Inference for Dependent Data with Cluster Learning

(PRELIMINARY AND INCOMPLETE)

Jianfei Cao

The University of Chicago
Booth School of Business
5807 S. Woodlawn, Chicago, IL 60637, USA
e-mail: jcao0@chicagobooth.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 proposes a cluster-based inferential procedure. Observations are grouped into clusters which are learned using a unsupervised learning algorithm given a dissimilarity measure. We consider a set of cluster-based inference procedures on the learned clusters. We give conditions under which our procedure asymptotically attains correct size. We illustrate the finite sample validity and apply our procedure to an empirical example.

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

[*] First version: October 2018. This version is of January 1, 2020. 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.

1. Introduction

Much empirical work in economics relies on observational data that seems likely to be dependent. For example, empirical work often considers serial dependence in the analysis of time series, and cross-sectional or spatial dependence in cross-sectional and panel data. Failing to account for dependence generally leads to invalid inferential statements. This paper studies inferential procedures that are robust to weak dependence.

To conduct valid inference, many solutions exist in the literature such as HAR variance estimation. Among all, one class of inferential methods, typically called cluster- or group-based inference procedures, are particularly popular among empirical researchers. Cluster-based methods work from taking a clustering structure $\mathcal{C} = \{\mathsf{C}_g\}_{g=1}^G$ which partitions the observations. For those methods, inference proceeds by ignoring covariance between observations that fall in different clusters. Conceptually, those methods are exactly analogous to truncation or downweighting of autocovariances in traditional time series HAR. In this paper, we consider group-based inference with small or moderate number of large groups. Much work has been done in this literature. For example, Bester, Conley, Hansen (2008) [5] (hereafter BCH) present a simple method for conducting inference using the cluster covariance matrix estimator (CCE) under asymptotics that treat the number of groups as fixed and the number of observations within a group as large. Ibragimov and Müller (2006) [14] (IM) provides a formal treatment of the famous Fama-Macbeth procedure by Fama and MacBeth (1973) [10], focusing upon properties of t-tests using point estimates from all clusters. Canay, Romano, Shaikh (2017) [6] (CRS) develops a theory of randomization tests under an approximate symmetry assumption, i.e., when the classes of transformations applied to the original data do not change the distribution. Those methods have been shown to be very robust in simulations.

However, one practical issue remaining is that the choice of clustering structures is typically ad-hoc and the chosen clustering often does not align with the "natural structure". For example, in the case of spatial dependence, it is not obvious how the researcher should cut the map, because many options seem to be plausible. In this case, we essentially have to pin down two tuning parameters, the number of groups G, and the partition of observations given G. The goal of this paper is to provide practical, data-driven methods to help make these choices. We propose to use clustering methods from the machine learning literature to partition observations given number of groups, and to use simulation to choose number of groups based on inferential properties.

To make sense of the usage of clustering methods, we consider a case where a measure of dissimilarity is available to the researcher, and this measure is informative about the underlying correlation structure in the sense that dependence between quantities goes away when the distance between them is large. There are many ways to construct groups from measures of dissimilarity. Such problems fall under "unsupervised learning." See [13] for a general review. Within unsupervised learning, there are several commonly used clustering algorithms. In this paper we will focus on k-medoids.

In the paper we propose a three-step inferential procedure: (i) Given observations locations, generate sequence of partitions $\{\mathcal{C}^{(G)}\}_{G=2}^{\bar{G}}$; (ii) estimate a parametric covariance model using scores and simulate inferential objects (e.g. size/power) to choose partition structure $\mathcal{C}^*$; (iii) perform partition-based inference using $\mathcal{C}^*$. Although our methods are readily generalizable, in this paper we focus on k-medoids in Step 1, a damped cosine covariance model in Step 2, and IM or CRS inference in Step 3. We first consider the estimation of means and then generalize it to the OLS and IV cases.

Our main theoretical results concern the behavior of cluster-based inference such as IM and CRS with fixed number of groups, each of which is determined by a data-driven clustering method. Our regularity conditions involve moment and mixing rate 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. Thus our assumptions are no stronger than those routinely made with the plug-in HAC approach or in the CRS approach.

We also give two important properties of clustering algorithm that can be of great use in inference and are new to the unsupervised learning literature. Namely, we show that the clustering structures produced by k-medoids are balanced and have small boundaries. We give formal definitions of those properties and show their usage in cluster-based methods.

Moreover, our theoretical results contribute to the growing literature on inference with spatial data; that is, data in which dependence is indexed in more than one dimension. Examples of papers in this literature are Conley (1996, 1999) [8], [9], Kelejian and Prucha (1999, 2001) [18], [19], Lee (2004, 2007a, 2007b) [23] [24] [25], and Jenish and Prucha (2007) [15]. We note that analysis of spatially dependent data is not a trivial extension of results for scalar-indexed (time series). Complications arise due to such concerns as set boundaries being of large order of magnitude relative to set sizes and the number of potential neighbors of any particular point increasing rapidly with the dimension in which dependence increases. We provide formal conditions under which inference based on the CCE remains valid in very general settings.

We present simulation evidence on the performance of our estimator in the panel data context in both OLS and IV cases. The simulations illustrate that inference procedures that ignore cross-sectional dependence, such as clustering based on only location, are severely size distorted: Even modest serial or spatial dependence needs to be accounted for in order to produce reliable inference. Moreover, it demonstrates that plug-in spatial-HAC inference procedures may suffer from substantial size distortions. However, provided the number of groups is small and correspondingly the number of observations per group is large, our proposed test procedure has actual size close to nominal size and good power properties. Finally, we apply our procedure to Condra, Long and Shaver (2018) [7] to investigate the effect of insurgent attack on voter turnouts.

The remainder of the paper is organized as follows. Section 2 presents our proposed inferential procedure. Section 3 studies the conditions under which cluster-based estimators are asymptotically normal and asymptotically independent. Section 4 studies the properties of clustering algorithms and gives formal results on our procedure. Section 5 presents simulation evidence regarding the performance of the proposed procedures and related methods. Section 6 applies our method to an empirical example. Section 7 concludes. Proofs are relegated to the appendix.

2. Methodology: Inference with Unsupervised Cluster Learning

In this section we present our proposed inferential procedure. Throughout, we consider the hypothesis of the type $H_0 : \beta_0 = 0$, where $\beta_0 \in \mathbb{R}$. The most important input into the inferential procedure is a dissimilarity measure d_n which captures key aspects of dependence of observables across observations. In the following sections, we will assume in a precise sense that the degree of dependence between observations decreases sufficiently quickly, as dissimilarity increases. There are many possible sources for generating dissimilarity measures. Leading examples include geographic distance, notions of economic distance ([8], [9], see also measures of economic distance constructed from input-output tables), displacement in time, and combinations of the above. As it is described in Algorithm 1 we use d_n to produce a fixed number of clusters by applying an unsupervised clustering procedure to the data (more details below). Then, we use an inferential procedure over fixed clusters, i.e., an inferential procedure which, if a good clustering of the data which captured relevant dependence in observations was known ex ante, could be applied to produce approximate inference.

