{D}_n \mapsto h\mathscr{D}_n$ for all $h \in \mathcal{H}_{\mathcal{C}}$. The action on the data is defined in such a way that

$$\mathscr{D}_n \mapsto h\mathscr{D}_n \;\text{ induces }\; S_{\mathcal{C}} \mapsto hS_{\mathcal{C}} \tag{5.9}$$

which respects the original action in the sense $S_{\mathcal{C}}(h\mathscr{D}_n) = hS_{\mathcal{C}}(\mathscr{D}_n)$. With slight abuse of notation, write $w(\mathscr{D}_n) = w(S_{\mathcal{C}})$.

In the context of cluster-based inference in this paper, the only type of group action of interest will be set of signs $\mathcal{H}_{\mathcal{C}} = \{-1, +1\}^{\mathcal{C}}$ that operates within clusters. Elements, h, which are tuples of signs $h_{\mathsf{C}} \in \{-1, +1\}$ indexed by $\mathsf{C} \in \mathcal{C}$ can be defined to act in a way such that $S_{\mathcal{C},\mathsf{C}} \mapsto h_{\mathsf{C}} S_{\mathcal{C},\mathsf{C}}$. In the context of the OLS and IV models, $(\widehat{\theta}_{\mathsf{C}} - \theta^{\circledast}) \mapsto h_{\mathsf{C}}(\widehat{\theta}_{\mathsf{C}} - \theta^{\circledast})$

To complete the general description of CRS, let $M = |\mathcal{H}_{\mathcal{C}}|$ and let $w^1(S_{\mathcal{C}}) \leqslant w^2(S_{\mathcal{C}}) \leqslant \cdots \leqslant w^M(S_{\mathcal{C}})$ denote the order statistics of the orbit $\{w\,(h\mathscr{D}_n) : h \in \mathcal{H}\}$. Let $j_a = M(1-a)$, $M^+(X) = |\{1 \leqslant j \leqslant M : w^j(S_{\mathcal{C}}) > w^{j_a} S_{\mathcal{C}}\}|$ and $M^0(X) = |\{1 \leqslant j \leqslant M : w^j(X) = w^{j_a}(S_{\mathcal{C}})\}|$. Let $\tilde{a}\,(S_{\mathcal{C}}) = \frac{Ma - M^+(S_{\mathcal{C}})}{M^0(S_{\mathcal{C}})}$. Set $T_{\text{CRS}(a),\mathcal{C}} =$ Reject if $w(S_{\mathcal{C}}) > w^{j_a}(S_{\mathcal{C}})$ or $\big(w(S_{\mathcal{C}}) = w^{j_a}(S_{\mathcal{C}})$ and $\tilde{a}(S_{\mathcal{C}}) = 1\big)$. $\widehat{T}_{\text{CRS}(\alpha)}$ is then defined as in Section 2.

Condition 7 (Regularity Conditions for CRS)**.**

(i) *$\{S_{\mathcal{C}}\}_{\mathcal{C} \in \cup_{n\in\mathbb{N}} \mathscr{C}_n}$ satisfy Condition 4.*

(ii) *$h\tilde{S}^*_{\mathcal{C}}$ has the same distribution as $\tilde{S}^*_{\mathcal{C}}$ for all $h \in \mathcal{H}_{\mathcal{C}}$, for all $\mathcal{C}$ in all $\mathscr{C}_n$; for distinct h, h', either $w \circ h = w \circ h'$ or $\tilde{\mathrm{P}}(w(h\tilde{S}^*_{\mathcal{C}}) \neq w(h'\tilde{S}^*_{\mathcal{C}})) = 1$; w is continuous and the action of h is continuous for each h in each $\mathcal{H}_{\mathcal{C}}$.*

(iii) $|\mathscr{C}_n|$ *and* $\max_{\mathcal{C}\in\mathscr{C}_n}|\mathcal{C}|$ *are each bounded by a constant independent of* n.

Theorem 7. *Suppose that* $\{S_{\mathcal{C}}\}_{\mathcal{C}\in\cup_{n\in\mathbb{N}}\mathscr{C}_n}$ *satisfies Condition 7 under* P_0. *Suppose* $\widehat{T}_{\mathrm{CRS}(\alpha)\to}$ *satisfies Condition 6. Then* $\limsup_{n\to\infty}\mathrm{P}_0(\widehat{T}_{\mathrm{CRS}(\alpha),n} = \text{Reject}) \leqslant \alpha$.

In Canay et al. (2017), it was shown that the conclusion of Theorem 7 holds under conditions which are subsumed by Condition 7, working in the case that $\mathscr{C}_n$ are singletons, i.e., $\mathscr{C}_n = \{\mathcal{C}_n\}$ consist of a unique predetermined partition of $\{1, ..., n\}$ without any described dependence on any spatial indexing X_n.

Condition 7(iii) is stronger than necessary. Discussion about weakening Condition 7(iii) is postponed until Subsection 5.11 about uniformity. In particular, no requirement on $|\mathscr{C}_n|$ are needed when $w(S) = |t(S)|$. The supplemental material proves Theorem 7 under a strictly weaker condition, Condition 7(iii)', which is stated formally there. The argument accommodating the ability to generalize to this more general case is new relative to Canay et al. (2017).

As was the case in previous section with the analysis of IM, Theorem 7 has an important corollary which specializes to the specific case of Algorithm 1 (with k-`medoids`) for test $H_0 : \theta_0 = \theta^{\circledast}$, where θ_0 is an unknown coefficient in a linear regression model. Once again, consider a sequence of regression datasets $\mathscr{D}_\to$ with $\mathscr{D}_n = \{\zeta_i\}_{i\in\mathsf{X}_n}$ with $\zeta_i = (Y_i, X_i, W_i, Z_i)$ as in Subsection 5.6 above satisfying $Y_i = \theta_0 X_i + W_i'\gamma_0 + U_i$.

Corollary 2. *(*CRS *and* k-`medoids` *in the linear model). Suppose that* $\mathsf{X}_\to$ *satisfies Condition 1. Suppose that* $\mathscr{D}_\to$ *satisfies Condition 5. Let* $H_{0\to} = (H_{0,n})_{n\in\mathbb{N}}$ *be the sequence of hypotheses* $H_{0,n} : \theta_0 = \theta^{\circledast}$. *Let* $\widehat{T}_{\mathrm{CRS}(\alpha)\to} = (\widehat{T}_{\mathrm{CRS}(\alpha),n})_{n\in\mathbb{N}}$ *be defined by Algorithm 1 applied to* $(\mathscr{D}_n, \mathsf{X}_n)$ *for hypotheses* $H_{0,n}$ *for each* n *with* k*-medoids and* $k_{\max}$ *fixed independent of* n. *Suppose* $\widehat{\mathrm{Err}}_{\text{Type-I}}$ *and* $\widehat{\mathrm{Err}}_{\text{Type-II},\mathbf{P}_{\text{alt}}}$ *are such that* $\widehat{T}_{\mathrm{CRS}(\alpha)\to}$ *satisfies Condition 6. Then* $\lim_{n\to\infty}\mathrm{P}_0(\widehat{T}_{\mathrm{IM}(\alpha),n} = \text{ Reject }) \leqslant \alpha$.

5.10. Discussion for BCH with learned clusters

Recall that the CCE test for $H_0 : \theta_0 = \theta^{\circledast}$ using a partition $\mathcal{C}_n$ with k clusters is defined by $T_{\mathrm{CCE}(a),n} = \text{ Reject if } \frac{\widehat{\theta}_{\mathsf{X}_n} - \theta^{\circledast}}{\widehat{V}^{1/2}_{\mathrm{CCE},\mathcal{C}_n}} > \sqrt{\frac{k}{k-1}} \times t_{1-a/2,k-1}$ where $\widehat{V}_{\mathrm{CCE},\mathcal{C}_n}$ is the standard cluster covariance estimator (without a degrees of freedom correction). BCH analyzed the CCE estimator under an asymptotic frame with a fixed, finite number of clusters k (i.e. k independent of n). The resulting inference is based on calculating a t statistic based on an estimated standard error. Under regularity conditions, BCH show that the resulting t statistic is asymptotically pivotal, but distributed according to $\sqrt{k/(k-1)} \times t_{k-1}$ where t_{k-1} is the t distribution with $k-1$ degrees of freedom.

The regularity conditions required in Bester et al. (2011) are strong, and in particular, require that the clusters have asymptotically equal numbers of observations. The k-`medoids` algorithm does not generally return such a partition. As a result, the CCE estimator is not anticipated to have asymptotically correct size.