The general procedure proposed by this paper can be described by three steps: (i) learn a finite sequence of clustering structures $\mathscr{C} = \{\mathcal{C}^{(2)}, \ldots, \mathcal{C}^{(\bar{G})}\}$ by some unsupervised learning method from a dissimilarity measure over the data, (ii) choose the optimal clustering $\mathcal{C}^* \in \mathscr{C}$ based on a correlation model and some criterion of size-power tradeoff, (iii) perform cluster-based hypothesis testing using $\mathcal{C}^*$. This procedure is described below and formally collected in Algorithm 1. Implementation details are in Appendix A.

Algorithm 1

Inputs. Data $\mathscr{D}_n = \{W_i\}_{i=1}^n$; an $n \times n$ dissimilarity measure over observations given by a metric space (X_n, d_n); a significance threshold α; a user specified upper bound $\bar{G}$ for the number of clusters.

Procedure.

1. Apply an unsupervised clustering using (X_n, d) and produce a sequence of clustering structures $\{\mathcal{C}^{(G)}\}_{G \in \mathcal{G}}$ with $\mathcal{G} = \{2, ..., \bar{G}\}$ where $G \leqslant \bar{G}$ is the number of clusters.
2. Simulate size and power (against interesting alternative or weighted average power against interesting family of alternatives) for each $\mathcal{C}^{(G)}$ and choose the one that minimizes loss over size and power. To achieve this, first estimate a (possibly misspecified) covariance structure of scores using some parametric model.
3. Apply CCE, CRS, or IM using the selected clustering.

Outputs. $T \in \{0, 1\}$.

2.1. Clustering

Given the dissimilarity measure d_n, we use k-medoids (KM) to produce a sequence of clustering structures $\{\mathcal{C}^{(2)}, \ldots, \mathcal{C}^{(\bar{G})}\}$.[1] k-medoids is a popular clustering technique which is related to k-means and in many cases produces very similar clustering results as k-means does. Because we deal with a dissimilarity which does not necessarily arise from a Euclidean space, k-means centroids are not necessarily defined. Instead assigning the geographic centers to be centroids of clusters as in k-means, k-medoids requires that cluster 'centers' be themselves elements of X_n. For a detailed description of k-medoids see Hastie, Tibshirani, and Friedman (2009) [13]. We reproduce a k-medoids algorithm here for convenience. Let the cost of a cluster C with center x be

$$\mathrm{cost}(C, x) = \sum_{x' \in C} d(x, x')^2.$$

The total cost is defined by

$$\text{total cost} = \sum_{C \in \mathcal{C}} \mathrm{cost}(C, x_C)$$

with x_C being the center of C. Then the algorithm of k-medoids is given by Algorithm 2.

Algorithm 2 k-medoids Clustering

```
Inputs. An $n \times n$ dissimilarity measure over a finite metric space $(\mathsf{X}_n, d_n)$; the number of groups $G$.
Procedure.
    1. Initialize cluster centroids $\{x_1, ..., x_G\} \subset \mathsf{X}_n$:
    2. While total cost decreases,
        a. for each $k \leqslant G$, for each $z \notin \{x_1, ..., x_G\}$ compute the cost with new medoids $\{x_1, ..., x_{k-1}, z, x_{k+1}, ...x_G\}$;
        b. assign new medoids and membership if the new set of medoids has lower total cost.
Outputs. A clustering structure $\mathcal{C}^{(G)}$ with $G$ groups.
```

Algorithm 2 produces one clustering structure each time. To produce a sequence of clusterings with different number of groups, we run k-medoids for each $G \in \{2, \ldots, \bar{G}\}$, where $\bar{G} = \lceil n^{1/3} \rceil$. The maximum number of clusters is chosen such that the size of each cluster is going to infinity.

2.2. Size-power tradeoff

To simulate size and power, we need an estimator for the covariance structure. This can be done by using QMLE to estimate a simple covariance model for scores s_i. In the example of linear regression, $s_i = \tilde{x}_i \widehat{u}_i / (n^{-1} \sum_j \tilde{x}_j \widehat{u}_j)$, where $\tilde{x}_i$ is typically the partialed-out parameter of interest and $\widehat{u}_i$ is the regression residual. We note that our results do not require the consistency of the estimated covariance function in order to control size. As a result,

[1] Other clustering algorithms apply as long as the resulting clustering structures satisfy Assumption C3. Results on a penalized version of hierarchical clustering is available on request.

misspecification in our model for dependence leads only to potential loss of power. We focus on a simple and intuitive specification with damped-cosine correlation structure. FOr example, in a spatial setting, we can consider the class of covariance functions parameterized by $b = (b_1, b_2, b_3)'$ and given by

$$\left\{ \text{cov}(s_i, s_j; b) = b_1 \exp\left(- \left(\frac{\|L_i - L_j\|_2}{b_2} \right) \right) \cos\left(\frac{\|L_i - L_j\|_2}{b_3} \right) \right\}_{b \in \mathbb{R}^3_+},$$

where L_i is the location of observation i. Each b gives rise to an implied covariance matrix Σ_b of $(s_1, \ldots, s_n)'$. Gaussian QMLE yields estimates

$$\widehat{b} \in \arg\min_{b \in \mathbb{R}^3_+} \det\left(\Sigma_b^{-1/2} \right) \exp\left(-\frac{1}{2} s' \Sigma_b^{-1} s \right).$$

We can then generate samples drawn from $N\left(0, \Sigma_{\widehat{b}}\right)$ or $N\left(b, \Sigma_{\widehat{b}}\right)$ and calculate the Type-I and Type-II error using the cluster-based testing procedure in Section 2.3 with the clustering $\mathcal{C}$. The framework can be readily extended to a panel setting, which will be illustrated in Appendix A.

2.3. Group-based hypothesis testing

In the development of this paper, we focus on the procedure of Ibragimov and Muller (2010, IM) and Canay Romano and Shaikh (2017, CRS) . Other examples of group-based procedure include the cluster covariance estimator as in Bester, Conley and Hansen (2011, CCE), whose performance is studied in the simulation section.

2.3.1. IM

The IM procedure by [14] uses properties of t-statistic of heterogeneous independent normal random variables. The idea is that when group-level estimates are asymptotically independent and normal, a t-test can be performed in order to control size under some assumptions.

2.3.2. CRS

CRS is an asymptotic analogue of a randomization testing procedure. In CRS, a randomization test is obtained by recomputing a test statistic over the set of transformations under which the distribution of the test statistic is invariant (e.g. permutation of the data). A full exposition of the CRS testing procedure is in Appendix A.3.2.

3. A Central Limit Theorem with a small number of groups

In this section, we give conditions under which group-level estimators are asymptotically normal and asymptotically independent. Our analysis is based on a simple condition over dissimilarity matrices, which is related to the notion of Ahlfors regularity. We give mixing conditions which are directly analogous to those in Jenis and Prucha (2009) [16]. We define a notion of cluster balance and boundary and show they are the key to the asymptotic properties of the group-level means.

We first consider a simple setting with $\{\{Z_{i,n}\}_{i=1}^{n}, \mathsf{X}_n\}_{n=1}^{\infty}$, where we observe scalar data $\{Z_{i,n}\}_{i=1}^{n}$ in the space X_n and are interested in the mean of $Z_{i,n}$. In Section 3.4 we extend to the OLS and IV case by relating the scores to $Z_{i,n}$.

3.1. Formal Conditions on an Increasing Sequence of Dissimilarity Measures

This section describes assumptions on increasing sequences of dissimilarity measures over which clustering will take place. We assume that the researcher has access to some set of dissimilarity measures which are denoted (X_n, d_n). We begin by making the assumption that the triangle inequality holds for each X_n. In other words, the dissimilarity measures are genuine metric spaces. In addition, they satisfy the following regularity condition.

Definition 1. Let $|\mathsf{X}|$ be the carnality of X. A finite metric space X is (C, δ)-finite-Ahlfors regular if $|\mathsf{X}| \vee C^{-1} r^{\delta} \leqslant |\mathbf{B}_{\mathsf{X},r}(x)| \leqslant C r^{\delta} \wedge 1$ for any $r > 0$, for any x in any X, where $\mathbf{B}_{\mathsf{X},r}(x)$ is the r-ball centered at x in the space X.

(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 (in which case $|\mathbf{B}_{\mathsf{X},r}(x)|$ is replaced by $\mu(\mathbf{B}_{,r}(x))$ and the conditions $|\mathsf{X}| \vee \ldots$ and $\ldots \wedge 1$ are dropped). This notion has several advantages, relative to assuming that X_n 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.

Not every dissimilarity measure X_n satisfying (C, δ)-finite-Ahlfors regularity admits an isometric embedding into $\mathbb{R}^r$ for some r. It is not even sufficient to guarantee that X admit a bi-Lipschitz embedding (defined so that the maximum distortion is bounded by a constant). However, the condition is strong enough to ensure that X can be "regularized." The derived space $(\mathsf{X}, d^{1-\varepsilon})$ 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 $\mathbb{R}^r$ with bounded distortion.

Throughout our analysis, we will refer to the following assumption. In particular, this assumption defines our spatial asymptotic frame.

Assumption C1. *(Ahlfors Regularity) The sequence of spaces X_n is a uniformly Ahlfors sequence of finite metric space with n elements.*

The utility of this notion is that gives enough structure to allow us to simultaneously (1) study the performance of several leading clustering techniques analytically, (2) derive dependent central limit theorems and laws of large numbers. It is also simple to express.

3.2. Mixing conditions

We first consider In this section, we list mixing conditions on the array of variables $\{\{Z_{i,n}\}_{i=1}^n, \mathsf{X}_n\}_{n=1}^\infty$.

Definition 2 (Mixing coefficients)**.** Let $\{\{Z_{i,n}\}_{i=1}^n, (\mathsf{X}_n, d_n)\}_{n=1}^\infty$ be an array of random variables on a probability space $(\Omega, \mathsf{F}, \mathrm{P})$ with spatial indeces given by X_n. Let A and B be two (sub-)σ-algebras of F and let $\alpha(\mathsf{A}, \mathsf{B}) = \sup\{|\mathrm{P}(A \cap B) - \mathrm{P}(A)\mathrm{P}(B)| : A \in \mathsf{A}, B \in \mathsf{B}\}$. For $U \subseteq \mathsf{X}_n$ let $\sigma_n(U) = \sigma(Z_{i,n}; i \in U)$, and let $\alpha_n(U,V) = \alpha(\sigma_n(U), \sigma_n(V))$. Let

$$\alpha_{k,l,n}(r) = \sup\left\{\alpha_n(U,V), |U| \leqslant k, |V| \leqslant l, d(U,V) \geqslant r\right\},$$
$$\bar{\alpha}_{k,l}(r) = \sup_n \alpha_{k,l,n(r)}.$$

Assumption C2. *(Mixing) Let $\{\{Z_{i,n}\}_{i=1}^n, (\mathsf{X}_n, d_n)\}_{n=1}^\infty$ be an array of random variables on a probability space $(\Omega, \mathsf{F}, \mathrm{P})$ with spatial indeces given by X_n satisfying Condition* **C1** *with Ahflors constants C, δ. Let $\nu = \frac{4}{3} \times 833\delta \log(3C^2)$. Alternatively, if (X_n, d_n) embed isometrically into $\mathbb{R}^N$ for some common N, then $\nu = N$ may be taken. Let $\sigma_n^2 = \mathrm{var}(\sum_{i=1}^n Z_{i,n})$. There exists an array of positive constants $\{\{c_{i,n}\}_{i=1}^n\}_{n=1}^\infty$ and a positive $\mu > 0$ such that*

(i) $\mathrm{E}[Z_{i,n}] = 0$.
(ii) $\lim_{k\to\infty} \sup_n \sup_{i\in\mathsf{X}_n} \mathrm{E}[|Z_{i,n}/c_{i,n}|^{2+\mu}\mathbf{1}_{|Z_{i,n}/c_{i,n}|>k}] = 0$.
(iii) $\sum_{m=1}^\infty \bar{\alpha}_{1,1}(m) m^{\nu\times\frac{\mu+2}{\mu}-1} < \infty$.
(iv) $\sum_{m=1}^\infty m^{\nu-1}\bar{\alpha}_{k,l}(m) < \infty$ *for* $k + l \leqslant 4$.
(v) $\bar{\alpha}_{1,\infty}(m) = O(m^{-\nu-\frac{4}{3}\mu})$.
(vi) $\inf_n |\mathsf{X}_n|^{-1}(\max_{i\in\mathsf{X}_n} c_{i,n}^{-2})\sigma_n^2 > 0$.

The conditions are similar to those given for the mixing central limit theorem in Jenis and Prucha (2012) [17], but with abstract index sets identified with X_n. By contrast, Jenis and Prucha (2009) require X_n be a possibly uneven lattice in a finite dimensional Euclidean space with a minimum separation between all points. In our formulation, appearance of the ambient dimension of the Euclidean space is in all cases replaced by ν. Note that our Condition **C2** automatically implies a minimum separation between points. Note also that in the case that X_n embed isometrically into $\mathbb{R}^\nu$, the only difference between our conditions and the conditions for Corollary 1 of JP are that we require $\bar{\alpha}_{1,\infty}(m) = O(m^{-\nu-\frac{4}{3}\mu})$ rather than the slightly weaker $\bar{\alpha}_{1,\infty}(m) = O(m^{-\nu-\mu})$.

Proposition 1. *Suppose that conditions* **C1** *and* **C2** *hold for $\{\{Z_{i,n}\}_{i=1}^n, (\mathsf{X}_n, d_n)\}_{n=1}^\infty$. Then*

$$\sigma_n^{-1}\sum_{i=1}^n Z_{i,n} \xrightarrow{d} \mathbf{N}(0,1).$$

The proposition can be proven by regularizing X_n using the snowflake construction described above with $\varepsilon = 1/4$, on which the JP central limit theorem can then be applied.