5.11. Discussion of uniformity

The high-level conditions used for proving Theorems 6 and 7 involve the assumption of uniform central limit theorems for convergences $\sup_{\mathcal{C}\in\mathscr{C}_n} \|\tilde{S}_{\mathcal{C}} - \tilde{S}^*_{\mathcal{C}}\|_2 \to 0$ for Gaussian random variables $\tilde{S}_{\mathcal{C}}$. Uniform convergence results may be explicitly derived in several cases of interest. First, when each $\mathscr{C}_n$ contains only finitely many partitions $\mathcal{C}$ all of which have cardinality bounded by some $k_{\max}$ independent of n, then uniform convergence follows immediately from a pointwise convergence result like that presented in Section 5.6.

Another case of interest involves increasing sequences $k_{\max,n}$. In particular, if $k_{\max,n}$ increases sufficiently slowly, then uniform analogues of the results in Theorems 6 and 7 may be anticipated by establishing Berry-Esseen-type bounds for dependent processes. Note, for instance, that Jirak (2016) establishes a unidimensional Berry-Esseen bound for sums of weakly dependent random variables. Note also that standard dimension-dependent Berry-Esseen bounds for independent data scale as $O(k^{1/4}n^{-1/2})$ for total variation distance, implying bounds $O(k^{3/4}n^{-1/2})$ when passing to Euclidean distance, i.e., $\|\tilde{S}_{\mathcal{C}} - \tilde{S}^*_{\mathcal{C}}\|_2$. Note also that Theorem 3 can be extended to sequences $k_{\max,n}$ sufficiently by the same argument as given in the supplemental material with $\check{n} := \lfloor n/k \rfloor$ replacing n in the appropriate places.

A final case to consider is to fix k but allow $\mathscr{C}_n$ to contain many different partitions. For example, this case could arise by considering the output of k-`medoids` algorithm with many different initial starting medoids. In this case, arguments from standard empirical process theory may be used. For brevity, this paper does not carry out such a program further. Note, however, that the set of k-tuples of points in a bounded subset of Euclidean space has known metric entropy properties. These metric entropies may be used to control the metric entropy of tuples in X_n after embedding $\mathsf{X}_n^{1-\varepsilon}$ into a Euclidean space for suitable $\varepsilon > 0$.

The conditions imposed for Theorem 6 do not explicitly require any bounds on $|\mathscr{C}_n|$ or $|\mathcal{C}|, \mathcal{C} \in \mathscr{C}_n$. Condition 7(iii) used in Theorem 7 does require bounds on the $|\mathscr{C}_n|$ or $|\mathcal{C}|, \mathcal{C} \in \mathscr{C}_n$ which are independent of n, but this requirement can be dispensed if $w(S)$ is sufficiently regular. This regularity condition is given in Condition 7(iii)' in the supplemental material. There also is provided a proposition showing that $w(S) = |t(S)|$ satisfies the required regularity condition.

Appendix A: Implementation Details

A.1. Implementation of k-`medoids` clustering

This section states the k-`medoids` algorithm used in the main text. For finite (X, d), let $\mathsf{C} \subseteq \mathsf{X}$ and define the cost of a cluster C with medoid $i \in \mathsf{X}$ to be $\mathrm{cost}(\mathsf{C}, i) = \sum_{j\in\mathsf{C}} \mathrm{d}(i,j)^2$. The total cost for a partition $\mathcal{C}$ and set of medoids $\{i_{\mathsf{C}}\}_{\mathsf{C}\in\mathcal{C}}$ is defined by summing over clusters total $\mathrm{cost}(\mathcal{C}, \{i_{\mathsf{C}}\}_{\mathsf{C}\in\mathcal{C}}) = \sum_{\mathsf{C}\in\mathcal{C}} \mathrm{cost}(\mathsf{C}, i_{\mathsf{C}}) = \sum_{\mathsf{C}\in\mathcal{C}} \sum_{j\in\mathsf{C}} \mathrm{d}(i,j)^2$.

k-`medoids`

Input. (X, d), k.

Procedure.

Initialize cluster centroids $\{i_1, ..., i_k\} \subset \mathsf{X}_n$ arbitrarily.

While total cost decreases,

(a) For each $g \leqslant k$, for each $j \notin \{i_1, ..., i_k\}$ compute the cost with new medoids $\{i_1, ..., i_{g-1}, j, i_{g+1}, ...i_k\}$;

(b) Update to new medoids if total cost decreases.

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

A.2. Implementation of cluster-based inference

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.

A.2.1. Cluster-based methods (CCE, IM, CRS)

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 $k = 2, ..., k_{max} = \lceil n^{1/3} \rceil$ where n is the total sample size, 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 $U \sim \mathrm{N}(0, \Sigma)$,

$$P_{M_{[X,W]}}\widehat{U} \sim \mathrm{N}(0, \Sigma_{PMU}),$$

where

$$\Sigma_{PMU} = P_{M_{[X,W]}} M_{[X,W]} \Sigma M_{[X,W]} P'_{M_{[X,W]}}.$$

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 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 $\widehat{\text{Err}}_{\text{Type-I}}(\text{IM}(a), k)$, $\widehat{\text{Err}}_{\text{Type-I}}(\text{CRS}(a), k)$, $\widehat{\text{Err}}_{\text{Type-I}}(\text{CCE}(a), k)$ by conducting the corresponding test $H_0 : \theta_0 = 0$ on each simulated dataset from *Step 3*. Set $\widehat{\alpha}_{\text{IM},k}$, $\widehat{\alpha}_{\text{IM},k}$, $\widehat{\alpha}_{\text{IM},k}$ to be the largest value $a \in [0, \alpha]$ such that $\widehat{\text{Err}}_{\text{Type-I}}(\text{IM}(a), k) \leqslant \alpha$, $\widehat{\text{Err}}_{\text{Type-I}}(\text{CRS}(a), k) \leqslant \alpha$, $\widehat{\text{Err}}_{\text{Type-I}}(\text{CCE}(a), k) \leqslant \alpha$.

For each partition $k = 2, ..., k_{\max} = 8$, compute simulated average Type II error rate $\widehat{\text{Err}}_{\text{Type-I}}(\text{IM}(\widehat{a}_{\text{IM},k}), k)$, $\widehat{\text{Err}}_{\text{Type-I}}(\text{CRS}(\widehat{a}_{\text{CRS},k}), k)$, $\widehat{\text{Err}}_{\text{Type-I}}(\text{CCE}(\widehat{a}_{\text{CCE},k}), k)$ using $\widehat{a}_{\text{IM},\mathcal{C}}, \widehat{a}_{\text{CRS},\mathcal{C}}, \widehat{a}_{\text{CCE},\mathcal{C}}$ and by conducting the corresponding test $H_0 : \theta_0 = \theta^{\text{alt}}$ for $\theta^{\text{alt}} \in \{-10/\sqrt{n}, -9/\sqrt{n}, \ldots \ \ldots, -1/\sqrt{n}, 1/\sqrt{n}, 2/\sqrt{n}, \ldots, 10/\sqrt{n}\}$ on each simulated dataset from *Step 3* and averaging the Type 2 error obtained at each hypothesized value.

Step 5. Solution to (2.9). Set $\widehat{k}_{\text{CCE}}$, $\widehat{k}_{\text{IM}}$, $\widehat{k}_{\text{CRS}}$ corresponding to the highest simulated average power. Set $\widehat{T}_{\text{IM}(\alpha)} = T_{\text{IM}(\widehat{a}_{\text{IM},k}),\mathcal{C}^{\widehat{k}_{\text{IM}}}}, \widehat{T}_{\text{CRS}(\alpha)} = T_{\text{CRS}(\widehat{a}_{\text{CRS},\mathcal{C}}),\mathcal{C}^{\widehat{k}_{CRS}}}, \widehat{T}_{\text{CCE}(\alpha)} = \widehat{T}_{\text{CCE}(\widehat{a}_{\text{CCE},\mathcal{C}}),\mathcal{C}^{\widehat{k}_{\text{CCE}}}}$.

A.2.2. *Spatial-HAC* (SK)

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 $K_1 = K_2 = \left[\sqrt{0.4N_{\text{pan}}^{2/3}}\right]$ where N_{pan} is the number of locations.

References

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. 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). Berry–Esseen 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.

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.

SUPPLEMENT TO: "Inference for Dependent Data with Learned Clusters"

Authors: Jianfei Cao, Christian Hansen, Damian Kozbur, Lucciano Villacorta
August 2, 2021

This supplement contains proofs as well as additional simulation results.

S.1. Proofs of Theorems 1–7

Before proving Theorems 1–7, the basic structural results about sequences of uniformly Ahlfors spaces are gathered. The first preliminary result is due to Assoud (1977) and ensures the existence of certain embedding of metric spaces of bounded doubling into Euclidean spaces.

Theorem 8 (Assoud)**.** *Let* (X, d) *be an arbitrary (not necessarily finite) metric space with finite doubling dimension* $dim_{2\times}(\mathsf{X}) < \infty$*. Let* $\varepsilon \in (0, 1)$*. Then there exists an* L*-bi-Lipschitz map* $(\mathsf{X}, \mathrm{d}^{\nu}) \to \mathbb{R}^r$ *for some* L, ν *which depend only on* ε *and* $dim_{2\times}(\mathsf{X})$*.*

Ahlfors regularity of X implies constant doubling as a result of standard arguments involving covering numbers of metric spaces. Thus the next proposition follows.

Proposition 1. *Suppose that* $\mathsf{X} \in \mathbf{Ahlf}_{C,\delta}$*. Then the doubling dimension of* X *is bounded by* $dim_{2\times}(\mathsf{X}) \leqslant \delta \log_2(3C^2)$*. Therefore, there exists an* L*-bi-Lipschitz map* $(\mathsf{X}, \mathrm{d}^{3/4}) \to \mathbb{R}^{\nu}$ *where* ν *and the Lipschitz constant* L *depend only on* C, δ*.*

S.1.1. Proof of Theorem 1

Proof. In the case that X_n do not embed isometrically, let L be the bi-Lipschitz constant from the maps $(\mathsf{X}_n, \mathrm{d}_n^{1-1/4}) \to \tilde{\mathsf{X}}_n \subseteq \mathbb{R}^{\nu}$. The array $\{\{\zeta_i\}_{i\in\mathsf{X}_n}\}_{n=1}^{\infty}$ indexed on X_n yields an array $\{\{\tilde{\zeta}_i\}_{i\in\tilde{\mathsf{X}}_n}\}_{n=1}^{\infty}$ indexed on $\tilde{\mathsf{X}}_n$. It is sufficient to check the conditions of Corollary 1 in Jenish and Prucha (2009) for this new process. Apply the same set array of constants c_i to $\tilde{\zeta}_i$. Assumption 1 in Jenish and Prucha (2009) is satisfied by the fact that distances are at least ρ_0 in X_n for some $\rho_0 > 0$ by Ahlfors regularity. Note that L depends only on C, δ and in particular, does not change with n. Then $\forall i, j$, $\tilde{\mathrm{d}}_n(i, j)$ is also bounded away from zero by a constant which does not depend on n. Condition 2(ii) is identical to Equation 3 in Jenish and Prucha (2009).

The next conditions in Jenish and Prucha (2009) are mixing conditions. To verify these, let $\tilde{\alpha}^{\text{mix}}_{g,l,n}(r)$ and $\bar{\tilde{\alpha}}^{\text{mix}}_{g,l,n}(r)$ denote the corresponding mixing coefficients for $\tilde{\zeta}_i$ over $\tilde{\mathsf{X}}_n$. Note that $\tilde{\mathrm{d}}(\mathsf{U}, \mathsf{V}) \geqslant r \Rightarrow \mathrm{d}(\mathsf{U}, \mathsf{V}) \geqslant L^{-1}\nu^{-1/2}r^{\frac{3}{4}}$. Let $c = L^{-1}\nu^{-1/2}$. Then $\bar{\tilde{\alpha}}^{\text{mix}}_{g,l}(r) \leqslant \bar{\alpha}^{\text{mix}}_{g,l}(cr^{3/4})$. To verify Equation 4 in Jenish and Prucha (2009), it is sufficient to show that $\sum_{m=1}^{\infty} \bar{\tilde{\alpha}}^{\text{mix}}_{1,1}(m) m^{\nu\times\frac{\mu+2}{\mu}-1} < \infty$. Note that $\bar{\tilde{\alpha}}^{\text{mix}}_{1,1}(m)$ is nonincreasing and defined for nonnegative real m and $m^{\nu\times\frac{\mu+2}{\mu}-1}$ is a polynomial in m. Thus, the above summation is bounded up to a constant by the corresponding integral $\int_{m=0}^{\infty} \bar{\bar{\alpha}}_{1,1}^{\text{mix}}(m) m^{\nu \times \frac{\mu+2}{\mu} - 1} dm$. Substituting $\bar{\bar{\alpha}}_{1,1}^{\text{mix}}(m) \leqslant \bar{\alpha}_{1,1}^{\text{mix}}(cm^{3/4})$ and a standard calculus change of variables $m' = cm^{3/4}$, $dm' = \frac{3}{4} cm^{-1/4} dm$ shows that it is sufficient to verify $\int_{m'=0}^{\infty} \bar{\alpha}_{1,1}^{\text{mix}}(m') m'^{\frac{4}{3}(\nu \times \frac{\mu+2}{\mu} - 1)} m'^{1/3} dm' < \infty$. This integral is in turn bounded by a constant times the summation

$$\sum_{m'=1}^{\infty} \bar{\alpha}_{1,1}^{\text{mix}}(m') m'^{\frac{4}{3}\nu \times \frac{\mu+2}{\mu} - 1} = \sum_{m'=1}^{\infty} \bar{\alpha}_{1,1}^{\text{mix}}(m') m'^{\nu \times \frac{\mu+2}{\mu} - 1} < \infty$$

which is assumed to be finite under Condition 2(iii), thus verifying Equation 4 in Jenish and Prucha (2009). Using similar arguments, Condition 2(iv) implies Assumption 4(2) in Jenish and Prucha (2009). Next, Condition 2(v) implies that

$$\bar{\bar{\alpha}}_{1,\infty}^{\text{mix}}(m) \leqslant \bar{\alpha}_{1,\infty}^{\text{mix}}(cm^{3/4}) = O((cm^{3/4})^{\frac{4}{3}(\nu - \mu)}) = O(m^{\nu - \mu})$$

thus verifying Assumption 4(3) in Jenish and Prucha (2009). Finally, Condition 2(vi) implies Assumption 5 in Jenish and Prucha (2009). This verifies the assumptions of Corollary 1 in Jenish and Prucha (2009).

□

S.1.2. Proof of Theorem 2

Proof. Theorem 2 is proven in the case $|\mathcal{C}| = 2$. In this case for $\mathsf{C} \subseteq \mathsf{X}_n$, take $\mathsf{D} = \mathsf{X}_n \setminus \mathsf{C}$. The general case follows by applying the same arguments, and by retracing boundaries when necessary. By arguments in Bester et al. (2011) which appeared previously in Jenish and Prucha (2009) and Bolthausen (1982), $\sigma(\mathsf{C})^{-1}\sigma(\mathsf{D})^{-1}n = O(1)$ and

$$\left|\text{cov}\Big(\sigma(\mathsf{C})^{-1} \sum_{i \in \mathsf{C}} \zeta_j, \sigma(\mathsf{D})^{-1} \sum_{j \in \mathsf{D}} \zeta_j\Big)\right| \leqslant \sigma(\mathsf{C})^{-1}\sigma(\mathsf{D})^{-1} \sum_{(i,j) \in \mathsf{C} \times \mathsf{D}} \bar{\alpha}_{1,1}(\lceil \mathrm{d}_n(i,j) \rceil)^{\mu/(2+\mu)}$$

$$= \sigma(\mathsf{C})^{-1}\sigma(\mathsf{D})^{-1} \sum_{g=1}^{\infty} |\{(i,j) \in \mathsf{C} \times \mathsf{D} : g - 1 \leqslant \mathrm{d}_n(i,j) < g\}| \bar{\alpha}_{1,1}(g)^{\mu/(2+\mu)}.$$

Next, to proceed in the current setting, note that by Condition 1,

$$|\{(i,j) \in \mathsf{C} \times \mathsf{D} : g - 1 \leqslant \mathrm{d}_n(i,j) < g\}| \leqslant |\{(i,j) \in \mathsf{X} \times \mathsf{X} : g - 1 \leqslant \mathrm{d}_n(i,j) < g\}| \leqslant nCg^{\delta}.$$

By Condition 3, it also follows that

$$\max_{g \leqslant r} |\{(i,j) \in \mathsf{C} \times \mathsf{D} : g - 1 \leqslant \mathrm{d}_n(i,j) < g\}| \leqslant o(n).$$