The proof is carried out in detail in the appendix. The same argument can be carried out more generally for other $\varepsilon \in (0, 1/2)$, however we reference the single $\varepsilon = 1/4$ snowflake for added concreteness.

3.3. Balance and Small Common Boundary Conditions

The key property which allows inference in BCH is that cluster sizes are balanced and that their boundaries are small. In this section, we formalize this requirement and then show that both unsupervised clustering methods satisfy the desired property.

Assumption C3. *A sequence of clusterings $\{\mathcal{C}_n\}_{n\geq 2}$ on X_n is asymptotically balanced with small boundaries if:*

(i) (Balanced) The ratio of minimial cluster size to n is bounded away from zero uniformly, i.e.,

$$\liminf_{n\to\infty} \min_{\mathsf{C}\in\mathcal{C}_n} \frac{|\mathsf{C}|}{n} > 0.$$

(ii) (Small Boundaries) There is a $\bar{r} = \bar{r}(n) \to \infty$ so that

$$\max_{\mathsf{C}\in\mathcal{C}_n} |\{x \in \mathsf{C} : d(x, \mathsf{X} \setminus \mathsf{C}) \leqslant \bar{r}\}| = o(n).$$

This definition differs slightly from the definition of small boundary given in BCH. In particular, BCH leverage the fact their 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 $\mathsf{X} \setminus \mathsf{C}$. BCH assume that their given spatial clusters are contiguous and use that fact to bound the number of higher order neighbors from C to $\mathsf{X} \setminus \mathsf{C}$. In this context, there is no available definition of first order neighbor since X_n can be irregular (even non-Euclidean). As a result, we work instead with an asymptotic notion of boundary which entails a sequence $\bar{r}$ which allows boundaries to widen as $n \to \infty$. A high-level implication of having asymptotically balanced with small boundaries clusterings is the following proposition.

Proposition 2. *Suppose that $\mathsf{X}_n, Z_{i,n}$, and $\mathcal{C}_n$ satisfy* **C1**, **C2**, *and* **C3**. *Consider any $\mathsf{C}_n, \mathsf{D}_n \in \mathcal{C}_n$ for each n with $\mathsf{C}_n \neq \mathsf{D}_n$. Let $\sigma^2_{n,\mathsf{C}_n} = \mathrm{var}\left(\sum_{i\in\mathsf{C}_n} Z_{i,n}\right)$ and $\sigma^2_{n,\mathsf{D}_n} = \mathrm{var}\left(\sum_{i\in\mathsf{D}_n} Z_{i,n}\right)$. Then*

$$\mathrm{cov}\left(\sigma^{-1}_{n,\mathsf{C}_n} \sum_{i\in\mathsf{C}_n} Z_{i,n},\ \sigma^{-1}_{n,\mathsf{D}_n} \sum_{i\in\mathsf{D}_n} Z_{i,n}\right) \to 0.$$

The proof is related but not identical to the argument BCH. Whereas BCH count points in "shells" around the boundaries of clusters, our arguments instead rely on the doubling structure implied by the fact that X_n are Ahfolrs regular. Both our argument and BCH leverage a bound on covariances $\mathrm{cov}(Z_{i,n}, Z_{j,n})$ for sufficiently far spatial as implied by the mixing conditions stated in **C2**.

The condition **C3** acting as a high-level assumption on clusterings will also be sufficient for asymptotically valid inference in the proposed procedure. In the next section we verify **C3** for various clusterings under the assumption that X_n satisfy **C1**.

3.4. Extension to OLS and IV

Previous propositions show the asymptotic normality and diminishing across-group correlation for *means*. Now we extend our results to OLS and IV estimators.

Consider a linear model where $\forall i \in \mathsf{X}_n$,

$$y_{i,n} = x_{i,n}'\beta_{0,n} + u_{i,n}$$

and

$$\mathrm{E}[w_{i,n}u_{i,n}] = 0,$$

for scalars $y_{i,n}$ and $u_{i,n}$, and K-dimensional vectors $x_{i,n}$ and $w_{i,n}$. We suppress the subscription n for notation simplicity, i.e., $x_i = x_{i,n}$. Note that everything is a function of n. The IV estimator for β_0 within some set $\mathsf{C} \subset \mathsf{X}_n$ is given by

$$\widehat{\beta}_{n,\mathsf{C}} = \left(\frac{1}{|\mathsf{C}|}\sum_{i\in\mathsf{C}} w_i x_i'\right)^{-1}\left(\frac{1}{|\mathsf{C}|}\sum_{i\in\mathsf{C}} w_i y_i\right).$$

The OLS model is a special case where $w_i = x_i$.

Let $\{\mathcal{C}_n\}_{n\geq 1}$ be a sequence of clusterings with G groups. For some n, let $\mathcal{C}_n = \{\mathsf{C}_1, \ldots, \mathsf{C}_G\}$. For some $\mathsf{C} \in \mathcal{C}_n$, define $V_{n,\mathsf{C}} = Q_{n,\mathsf{C}}^{-1}\Omega_{n,\mathsf{C}}(Q_{n,\mathsf{C}}^{-1})'$, where

$$Q_{n,\mathsf{C}} = \mathrm{E}\left[\frac{1}{|\mathsf{C}|}\sum_{i\in\mathsf{C}} w_i x_i'\right]$$

and

$$\Omega_{n,\mathsf{C}} = \mathrm{var}\left[\frac{1}{|\mathsf{C}|}\sum_{i\in\mathsf{C}} w_i u_i\right].$$

Let $V_n = \mathsf{Diag}(V_{n,\mathsf{C}_1}, \ldots, V_{n,\mathsf{C}_G})$, i.e., V_n is an $(GK)\times(GK)$ block-diagonal matrix with the i-th diagonal block being V_{n,C_i}. Let

$$S_n = \begin{pmatrix} \sqrt{|\mathsf{C}_1|}(\widehat{\beta}_{n,\mathsf{C}_1} - \beta_0) \\ \vdots \\ \sqrt{|\mathsf{C}_G|}(\widehat{\beta}_{n,\mathsf{C}_G} - \beta_0) \end{pmatrix}$$

and

$$s_n = \begin{pmatrix} \frac{1}{\sqrt{|\mathsf{C}_1|}}\sum_{i\in\mathsf{C}_1} w_i u_i \\ \vdots \\ \frac{1}{\sqrt{|\mathsf{C}_G|}}\sum_{i\in\mathsf{C}_G} w_i u_i \end{pmatrix}.$$

Define $\eta_i = w_i x_i' - \mathrm{E}[|\mathsf{C}|^{-1} \sum_{i \in \mathsf{C}} w_i x_i']$ and

$$q_n = \begin{pmatrix} \frac{1}{\sqrt{|\mathsf{C}_1|}} \sum_{i \in \mathsf{C}_1} \eta_i \\ \vdots \\ \frac{1}{\sqrt{|\mathsf{C}_G|}} \sum_{i \in \mathsf{C}_G} \eta_i \end{pmatrix}.$$

Let $\eta_{i,j}$ and $q_{n,j}$ be the j-th column of η_i and q_n, respectively.