Then the original covariance is bounded by

$$\sigma(\mathsf{C})^{-1}\sigma(\mathsf{D})^{-1} \left[\sum_{g=1}^{r} o(n)\bar{\alpha}_{1,1}(g) + \sum_{m=r}^{\infty} nCg^{\delta}\bar{\alpha}_{1,1}(g)\right]$$

The first term satisfies $\sigma(\mathsf{C})^{-1}\sigma_{\mathsf{D}}^{-1} \sum_{g=1}^{r} o(n)\bar{\alpha}_{1,1}(g) \to 0$ by $\sum_{g=1}^{r} o(n)\bar{\alpha}_{1,1}(g) \leqslant \sum_{g=1}^{\infty} o(n)\bar{\alpha}_{1,1}(g) < \infty$ and $\sigma_{\mathsf{C}}^{-1}\sigma(\mathsf{D})^{-1}o(n) \to 0$. The second term satisfies

$\sigma_{\mathsf{C}}^{-1}\sigma_{\mathsf{D}}^{-1}\sum_{m=r}^{\infty} nCg^{\delta}\bar{\alpha}_{1,1}(g) \to 0$ by $n\sigma(\mathsf{C})^{-1}\sigma(\mathsf{D})^{-1} = O(1)$, $\sum_{m=1}^{\infty} nCg^{\delta}\bar{\alpha}_{1,1}(g) < \infty$ and $r \to \infty \implies \sum_{m=r}^{\infty} Cg^{\delta}\bar{\alpha}_{1,1}(g) \to 0$. The proposition follows.

□

S.1.3. Proof of Theorem 3

Proof. Let $\bar{r}_n = \log n$. From this point on in this proof, n is excluded from notation when used as an index in a subscript and when it provides no added clarity. Consider two arbitrary (sequences of) points $i, j \in \mathsf{X}$. Let $M = \{l : |\mathrm{d}(i,l) - \mathrm{d}(j,l)| \leqslant \bar{r}\}$. Let M_0 be an $\bar{r}^2$-net of M. Note that by construction, if $|M| \neq o(n)$ then $|M_0| \neq o(n/\bar{r}^{2\delta})$. Let $A \subseteq [\frac{1}{2}, \frac{9}{10}]$ satisfy $|a - a'| \geqslant \bar{r}^3/\mathrm{d}(i,j)$ for each $a, a' \in A$. Take $|A| \geqslant \frac{4}{10}\frac{\mathrm{d}(i,j)}{2}/\bar{r}^3$. For $a \in A$, let M_a consist of interpolants l_a such that $|\mathrm{d}(i,l_a) - a\mathrm{d}(i,l)| \leqslant K$ and $|\mathrm{d}(l_a,l) - (1-a)\mathrm{d}(i,l)| \leqslant K$ for each $l \in M_0$. Then by trigonometry, $\cup_{a\in A} M_a$ is $3\bar{r}$-separated for n sufficiently large and contains $|A| \times |M_0|$ elements. This shown by constructing the line segments $\overline{\iota(i)\iota(l)}$ and $\overline{\iota(i)\iota(l')}$ where ι is the coarse isometry to Euclidean space E, which gives points u, u' belonging to the above constructed line segments with distances $d_{\mathsf{E}}(\iota(l_a), u)$, $d_{\mathsf{E}}(\iota(l_a'), u')$ bounded, implying that u, u' are sufficiently separated to yield the claim. By $3\bar{r}$-separation and $|\mathsf{B}_{\bar{r}}(l_a)| \geqslant C^{-1}\bar{r}^{\delta}$, it follows that

$$\left|\bigcup_{a\in A, l_a \in M_a} \mathsf{B}_{\bar{r}}(l_a)\right| \geqslant |A| \times |M_0| C^{-1}\bar{r}^{\delta} \geqslant \frac{4}{20}\frac{\mathrm{d}(i,j)}{\bar{r}^3}|M_0|C^{-1}\bar{r}^{\delta}.$$

As the left hand side above is $\leqslant n$, it follows that the two statements (1.) $d(i,j) \geqslant \bar{r}^{4+\delta}$ for n sufficiently large, and (2.) $|M_0| \neq o(n/\bar{r}^{2\delta})$ cannot be true simultaneously.

Next, it is shown that k-`medoids` terminates with medoids $i_1, ..., i_k$ such that $\mathrm{d}(i_g, i_l) \geqslant (\log n)^{4+\delta}$ for n sufficiently large. This implies the small boundaries statement of Theorem 3, as a violation of this statement would require that M corresponding to some pair i_g, i_l satisfies $M \neq o(n)$, subsequently implying $M_0 \neq o(n/\bar{r}^{2\delta})$, violating the above dicotomy.

Again for contradiction, suppose there is a sequence $\ell \to 0$ such that for infinitely many n, there are two clusters $\mathsf{C}_1, \mathsf{C}_2$ with medoids i_1, i_2 satisfying $\mathrm{d}(i_1, i_2) < \ell n^{1/\delta}$. By the pigeonhole principal, there must be a cluster C_3 with n/k members. Then $diam(\mathsf{C}_3)$ must be at least $C^{-1}(n/k)^{1/\delta}$. Let x_3 be the corresponding medoid. Then there must be $i_3' \in \mathsf{C}_3$ such that $\mathrm{d}(i_3, i_3') \geqslant \frac{1}{4}C^{-1}(n/k)^{1/\delta}$ and $\mathrm{d}(i_3', i_g) \geqslant \frac{1}{4}C^{-1}(n/k)^{1/\delta}$ for any other medoid i_g. Then consider the update in the partitioned medoid algorithm given by $i_2 \leftarrow i_3'$. This update is cost reducing for n sufficiently large. To see this, note that for elements, $i \in \mathsf{B}_{\frac{1}{4}C^{-1}(n/k)^{1/\delta}}(i_3')$ the total cost reduction from being reassigned from a medoid centered around i_2 to a medoid centered around i_3' is at least $|\mathsf{B}_{\frac{1}{4}C^{-1}(n/k)^{1/\delta}}(i_3')|\frac{1}{4}C^{-1}(n/k)^{1/\delta} \geqslant \frac{1}{4}C^{-1}(n/k)^{1/\delta}C^{-1}(\frac{1}{4}C^{-1}(n/k)^{1/\delta})^{\delta}$. The total cost increase from reassigning elements in C_2 to C_1 is at most $\ell n^{1/\delta}|\mathsf{C}_2| \leqslant \ell n^{1/\delta} n$. The difference between the above two quantities is a lower bound on the cost reduction for the update. Comparing the above to quantities for n sufficiently large, the k-`medoids` algorithm could not have stopped at a step with $\mathrm{d}(i_1, i_2) < \ell n^{1/\delta}$ giving the desired contradiction.

Finally, note that $\mathrm{d}(i_g, i_l) \geqslant \ell n^{1/\delta}$ for all medoids i_g, i_l, some ℓ bounded uniformly away from 0, and for n sufficiently large implies the balanced clusters condition after applying Ahlfors regularity. □

S.1.4. Proof of Theorem 4

Proof. Fix $k \leqslant k_{\max}$ and consider the subsequence of $\mathcal{C}_{\rightarrow}$ for which $|\mathcal{C}_n| = k$. Let V_k be a k-dimensional real vector space with distinguished basis indexed by $\mathcal{C}_{\infty,k}$. Choose bijections $\mathcal{C}_n \rightarrow \mathcal{C}_{\infty,k}$ and induce linear maps $\ell_n : \mathbb{R}^{\mathcal{C}_n} \rightarrow V \rightarrow$ such that the image of S_n in the direction of any vector corresponding to an element of $\mathcal{C}_\infty$ has variance 1. Applying Theorems 1 and 2 to S_n, the Cramèr-Wold device to $\ell(S_n)$ and the Almost-Sure Representation theorem (Theorem 2.19 in van der Vaart (1998)) then gives the $\breve{S}^*_{\infty,k}$ and $\breve{S}_{\mathcal{C}_n}$ such that $\|\breve{S}^*_{\infty,k}\|_2 \rightarrow 0$, $\tilde{\mathrm{P}}_k$- a.s. for some $\tilde{\mathrm{P}}_k$ for all n in the subsequence. Set for n in the subsequence $\tilde{S}^*_{\mathcal{C}_n} = \ell_n^{-1}\breve{S}^*_{\infty,k}$ and $\tilde{S}_{\mathcal{C}_n} = \ell_n^{-1}\breve{S}_{\mathcal{C}_n}$. Repeat this construction for all k and take all $\tilde{\mathrm{P}}_k$ to be a common $\tilde{\mathrm{P}}$. □

S.1.5. Proof of Theorem 5

Proof. The proof follows directly from the definition of $\widehat{\theta}_{\mathsf{C}}$ along with standard arguments and the application of Theorems 1–4. □

S.2. Proof of Theorem 6

Proof. Define the function ψ by $\psi(S) = \mathbf{1}_{|t(S)|>t_{1-\alpha/2,k-1}}$, where S has k components. Let $\tilde{\mathrm{E}}$ be expectation with respect to the measure $\tilde{\mathrm{P}}$. Note that for a partition $\mathcal{C}$, $T_{\mathrm{IM}(\alpha),\mathcal{C}} = \psi(S_{\mathcal{C}})$. By construction of $\tilde{S}^*_{\mathcal{C}}$ and Theorem 1 of Ibragimov and Müller (2010), for any n and any $\mathcal{C} \in \mathscr{C}_n$, $\tilde{\mathrm{E}}[\psi(\tilde{S}^*_{\mathcal{C}})] \leq \alpha$.