In Definition 2, we redefine $\sigma_n(U) = \sigma(\{y_{i,n}, x_{i,n}, w_{i,n}, u_{i,n}\}; i \in U)$.

Assumption C2*. Let $\{\{y_{i,n}, x_{i,n}, w_{i,n}, u_{i,n}\}_{i=1}^n, (\mathsf{X}_n, d_n)\}_{n=1}^\infty$ be an array of random elements on a probability space $(\Omega, \mathsf{F}, \mathrm{P})$ with spatial indeces given by X_n satisfying Condition **C1** with Ahflors constants C, δ. Let $\nu = \frac{4}{3} \times 833\delta \log(3C^2)$. Alternatively, if (X_n, d_n) embed isometrically into $\mathbb{R}^N$ for some common N, then $\nu = N$ may be taken. Let $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ be the minimal and maximal eigenvalue of a matrix, respectively. There exists an array of positive constants $\{\{c_{i,n}\}_{i=1}^n\}_{n=1}^\infty$ and a positive $\mu > 0$ such that

(i) $\mathrm{E}[w_{i,n} u_{i,n}] = 0$.

(ii) For some $z_i = w_i u_i$ or $q_{i,j}$, $\forall j$,

$$\lim_{k \to \infty} \sup_n \sup_{i \in \mathsf{X}_n} \mathrm{E}[\|n^{-1/2} z_i / c_{i,n}\|^{2+\mu} \mathbb{1}\{\|n^{-1/2} z_i / c_{i,n}\| > k\}] = 0.$$

(iii) $\sum_{m=1}^\infty \bar{\alpha}_{1,1}(m) m^{\nu \times \frac{\mu+2}{\mu} - 1} < \infty$.

(iv) $\sum_{m=1}^\infty m^{\nu-1} \bar{\alpha}_{k,l}(m) < \infty \ for \ k + l \leqslant 4$.

(v) $\bar{\alpha}_{1,\infty}(m) = O(m^{-\nu - \frac{4}{3}\mu})$.

(vi) $\inf_n n|\mathsf{X}_n|^{-1}(\max_{i \in \mathsf{X}_n} c_{i,n}^{-2}) > 0$.

(vii) $\inf_n \lambda_{\min}(\mathrm{var}(s_n)) > 0$ and $\mathrm{var}(s_n) = O(1)$.

(viii) $\inf_n \lambda_{\min}(\mathrm{var}(q_{n,j})) > 0$ and $\mathrm{var}(q_{n,j}) = O(1)$, $\forall j$.

Proposition 3. *(Asymptotic normality) Under the model, C1, C2*, C3 imply asymptotic normality*

$$V_n^{-1/2} S_n \xrightarrow{d} N(0, I_{GK}).$$

4. Group-based Testing with Learned Clustering

In this section, we formally describe the clustering procedure that yields the desired clustering properties. Together, these ingredients give the main theorem which verifies asymptotic validity of the inferential procedure described above.

4.1. Clustering Procedures and their Propoerties

We investigate k-medoids in simulations and in the empirical application below. We show that the small boundary and balance conditions are satisfied under regularity conditions. The next proposition derives relevant properties of this version of k-medoids algorithm.

Proposition 4. *Assume condition* **C2** *holds for* X_n. *Assume the following additional convexity condition holds. There is a constant* K *independent of* n *such that (1)* X_n *is* K*-coursely isometric*[2] *to a subset of a Euclidean space with dimension* $N = N(K)$, *and (2) for any two point* $x, y \in \mathsf{X}_n$ *and any* $a \in [0,1]$ *there is an interpolant* $z \in \mathsf{X}_n$ *such that* $|d(x,z) - ad(x,y)| \leqslant K$ *and* $|d(y,z) - (1-a)d(x,y)| \leqslant K$. *Then the* k*-medoids algorithm described in the text satisfies* **C3**.

4.2. Analysis of Group-based Inference

In this section, we provide formal results on the asymptotic validity of the proposed inferential procedures with unsupervised cluster learning.

Given some clustering structure $\mathcal{C} = \{\mathsf{C}_1, \ldots, \mathsf{C}_G\}$, some scalar parameter θ_0, and its within-group estimators $\{\widehat{\theta}_{n,\mathsf{C}_g}\}_{g=1}^G$, define

$$S_{n,\mathcal{C}} = \begin{pmatrix} \sqrt{|\mathsf{C}_1|}(\widehat{\theta}_{n,\mathsf{C}_1} - \theta_0) \\ \vdots \\ \sqrt{|\mathsf{C}_G|}(\widehat{\theta}_{n,\mathsf{C}_G} - \theta_0) \end{pmatrix}.$$

For example, θ_0 can the an entry of β_0 in the IV model and $\widehat{\theta}_{n,\mathsf{C}}$ is the corresponding entry of the IV estimator using only observations in C.

Given some $\mathsf{C} \in \mathcal{C} = (\mathsf{C}_1, \ldots, \mathsf{C}_G)$, let $\tau_{n,\mathsf{C}}^2 = \mathrm{var}(\sqrt{|\mathsf{C}|}(\widehat{\theta}_{n,\mathsf{C}} - \theta_0))$ be the variance of some entry of $S_{n,\mathcal{C}}$. Let $R_G^* \sim N(0, I_G)$ be G-dimensional standard joint normal random vector with independent entries. Note that R_G^* is not a function of n. For $R_G^* = (r_1, \ldots, r_G)$, define $S_{n,\mathcal{C}}^* = (\tau_{n,\mathsf{C}_1} r_1, \ldots, \tau_{n,\mathsf{C}_G} r_G)$, i.e., $S_{n,\mathcal{C}}^*$ is obtained by multiplying each entry of R_G^* by the standard deviation of the corresponding entry in $S_{n,\mathcal{C}}$.

4.2.1. IM

Now we give regularity assumptions for IM and the formal theorem:

Assumption R1. (Regularity assumptions for IM)

(i) The significance level $\alpha \leq 2\Phi(-\sqrt{3}) = 0.08326\ldots$, where Φ is the cumulative distribution function of a standard normal random variable.

(ii) With probability one,

$$\limsup_{n\to\infty} \sup_{\mathcal{C}\in\mathscr{C}_n} |\bar{S}_{n,\mathcal{C}}^*| < \infty,$$

$$\liminf_{n\to\infty} \inf_{\mathcal{C}\in\mathscr{C}_n} |\mathrm{se}(S_{n,\mathcal{C}}^*)| > 0,$$

and

$$\limsup_{n\to\infty} \sup_{\mathcal{C}\in\mathscr{C}_n} |\mathrm{se}(S_{n,\mathcal{C}}^*)| < \infty.$$

[2] $f : (\mathsf{Y}, d_{\mathsf{Y}}) \to (\mathsf{Z}, \mathsf{d}_{\mathsf{Z}})$ is a K-course isometry if $d_{\mathsf{Z}}(f(y), f(y')) - K \leqslant d_{\mathsf{Y}}(y, y') \leqslant d_{\mathsf{Z}}(f(y), f(y')) + K$.