Next, show $\sup_{\mathcal{C}\in\mathscr{C}_n} |\tilde{\mathrm{E}}[\psi(\tilde{S}_{\mathcal{C}})] - \tilde{\mathrm{E}}[\psi(\tilde{S}^*_{\mathcal{C}})]| \rightarrow 0$. Define the events $\mathcal{E}_{1,\mathcal{C}} = \{|t(\tilde{S}^*_{\mathcal{C}}) - t_{1-\alpha/2,|\mathcal{C}|-1}| > b_{\mathcal{C}}\}$, $\mathcal{E}_{2,\mathcal{C}} = \{|\mathrm{sd}(\tilde{S}^*_{\mathcal{C}})| > b'_{\mathcal{C}}\}$ for $b_{\mathcal{C}}, b'_{\mathcal{C}} \in \mathbb{R}$ specified below. Note that by the upper and lower bounds on the variances of $\tilde{S}^*_{\mathcal{C}}$ in Condition 7, and continuity properties of $t(\cdot)$ and $\mathrm{sd}(\cdot)$ one can select a common positive function, denoted M which is independent of $\mathcal{C}$, such that $\lim_{b\rightarrow 0} M(b, b') = \lim_{b'\rightarrow 0} M(b, b') = 0$ and such that $\tilde{\mathrm{P}}(\mathcal{E}_{1,\mathcal{C}} \cap \mathcal{E}_{2,\mathcal{C}}) \geqslant 1 - M(b_{\mathcal{C}}, b'_{\mathcal{C}})$ whenever $b_{\mathcal{C}}, b'_{\mathcal{C}}$ are sufficiently small. Then on $\mathcal{E}_{1,\mathcal{C}}$, a sufficient condition for $\psi(\tilde{S}_{\mathcal{C}}) = \psi(\tilde{S}^*_{\mathcal{C}})$ is that $|t(\tilde{S}_{\mathcal{C}}) - t(\tilde{S}^*_{\mathcal{C}})| < b_{\mathcal{C}}$. By $\sup_{\mathcal{C}\in\mathscr{C}_n} \|\tilde{S}^*_{\mathcal{C}} - \tilde{S}_{\mathcal{C}}\|_2 \rightarrow 0$, and by continuity properties of $t(\cdot)$ outside of $\mathcal{E}_{2,\mathcal{C}}$, there are $b_{\mathcal{C}} < b_n \rightarrow 0, b'_{\mathcal{C}} < b'_n \rightarrow 0$ such that $|t(\tilde{S}_{\mathcal{C}}) - t(\tilde{S}^*_{\mathcal{C}})| < b_{\mathcal{C}}$ is satisfied on $\mathcal{E}_{1,\mathcal{C}} \cap \mathcal{E}_{2,\mathcal{C}}$. Therefore, $|\tilde{\mathrm{E}}[\psi(\tilde{S}_{\mathcal{C}})] - \tilde{\mathrm{E}}[\psi(\tilde{S}^*_{\mathcal{C}})]| \leqslant M(b_n, b'_n)$, for all $\mathcal{C} \in \mathscr{C}_n$ from which the above desired uniform limit result follows.

Finally, $\mathrm{P}_0(\widehat{T}_{\mathrm{IM}(\alpha),n} =$ Reject $) = \mathrm{P}_0(T_{\mathrm{IM}(\widehat{\alpha}),\widehat{\mathcal{C}}} =$ Reject $) \leqslant \mathrm{P}_0(T_{\mathrm{IM}(\alpha),\widehat{\mathcal{C}}} =$ Reject $)$ $= \sum_{\mathcal{C}\in\mathscr{C}_n} \mathrm{P}_0(T_{\mathrm{IM}(\alpha),\mathcal{C}} =$ Reject $|\widehat{\mathcal{C}} = \mathcal{C})\mathrm{P}_0(\widehat{\mathcal{C}} = \mathcal{C})$. Bound the above expression by $o(1) + \sum_{\mathcal{C}\in\mathscr{C}_{0,n}} \mathrm{P}_0(T_{\mathrm{IM}(\alpha),\mathcal{C}} =$ Reject $)\mathrm{P}_0(\widehat{\mathcal{C}} = \mathcal{C}) \leqslant o(1) + \alpha$ by using both Condition 6(i) and Condition 6(ii) and the previous bounds. This concludes the proof of Theorem 1. □

S.3. Proof of Theorem 7

Before giving the proof, note that the following more general condition is used in place of Condition 7(iii) in the main text.

Condition 7. *(iii)' Either*

a. $|\mathscr{C}_n|$ *and* $\max_{\mathcal{C}\in\mathscr{C}_n}|\mathcal{C}|$ *are each bounded by a constant independent of* n*; or*
b. For any sequence $\check{k}_n \to \infty$ *sufficiently slowly, the following hold. There is a sequence* $\check{\delta}_n \to 0$ *and sets* $\mathcal{A}_\mathcal{C} \subseteq \mathbb{R}^\mathcal{C}$ *for each* $\mathcal{C} \in \mathscr{C}_n$ *with* $|\mathcal{C}| > \check{k}_n$ *which are closed under the action of* $\mathcal{H}_\mathcal{C}$ *such that,* $\lim_{n\to\infty}\inf_{\mathcal{C}\in\mathscr{C}_n,|\mathcal{C}|>\check{k}}\tilde{\mathrm{P}}(\mathcal{A}_\mathcal{C}) = 1$*,* $w(\cdot)$ *can be renormalized to have Lipschitz constant 1 respect to the Euclidean norm on all* $\mathcal{A}_\mathcal{C}$*, and it holds that* $\sup_{\mathcal{C}\in\mathscr{C}_n}\sup_{r\in\mathbb{R}}\mathrm{P}_\mathcal{C}(|r - w(h\tilde{S}^*_\mathcal{C})| < \check{\delta}_n) \to 0$ *where* $\mathrm{P}_\mathcal{C}$ *be the uniform probability measure over* $h \in \mathcal{H}_\mathcal{C}$*. Finally,* $|\{\mathcal{C} \in \mathscr{C}_n : |\mathcal{C}| \leqslant \check{k}_n\}|$ *depends only* $\check{k}_n$*.*

The next proposition verifies that the test statistic, w, defined in the text above and used for the empirical example and simulation study satisfy the anticoncentration requirement of Condition 7(iii)'.

Proposition 2. *For* $S \in \mathbb{R}^\mathcal{C}$*, let* $w(S) = |t(S)|$*. Let* $S^*_\mathcal{C}$ *be a Gaussian random variables with independent components with variances bounded from below and above (as in Condition 7(i)). Suppose that* $\mathscr{C}_n = \{\mathcal{C}^{(2)}, ..., \mathcal{C}^{(k_{\max,n})}\}$ *where* $k_{\max,n}$ *may be an arbitrary function of* n*. Then Condition 7(iii)'*a *is satisfied if* $k_{\max,n}$ *is bounded and Condition 7(iii)'*b *is satisfied if* $k_{\max,n} \to \infty$*.*

The proof below is given assuming Condition 7(iii)'.