(iii) With probability one,

$$\liminf_{n\to\infty} \inf_{\mathcal{C}\in\mathscr{C}_n} ||t(S^*_{n,\mathcal{C}})| - \mathrm{cv}_G(\alpha)| > 0,$$

where $G = |\mathcal{C}|$.

Theorem 1. *Let* $\mathsf{X}_n, Z_{i,n}, \mathscr{C}_n$ *satisfy* **C1**, **C2**, *and* **C3**. *Suppose* **R2** *holds. Suppose* $Z_{i,n}$ *are observable functions of data* $\mathscr{D}_n \sim P_n$. *Suppose further that* ψ *is the IM test result and is based on* $S_{n,\mathcal{C}} = S_{n,\mathcal{C}}(\mathscr{D}_n)$ *for some* $\mathcal{C} \in \mathscr{C}_n$. *Then under Algorithm 1,*

$$\sup_{\mathcal{C}\in\mathscr{C}_n} (\mathrm{E}_{P_n}[\psi(S_{n,\mathcal{C}})] - \alpha)_+ \to 0.$$

4.2.2. CRS

We formalize assumptions for CRS to be a valid inferential procedure by the following set of conditions:

Assumption R2. (Regularity assumptions for CRS)

(i) $\sup_n \sup_{\mathsf{C}\subset\mathsf{X}_n} \tau^2_{n,\mathsf{C}} < \infty$.

(ii) With probability one,

$$\liminf_{n\to\infty} \inf_{\mathcal{C}\in\mathscr{C}_n} \inf_{g\neq g'} |T(g(S^*_{n,\mathcal{C}})) - T(g'(S^*_{n,\mathcal{C}}))| > 0,$$

where $g, g' \in \mathcal{G}_{\mathcal{C}}$ and $\mathcal{G}_{\mathcal{C}}$ is the set of transformations associated with clustering $\mathcal{C}$.

Theorem 2. *Let* $\mathsf{X}_n, Z_{i,n}, \mathscr{C}_n$ *satisfy* **C1**, **C2**, *and* **C3**. *Suppose* **R1** *holds. Suppose* $Z_{i,n}$ *are observable functions of data* $\mathscr{D}_n \sim P_n$. *Suppose further that* ϕ *is the CRS test result and is based on* $S_{n,\mathcal{C}} = S_{n,\mathcal{C}}(\mathscr{D}_n)$. *Then under Algorithm 1,*

$$\sup_{\mathcal{C}\in\mathscr{C}_n} |\mathrm{E}_{P_n}[\phi(S_{n,\mathcal{C}})] - \alpha| \to 0.$$

5. Simulation

In this section, we study the finite sample performance of inference with learned clusters in a series of simulation experiments. We consider a $N \times T$ panel from the following process:

$$y_{it} = \alpha_0 + \theta_0 x_{it} + w'_{it}\gamma_0 + u_{it},$$

where the parameter of interest is θ_0, and w_{it} is a vector of control variables. For each i, we observe a coordinate vector L_i, which is taken from our empirical example. Details of the data generating process will be provided below.

5.1. Inferential procedures

As benchmarks, we consider several alternative inferential methods: cluster covariance estimator only by location i (LOCA), spatial HAC estimator from [33] with optimal bandwidth selection described in [22] (SK), and U-statistic type subsampling test as in [32] and [26] (SL).

Out proposed method is described by Algorithm 1 and consist of three steps: (i) learning a sequence of clustering structures $\{\mathcal{C}^{(2)}, \ldots, \mathcal{C}^{(\bar{G})}\}$, (ii) choosing the optimal clustering $\mathcal{C}^*$ with the optimal cluster number G^* based on some criterion of size-power tradeoff, (iii) given $\mathcal{C}^*$, performing cluster-based hypothesis testing including IM, and CRS. Details on implementation are given in Appendix A. We also consider the commonly used method CCE with the learned clustering, which often turns out to be over-rejecting. Besides, we consider an procedure (OPTI) that gives the highest simulated power among CCE, IM, and CRS.

5.2. Settings and results

We consider both OLS and IV estimation. For each setting, we implement 1000 replications. The nominal rejection rate is $\alpha = 0.05$.

5.2.1. OLS

The results for OLS are in Table 1. We consider the following configurations:

BENCHMARK In this benchmark case, we let $(N, T, p) = (820, 2, 10)$. The empirical example only has 205 locations. We manually generate three more copies of the original map in the neighboring area in order to keep the feature of naturally-formed locations. For some generic variable z (u or any entry of $(x, w')'$), the correlation between two observations is given by

$$\operatorname{cov}(z_{it}, z_{js}) = \exp\left(-\left(\frac{\|L_i - L_j\|_2}{\kappa} + \frac{|t-s|}{\rho}\right)\right), \tag{5.1}$$

where $\kappa = 3$ is the strength of spatial correlation and $\rho = 1$ is strength of serial correlation. Let the corresponding correlation matrix of vectorization of z be Σ. The correlation between any pair of regressors in $(x_{it}, w_{it}')'$, is set to 0.5. The marginal distribution of any regressor or the error term u_{it} is $N(0,1)$. The error term u is exogenous. The data is generated by Equation (5), with the parameter of interest $\theta_0 = 0$ and $(\alpha_0, \gamma_0')' = (1, \ldots, 1)'$. Other configurations are deviations from this benchmark case.

NON-STA This non-stationary case differs from the benchmark case only for the marginal distribution of the regressors. Instead of being one, the variance of each regressor is now $(2\sqrt{\|L_i - \bar{L}\|_2} - 4)^2$ with $\bar{L}$ being the "center" of all locations.

SAR In this setting, we assume the cross-sectional interdependence comes from a spatial auto-regression process (SAR). Define W to be an $N \times N$ weighting matrix with entry $W_{i,j} = 1\{\|L_i - L_j\|_2 < 0.3\}$. Let $U_t = (u_{1t}, \dots, u_{Nt})'$ for $t = 1, 2$, where

$$\begin{cases} U_1 = (I - \kappa W)^{-1} V_1 \\ U_2 = \rho U_1 + \sqrt{1-\rho^2}(I - \kappa W)^{-1} V_2. \end{cases}$$

and

$$\begin{pmatrix} V_1 \\ V_2 \end{pmatrix} \sim N\left(0, I_{NT}\right).$$

We set the spatial correlation $\kappa = 0.14$ and the serial correlation $\rho = \exp(-1)$. The spatial correlation κ is chosen to match the benchmark case in terms of the performance of the naïve method (LOCA). Each regressor follows the same SAR process and the correlation between any pair of V_1's (or V_2's) of different regressors is set to 0.5.

DISCRETE This setting differs from the benchmark case only for the marginal distribution of x_{it}. We transform the standard normal variable into a discrete one using a monotonic function such that the marginal distribution coincides with that in the empirical example.

SMALL-N We only use the original 205 locations in this setting.

LOW-SPA We let $\kappa = 0.5$ in the setting.

LOW-D We generate only $p = 1$ control variable in this setting. The coefficient is then $(\alpha_0, \beta_0, \gamma_0) = (1, 0, 1)$.

From Table 1, we find IM and CRS perform well in general and are correctly-sized in the low spatial correlation case. In the cases with high spatial correlation, IM and CRS are typically slightly over-sized while being powerful. CCE in general over-rejects unless the spatial correlation is low. For the three benchmark cases, LOCA and SK are always over-sized, and SL essentially has no power except for the case with only one control variable.

5.2.2. IV

The results for the IV cases are in Table 2. We consider the following configurations:

BENCHMARK In this benchmark case of IV, the data is generated using the following process:

$$\begin{cases} y_{it} = \alpha_0^y + \theta_0 x_{it} + w_{it}' \gamma_0^y + u_{it}^y, \\ x_{it} = \alpha_0^x + \pi_0 z_{it} + w_{it}' \gamma_0^x + u_{it}^x, \end{cases}$$

where x_{it} is the endogenous variable and z_{it} is the instrument. The marginal distribution of z_{it} and each entry of the control vector w_{it} is $N(0, 1)$. For the error terms, we first generate

$$\begin{pmatrix} V^y \\ V^x \end{pmatrix} \sim N\begin{pmatrix} I_{NT} & \rho^v I_{NT} \\ \rho^v I_{NT} & I_{NT} \end{pmatrix}$$

Table 1
OLS

					Power				Size-adjusted Power			
	Method	Mean	RMSE	Size	-4se	-2se	+2se	+4se	-4se	-2se	+2se	+4se
BENCHMARK	LOCA	0.006	0.20	0.42	1.00	0.85	0.86	1.00	0.91	0.50	0.53	0.93
	SK	0.006	0.20	0.36	1.00	0.83	0.84	1.00	0.94	0.49	0.52	0.94
	SL	0.006	0.20	0.03	0.04	0.03	0.03	0.04	0.06	0.04	0.05	0.05
	CCE	0.006	0.20	0.15	0.96	0.59	0.64	0.97	0.88	0.42	0.41	0.89
	IM	0.000	0.12	0.10	1.00	0.84	0.84	1.00	1.00	0.77	0.76	1.00
	CRS	0.000	0.12	0.08	1.00	0.83	0.82	1.00	1.00	0.77	0.75	1.00
	OPTI	0.004	0.18	0.14	0.97	0.70	0.72	0.98	0.91	0.48	0.50	0.92
NON-STA	LOCA	0.002	0.17	0.39	0.99	0.82	0.83	0.99	0.91	0.49	0.51	0.92
	SK	0.002	0.17	0.38	1.00	0.82	0.83	0.99	0.93	0.50	0.53	0.93
	SL	0.002	0.17	0.03	0.04	0.03	0.03	0.03	0.05	0.04	0.04	0.04
	CCE	0.002	0.17	0.13	0.94	0.58	0.59	0.94	0.83	0.36	0.36	0.83
	IM	-0.002	0.11	0.07	1.00	0.81	0.79	1.00	1.00	0.78	0.76	1.00
	CRS	-0.001	0.11	0.07	1.00	0.79	0.78	1.00	1.00	0.70	0.71	0.99
	OPTI	0.000	0.16	0.13	0.95	0.62	0.62	0.95	0.85	0.38	0.39	0.86
SAR	LOCA	0.000	0.11	0.46	1.00	0.91	0.91	1.00	0.97	0.52	0.54	0.97
	SK	0.000	0.11	0.29	1.00	0.82	0.84	0.99	0.95	0.51	0.54	0.95
	SL	0.000	0.11	0.04	0.06	0.04	0.04	0.06	0.08	0.06	0.06	0.08
	CCE	0.000	0.11	0.13	0.95	0.61	0.64	0.95	0.90	0.46	0.45	0.88
	IM	-0.001	0.06	0.05	1.00	0.94	0.93	1.00	1.00	0.94	0.92	1.00
	CRS	-0.001	0.06	0.05	1.00	0.92	0.91	1.00	1.00	0.93	0.92	1.00
	OPTI	-0.001	0.10	0.12	0.98	0.71	0.72	0.98	0.95	0.54	0.54	0.94
DISCRETE	LOCA	-0.001	0.11	0.41	1.00	0.89	0.87	1.00	0.95	0.55	0.55	0.95
	SK	-0.001	0.11	0.34	0.99	0.86	0.86	0.99	0.96	0.57	0.56	0.95
	SL	-0.001	0.11	0.05	0.03	0.04	0.04	0.03	0.03	0.04	0.04	0.03
	CCE	-0.001	0.11	0.15	0.97	0.67	0.66	0.96	0.87	0.44	0.43	0.88
	IM	0.000	0.07	0.07	0.99	0.87	0.87	0.99	0.99	0.84	0.84	0.99
	CRS	-0.001	0.06	0.08	1.00	0.88	0.87	1.00	1.00	0.84	0.83	1.00
	OPTI	-0.002	0.10	0.14	0.98	0.76	0.75	0.97	0.92	0.53	0.54	0.91
SMALL-N	LOCA	-0.003	0.29	0.51	1.00	0.91	0.90	1.00	0.95	0.51	0.52	0.94
	SK	-0.003	0.29	0.38	1.00	0.87	0.86	0.99	0.94	0.45	0.47	0.92
	SL	-0.003	0.29	0.04	0.05	0.04	0.04	0.05	0.06	0.06	0.06	0.08
	CCE	-0.003	0.29	0.19	0.96	0.64	0.66	0.95	0.80	0.36	0.36	0.81
	IM	-0.001	0.20	0.08	0.95	0.71	0.72	0.94	0.92	0.58	0.59	0.92
	CRS	0.000	0.20	0.07	0.96	0.67	0.67	0.96	0.97	0.64	0.64	0.96
	OPTI	-0.004	0.29	0.18	0.96	0.67	0.67	0.95	0.79	0.37	0.37	0.79
LOW-SPA	LOCA	0.002	0.07	0.21	1.00	0.78	0.79	1.00	0.98	0.52	0.55	0.98
	SK	0.002	0.07	0.13	0.99	0.68	0.70	0.99	0.98	0.52	0.54	0.98
	SL	0.002	0.07	0.04	0.05	0.04	0.05	0.04	0.05	0.05	0.05	0.06
	CCE	0.002	0.07	0.05	0.94	0.44	0.47	0.93	0.95	0.45	0.48	0.94
	IM	0.001	0.06	0.04	0.98	0.54	0.56	0.98	0.99	0.59	0.60	0.99
	CRS	0.000	0.06	0.04	0.97	0.52	0.53	0.97	0.98	0.55	0.55	0.98
	OPTI	0.002	0.06	0.05	0.96	0.48	0.52	0.96	0.96	0.47	0.50	0.95
LOW-D	LOCA	-0.004	0.21	0.34	0.99	0.84	0.83	0.99	0.94	0.51	0.52	0.93
	SK	-0.004	0.21	0.33	1.00	0.85	0.84	0.99	0.96	0.55	0.56	0.94
	SL	-0.004	0.21	0.04	0.34	0.06	0.06	0.35	0.41	0.09	0.09	0.40
	CCE	-0.004	0.21	0.09	0.96	0.56	0.56	0.94	0.92	0.46	0.47	0.91
	IM	-0.005	0.17	0.07	0.99	0.74	0.73	1.00	0.99	0.67	0.67	0.99
	CRS	-0.005	0.17	0.06	0.99	0.71	0.70	0.99	0.99	0.65	0.64	0.98
	OPTI	-0.001	0.21	0.09	0.97	0.57	0.59	0.94	0.93	0.50	0.51	0.92