Proof. Define functions $\phi(S)$ and $\tilde{\phi}(S, U)$ for $S \in \mathbb{R}^\mathcal{C}$ and $U \in \mathbb{R}$ according to

$$\phi(S) = \begin{cases} 1, & \text{if } w(S) > w^{(j_\alpha)}(S) \\ \tilde{a}(S), & \text{if } w(S) = w^{(j_\alpha)}(S) \\ 0, & \text{if } w(S) < w^{(j_\alpha)}(S) \end{cases}$$

and define $\tilde{\phi}(S, U) = \phi(S)$ if $W(S) \neq W^{(j_\alpha)}(S)$ and $\mathbf{1}_{U<a(S)}$ otherwise.

Without loss of generality, assume that $\tilde{\Omega}$ has a uniformly distributed random variable $\tilde{U} \in [0, 1]$ independent of all previously defined random variables defined on it. By Theorem 2.1 in Canay et al. (2017), using the fact that $h(\tilde{S}^*_\mathcal{C}) =_d \tilde{S}^*_\mathcal{C}$ for all h by the fact that h acts as component-wise multiplication by signs, it follows that $\mathrm{E}_0[\phi(\tilde{S}^*_\mathcal{C})] = \alpha$. Note,

$$\mathrm{E}_0[\phi(S_\mathcal{C})] - \alpha = \tilde{\mathrm{E}}[\tilde{\phi}(\tilde{S}_\mathcal{C}, \tilde{U}) - \tilde{\phi}(\tilde{S}^*_\mathcal{C}, \tilde{U})].$$

Therefore, it is sufficient to show $\limsup_{n\to\infty}\sup_{\mathcal{C}\in\mathscr{C}_n}\tilde{\mathrm{E}}[\tilde{\phi}(\tilde{S}_\mathcal{C}, \tilde{U}) - \tilde{\phi}(\tilde{S}^*_\mathcal{C}, \tilde{U})] = 0$. Consider an integer $\check{k}$, and decompose $\mathscr{C}_n = \mathscr{C}_{n,\check{k}} \cup \mathscr{C}^c_{n,\check{k}}$ where $\mathscr{C}_{n,\check{k}} = \{\mathcal{C} \in \mathscr{C}_n : |\mathcal{C}| \leqslant \check{k}\}$ and $\mathscr{C}^c_{n,\check{k}} = \{\mathcal{C} \in \mathscr{C}_n : |\mathcal{C}| > \check{k}\}$, which are handled separately. For each $\mathcal{C} \in \mathscr{C}_{n,\check{k}}$, let $\mathcal{E}_\mathcal{C}$ be the event in which the order statistics of $\{w(h\tilde{S}_\mathcal{C}) : h \in \mathcal{H}_\mathcal{C}\}$ and $\{w(h\tilde{S}^*_\mathcal{C}) : h \in \mathcal{H}_\mathcal{C}\}$

correspond to the same transformations $h^{(1)}, \dots, h^{(|\mathcal{H}_\mathcal{C}|)}$. Then the previous expression can be controlled by

$$\begin{aligned}|\tilde{\mathrm{E}}[\tilde{\phi}(\tilde{S}_\mathcal{C}, \tilde{U}) - \tilde{\phi}(\tilde{S}^*_\mathcal{C}, \tilde{U})]| &= |\tilde{\mathrm{E}}[\tilde{\phi}(\tilde{S}_\mathcal{C}, \tilde{U})\mathbf{1}_{\mathcal{E}_\mathcal{C}} + \tilde{\phi}(\tilde{S}_\mathcal{C}, \tilde{U})\mathbf{1}_{\mathcal{E}^c_\mathcal{C}} - \tilde{\phi}(\tilde{S}^*_\mathcal{C}, \tilde{U})\mathbf{1}_{\mathcal{E}_\mathcal{C}} - \tilde{\phi}(\tilde{S}^*_\mathcal{C}, \tilde{U})\mathbf{1}_{\mathcal{E}^c_\mathcal{C}}]| \\ &= |\tilde{\mathrm{E}}[\tilde{\phi}(\tilde{S}_\mathcal{C}, \tilde{U})\mathbf{1}_{\mathcal{E}^c_\mathcal{C}} - \tilde{\phi}(\tilde{S}^*_\mathcal{C}, \tilde{U})\mathbf{1}_{\mathcal{E}^c_\mathcal{C}}]| \\ &= |\tilde{\mathrm{E}}[(\tilde{\phi}(\tilde{S}_\mathcal{C}, \tilde{U}) - \tilde{\phi}(\tilde{S}^*_\mathcal{C}, \tilde{U}))\mathbf{1}_{\mathcal{E}^c_\mathcal{C}}]| \leq 2\tilde{\mathrm{E}}[\mathbf{1}_{\mathcal{E}^c_\mathcal{C}}].\end{aligned}$$

Furthermore, $\sup_{\mathcal{C}\in\mathscr{C}_{n,\check{k}}} \tilde{\mathrm{E}}[\mathbf{1}_{\mathcal{E}^c_\mathcal{C}}]$ can be controlled using Fatou's Lemma by

$$\limsup_{n\to\infty} \sup_{\mathcal{C}\in\mathscr{C}_{n,\check{k}}} \tilde{\mathrm{E}}[\mathbf{1}_{\mathcal{E}^c_\mathcal{C}}] \leq \limsup_{n\to\infty} \tilde{\mathrm{E}}[\sup_{\mathcal{C}\in\mathscr{C}_{n,\check{k}}} \mathbf{1}_{\mathcal{E}^c_\mathcal{C}}] \leqslant \tilde{\mathrm{E}}[\limsup_{n\to\infty} \sup_{\mathcal{C}\in\mathscr{C}_{n,\check{k}}} \mathbf{1}_{\mathcal{E}^c_\mathcal{C}}].$$

To further bound the right-hand side of the above expression, consider any fixed $\omega \in \tilde{\Omega}$ which satisfies the conditions $\sup_{\mathcal{C}\in\mathscr{C}_n} \|\tilde{S}_\mathcal{C}(\omega) - \tilde{S}^*_\mathcal{C}(\omega)\|_2 \to 0$ and $\inf_{\mathcal{C}\in\mathscr{C}_{n,\check{k}}} \inf_{h\nsim h'\in\mathcal{H}_\mathcal{C}} |w_\mathcal{C}(h(\tilde{S}^*_\mathcal{C})(\omega)) - w_\mathcal{C}(h'(\tilde{S}^*_\mathcal{C})(\omega))| > \delta_\omega > 0$ for some $\delta_\omega > 0$. Here, the equivalence $\sim$ is defined by $h \sim h'$ whenever $w \circ h = w \circ h'$. By Condition 7, the set of such $\omega \in \tilde{\Omega}$ has $\tilde{\mathrm{P}}$-measure 1. Note, more explicitly, choosing such $\delta_\omega > 0$ is possible due to (1) the continuity of the function w and the continuity of the action of h outside a set of $\tilde{P}$-measure 0, and (2) the fact that $|\mathscr{C}_{n,\check{k}}|$ depends only on $\check{k}$.

Let $h^{(1)}_\mathcal{C}(\omega), \dots, h^{(|\mathcal{H}_\mathcal{C}|)}_\mathcal{C}(\omega)$ be such that $w(h^{(1)}_\mathcal{C}(\omega)\tilde{S}^*_\mathcal{C}(\omega)) \leqslant \dots \leqslant w(h^{(|\mathcal{H}_\mathcal{C}|)}_\mathcal{C}(\omega)\tilde{S}^*_\mathcal{C}(\omega))$. Then the expression

$$\begin{aligned}w(h^{(j+1)}_{n,\mathcal{C}}(\omega)\tilde{S}_\mathcal{C}(\omega)) - w(h^{(j)}_{n,\mathcal{C}}(\omega)\tilde{S}_\mathcal{C}(\omega)) = {}& w(h^{(j+1)}_{n,\mathcal{C}}(\omega)\tilde{S}_\mathcal{C}(\omega)) - w(h^{(j+1)}_{n,\mathcal{C}}(\omega)\tilde{S}^*_\mathcal{C}(\omega)) \\ &+ w(h^{(j+1)}_{n,\mathcal{C}}(\omega)\tilde{S}^*_{n,\mathcal{C}}(\omega)) - w(h^{(j)}_{n,\mathcal{C}}(\omega)\tilde{S}^*_{n,\mathcal{C}}(\omega)) \\ &+ w(h^{(j)}_{n,\mathcal{C}}(\omega)\tilde{S}^*_{n,\mathcal{C}}(\omega)) - w(h^{(j)}_{n,\mathcal{C}}(\omega)\tilde{S}_\mathcal{C}(\omega))\end{aligned}$$

is nonnegative and is furthermore strictly positive for at least one j unless $h \sim h'$ for all $h, h' \in \mathcal{H}_\mathcal{C}$ in which case the theorem holds trivially. Positivity of the above expression in the case of non-equivalent-under-$\sim$ action follows from noting that the first and the third terms are smaller than $\delta_\omega/2$ in absolute value for n sufficiently large, while the second term is greater than δ_ω.

The claim $\limsup_{n\to\infty} \sup_{\mathcal{C}\in\mathscr{C}_n} \mathbf{1}_{\mathcal{E}^c_\mathcal{C}}(\omega) = 0$ $\tilde{\mathrm{P}}$-a.s. follows. Thus, $\limsup_{n\to\infty} \sup_{\mathcal{C}\in\mathscr{C}_{n,\check{k}}} \tilde{\mathrm{E}}[\mathbf{1}_{\mathcal{E}^c_\mathcal{C}}] = 0$. Therefore, there is a sequence $\check{k}_n \to \infty$ sufficiently slowly such that $\limsup_{n\to\infty} \sup_{\mathcal{C}\in\mathscr{C}_{n,\check{k}_n}} \tilde{\mathrm{E}}[\mathbf{1}_{\mathcal{E}^c_\mathcal{C}}] = 0$.

Next consider the partitions in the complement $\mathcal{C} \in \mathscr{C}^c_{\check{k}_n}$. Let $\mathcal{A}_s \subseteq \mathbb{R}^\mathcal{C}$ be of the form $\mathcal{A}_S = \{hS : h \in \mathcal{H}_\mathcal{C}\}$ for some $S \in \mathbb{R}^\mathcal{C}$. Then w^{j_α} is constant over $\mathcal{A}_S$ and thus $w^{j_\alpha}(\mathcal{A}_S)$ is well defined. Then for any $\delta > 0$,

$$\tilde{\mathrm{P}}(|w^{j_\alpha}(\mathcal{A}_S) - w(\tilde{S}^*_\mathcal{C})| < \delta | \tilde{S}^*_\mathcal{C} \in \mathcal{A}_S) \leqslant \sup_{r\in\mathbb{R}} \tilde{\mathrm{P}}(|r - w(\tilde{S}^*_\mathcal{C})| < \delta | \tilde{S}^*_\mathcal{C} \in \mathcal{A}_S).$$

Let $\mathrm{P}_\mathcal{C}$ be the uniform probability over $h \in \mathcal{H}_\mathcal{C}$. Then the above expression is bounded by

$$p_{\mathcal{C},S}(\delta) := \sup_{r\in\mathbb{R}} \mathrm{P}_\mathcal{C}(|r - w(hS)| < \delta).$$

Next, for any $\mathcal{A}_\mathcal{C}$ which is a disjoint union of sets of the form $\mathcal{A}_S$. Then

$$\tilde{\mathrm{P}}(|w^{j_\alpha}(\tilde{S}^*_\mathcal{C}) - w(\tilde{S}^*_\mathcal{C})| \geqslant \delta) \geqslant \tilde{\mathrm{P}}(|w^{j_\alpha}(\tilde{S}^*_\mathcal{C}) - w(\tilde{S}^*_\mathcal{C})| \geqslant \delta | \tilde{S}^*_\mathcal{C} \in \mathcal{A}_\mathcal{C})\tilde{\mathrm{P}}(\mathcal{A}_\mathcal{C})$$
$$\geqslant (1 - \sup p_{\mathcal{C},S}(\delta))\tilde{\mathrm{P}}(\mathcal{A}_\mathcal{C})$$

where the supremum runs over S indexing $\mathcal{A}_S \subseteq \mathcal{A}_\mathcal{C}$.

Note that a sufficient condition for $\tilde{\phi}(\tilde{S}^*_\mathcal{C}, \tilde{U}) = \tilde{\phi}(\tilde{S}_\mathcal{C}, \tilde{U})$ is that $\mathrm{sign}(w^{j_\alpha}(\tilde{S}^*_\mathcal{C}) - w(\tilde{S}^*_\mathcal{C})) = \mathrm{sign}(w^{j_\alpha}(\tilde{S}_\mathcal{C}) - w(\tilde{S}_\mathcal{C}))$. A further sufficient condition for this is that for some $\delta > 0$, $|w^{j_\alpha}(\tilde{S}^*_\mathcal{C}) - w(\tilde{S}^*_\mathcal{C})| \geqslant \delta$, $|w^{j_\alpha}(\tilde{S}^*_\mathcal{C}) - w^{j_\alpha}(\tilde{S}_\mathcal{C})| < \delta/2$, and $|w(\tilde{S}^*_\mathcal{C}) - w(\tilde{S}_\mathcal{C})| < \delta/2$. By Condition 7(iii)', w has Lipschitz constant 1 with respect to Euclidean norm on $\mathbb{R}^\mathcal{C}$ on a suitably chosen sequence of events of probability approaching 1. Then $|w^{j_\alpha}(\tilde{S}^*_\mathcal{C}) - w^{j_\alpha}(\tilde{S}_\mathcal{C})| < \delta/2$ whenever $\|\tilde{S}^*_\mathcal{C} - \tilde{S}_\mathcal{C}\|_2 < \delta/2$ in which case it also holds that $|w(\tilde{S}^*_\mathcal{C}) - w(\tilde{S}_\mathcal{C})| < \delta/2$. A sequence $\check{\delta}_n$ may be chosen such that $\check{\delta}_n \to 0$ sufficiently slowly such that for n sufficiently large, each of the above three inequalities may be achieved with common bounds on an event with probability $1 - o_{\tilde{\mathrm{P}}}(1)$.

Finally, using the same reasoning as at the conclusion of Theorem 6, and invoking Condition 6, $\mathrm{P}_0(\widehat{T}_{\mathrm{CRS}(\alpha),n} = \text{ Reject }) = o(1) + \sum_{\mathcal{C} \in \mathscr{C}_{0,n}} \mathrm{P}_0(T_{\mathrm{CRS}(\alpha),\mathcal{C}} = \text{ Reject })\mathrm{P}_0(\widehat{\mathcal{C}} = \mathcal{C}) \leqslant o(1) + \alpha$. concluding the proof of Theorem 2. □

S.4. Proof of Proposition 2

Proof. Fix $S \in \mathbb{R}^\mathcal{C}$. Let D be a constant chosen later, and consider a rescaled version of w defined by $w(S) = |\mathcal{C}|^{-1/2}|\bar{S}|/(D\mathrm{sd}(S))$. Next bound the following anticoncentration quantity:

$$\sup_{r \in \mathbb{R}} \mathrm{P}_\mathcal{C}(w(hS) \in [r, r+\delta]).$$

Because $\mathcal{H}_\mathcal{C}$ is finite, the range of the supremum above can be restricted and the above quantity is further reduced by

$$= \sup_{r \in [\min_{h \in \mathcal{H}_\mathcal{C}} w(hS), \max_{h \in \mathcal{H}_\mathcal{C}} w(hS)]} \mathrm{P}_\mathcal{C}(w(hS) \in [r, r+\delta])$$
$$\leqslant \sup_{r \in [\min_{h \in \mathcal{H}_\mathcal{C}} w(hS), \max_{h \in \mathcal{H}_\mathcal{C}} w(hS)]} \mathrm{P}_\mathcal{C}(|\bar{hS}| \in [\min_{h \in \mathcal{H}_\mathcal{C}} aD\mathrm{sd}(hS), \max_{h \in \mathcal{H}_\mathcal{C}} (a+\delta)D\mathrm{sd}(hS)])$$

Note that the length of the interval inside the $\mathrm{P}_\mathcal{C}$ never exceeds

$$\ell = \ell(\delta, D, S) := \sup_{r \in [\min_{h \in \mathcal{H}_\mathcal{C}} w(hS), \max_{h \in \mathcal{H}_\mathcal{C}} w(hS)]} \max_{h \in \mathcal{H}_\mathcal{C}} (r+\delta)D\mathrm{sd}(hS) - \min_{h \in \mathcal{H}_\mathcal{C}} rD\mathrm{sd}(hS).$$

Therefore, $\sup_{r \in \mathbb{R}} \mathrm{P}_\mathcal{C}(w(hS) \in [r, r+\delta]) \leqslant \sup_{r \in \mathbb{R}} \mathrm{P}_\mathcal{C}(|\mathcal{C}|^{-1/2}|\bar{hS}| \in [r, r+\ell])$. By considering both branches of the absolute value function, (which gives an additional factor of 2), and expressing the above quantity more explicitly in terms of components h_c, it follows that

$$\sup_{r \in \mathbb{R}} \mathrm{P}_\mathcal{C}(|r - |\mathcal{C}|^{-1/2}|\sum_{\mathsf{C} \in \mathcal{C}} h_\mathsf{C} S_\mathsf{C})|| < \ell) \leqslant 2 \sup_{r \in \mathbb{R}} \mathrm{P}_\mathcal{C}(|r - |\mathcal{C}|^{-1/2} \sum_{\mathsf{C} \in \mathcal{C}} h_\mathsf{C} S_\mathsf{C})| < \ell).$$

By Corollary 2.9 in Rudelson and Vershynin (2008), which states an anticoncentration bound for bernoulli sums, this is bounded by

$$2 \sup_{r \in \mathbb{R}} \mathrm{P}_{\mathcal{C}}(|r - |\mathcal{C}|^{-1/2} \sum_{\mathsf{C} \in \mathcal{C}} h_{\mathsf{C}} S_{\mathsf{C}}| < \ell) \leqslant 2 \left[\sqrt{\frac{2}{\pi}} \frac{\ell}{\| |\mathcal{C}|^{-1/2} S \|_2} + C_1 B \left(\frac{\| |\mathcal{C}|^{-1/2} S \|_3}{\| |\mathcal{C}|^{-1/2} S \|_2} \right)^3 \right]$$

where C_1 is an absolute constant, and $B = 1$ is the third moment of a Bernoulli random variable.

For a sequence $\check{k} \to \infty$, Gaussian concentration properties coupled with the assumed bounds on the variances of the components of $\tilde{S}^*_{\mathcal{C}}$ imply that there is are sets $\mathcal{A}_{\mathcal{C}}$ closed under the action of $\mathcal{H}_{\mathcal{C}}$ with $\tilde{\mathrm{P}}(\mathcal{A}_{\mathcal{C}}) \to_{\check{k} \to \infty} 1$ and a fixed choice D which makes w Lipschitz with constant 1 on all of $\mathcal{A}_{\mathcal{C}}$, such that ℓ is bounded by a fixed constant times δ for all S in all $\mathcal{A}_{\mathcal{C}}$. Gaussian concentration properties for $\| |\mathcal{C}|^{-1/2} \tilde{S}^*_{\mathcal{C}} \|_2$ and $(\| |\mathcal{C}|^{-1/2} \tilde{S}^*_{\mathcal{C}} \|_3 / \| |\mathcal{C}|^{-1/2} \tilde{S}^*_{\mathcal{C}} \|_2)^3$ then also allow the construction of $\check{\delta}_n \to 0$ such that the right hand of the above expression side also $\to 0$ provided that $\check{k}$ grows sufficiently slowly.

□

S.5. *Additional Simulation Results*

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

The same settings as in Section 4 are considered; though here we provide results with number of locations given by $N_{\mathrm{pan}} = 820$ in addition to $N_{\mathrm{pan}} = 205$. In the $N_{\mathrm{pan}} = 820$ 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 4. The maximal number of groups to be considered in CCE, IM, and CRS is chosen to be $G_{\max} = 12$. 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 $N_{\mathrm{pan}} = 820$ and $N_{\mathrm{pan}} = 205$ cases.

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

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 2. 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 7
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 2. 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 8
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 2. 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 9
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 2. 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 10
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 2. 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. The rows "$\widehat{k}$ frequency" is the frequency of a particular k achieving the highest simulated power among candidate k's in the setting.

TABLE 11
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 2. 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. The rows "$\widehat{k}$ frequency" is the frequency of a particular k achieving the highest simulated power among candidate k's in the setting.

TABLE 12
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 2. 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. The rows "$\widehat{k}$ frequency" is the frequency of a particular k achieving the highest simulated power among candidate k's in the setting.

TABLE 13
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 2. 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. The rows "$\widehat{k}$ frequency" is the frequency of a particular k achieving the highest simulated power among candidate k's in the setting.

TABLE 14
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 2. 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 15
Summary: OLS- SAR ($N_{\text{pan}} = 820$)

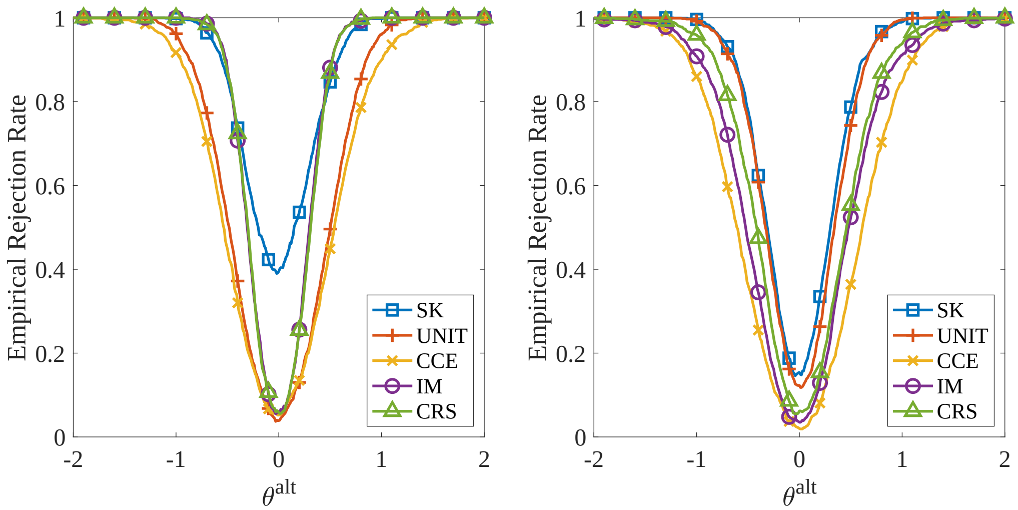
Method	Estim. Mean	Estim. RMSE	Size	Power			
				-1	-0.5	0.5	1
SK	0.003	0.229	0.154	0.996	0.780	0.787	0.990
UNIT-U	0.003	0.229	0.405	0.999	0.909	0.903	1.000
UNIT	0.003	0.229	0.123	0.992	0.756	0.743	0.995
CCE	0.003	0.229	0.020	0.860	0.363	0.364	0.848
IM	-0.013	0.249	0.035	0.908	0.465	0.524	0.911
CRS	-0.013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

Notes: Simulation results for estimation in the design described in Section 2. 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 16
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 2. 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).



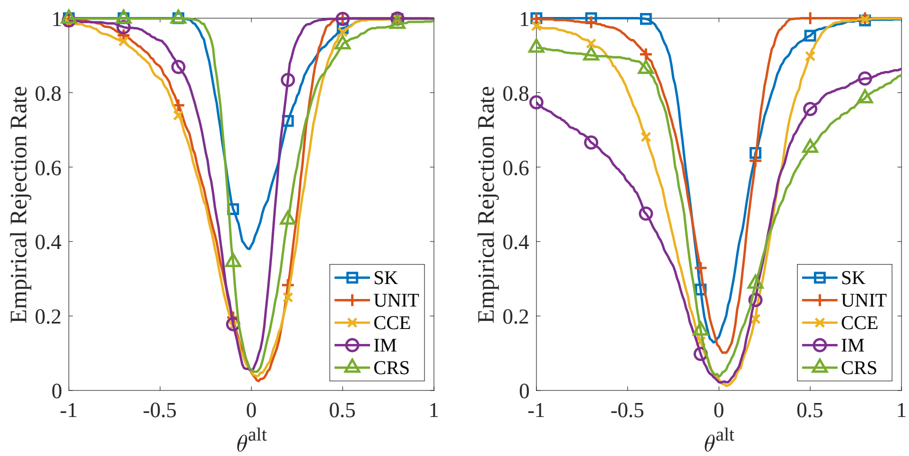
(a) BASELINE (b) SAR

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

TABLE 17
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 2. 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 5: IV power curves ($N_{\text{pan}} = 820$).

TABLE 18
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 2. 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. The rows "$\widehat{k}$ frequency" is the frequency of a particular k achieving the highest simulated power among candidate k's in the setting.

Table 19

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 2. 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. The rows "$\widehat{k}$ frequency" is the frequency of a particular k achieving the highest simulated power among candidate k's in the setting.

TABLE 20
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 2. 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. The rows "$\widehat{k}$ frequency" is the frequency of a particular k achieving the highest simulated power among candidate k's in the setting.

TABLE 21

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 2. 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. The rows "$\widehat{k}$ frequency" is the frequency of a particular k achieving the highest simulated power among candidate k's in the setting.