Notes: All other settings are deviations from *BENCHMARK*. The true parameter is 0. Alternatives for power calculation depend on the standard deviation of the full-sample least square estimator and are thus different across settings. Size-adjusted power is calculated by adjusted p-value threshold such that the hypothesis is rejected in 5% of the replications.

and then form

$$\begin{cases} U^y = \Sigma^{-1/2} V^y, \\ U^x = \Sigma^{-1/2} V^x, \end{cases}$$

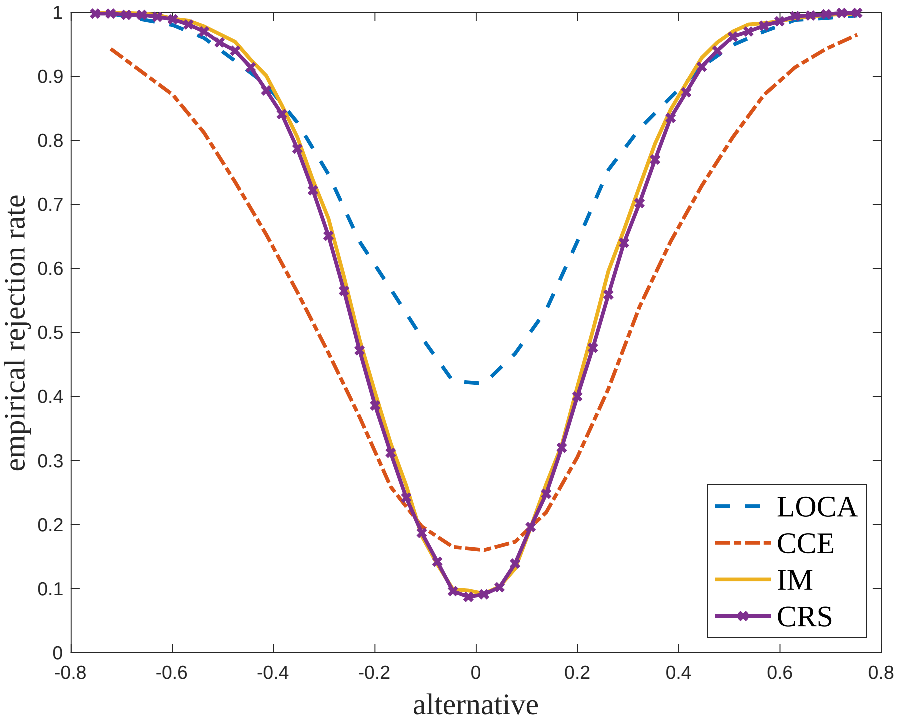

Fig 1. *Power curves in BENCHMARK of OLS with K-medoids clustering.*

where U^y and U^x are the vectorization of u_{it}^y and u_{it}^x, respectively. The level of endogeneity ρ^v is set to 0.6, and Σ is the correlation matrix specified in Equation (5.1). The level of first stage strength π_0 is set to 0.5. Other settings are the same as in the benchmark case of OLS.

HIGH-ENDO This case is different from the benchmark case only for $\rho^v = 0.9$.

HIGH-RELEV This case is different from the benchmark case only for $\pi_0 = 1$.

EMPIRICAL This case is similar to the setting of our empirical application. In this case, the endogenous variable x_{it} is transformed into a discrete variable, whose marginal distribution matches that of the empirical example. Also, the number of location is $N = 205$.

In the results table, we provide the median and the median absolute deviation (MAD) of the simulated estimates because the IV estimator does not have finite sample moments under just identification. In general, sizes are well-controlled by our cluster-based methods. A possible explanation is that tests are conservative under just-identification in the IV settings. Compared with CCE and CRS, IM is less powerful. CRS is slightly more powerful than CCE at large alternatives, which is shown in Figure 2. As in the OLS cases, LOCA and SK are in general over-sized except for SK in *EMPIRICAL*, and SL is correctly-sized but essentially has no power.

Table 2
IV

					Power				Size-adjusted Power			
	Method	Median	MAD	Size	-4mad	-2mad	+2mad	+4mad	-4mad	-2mad	+2mad	+4mad
BENCHMARK	LOCA	0.024	0.70	0.20	0.90	0.86	0.86	0.90	0.88	0.81	0.82	0.88
	SK	0.024	0.70	0.09	0.90	0.84	0.85	0.90	0.89	0.82	0.83	0.88
	SL	0.024	0.70	0.03	0.25	0.06	0.07	0.25	0.29	0.07	0.10	0.29
	CCE	0.024	0.70	0.02	0.86	0.79	0.79	0.87	0.88	0.83	0.83	0.89
	IM	0.031	1.87	0.02	0.90	0.78	0.78	0.88	0.91	0.83	0.83	0.90
	CRS	0.042	1.84	0.06	0.93	0.84	0.84	0.92	0.93	0.84	0.84	0.92
	OPTI	0.036	1.53	0.04	0.84	0.75	0.75	0.84	0.85	0.77	0.78	0.85
HIGH-ENDO	LOCA	0.029	0.83	0.20	0.89	0.83	0.85	0.89	0.86	0.79	0.80	0.87
	SK	0.029	0.83	0.10	0.88	0.83	0.83	0.89	0.87	0.81	0.81	0.88
	SL	0.029	0.83	0.03	0.17	0.04	0.05	0.17	0.20	0.05	0.07	0.21
	CCE	0.029	0.83	0.03	0.85	0.78	0.76	0.85	0.86	0.80	0.80	0.88
	IM	0.032	1.10	0.02	0.88	0.75	0.75	0.88	0.90	0.79	0.78	0.90
	CRS	0.028	1.59	0.06	0.91	0.82	0.80	0.90	0.91	0.82	0.80	0.89
	OPTI	0.037	1.43	0.04	0.81	0.73	0.70	0.80	0.82	0.75	0.72	0.83
HIGH-RELEV	LOCA	0.006	0.21	0.31	0.92	0.86	0.84	0.90	0.88	0.78	0.76	0.87
	SK	0.006	0.21	0.15	0.91	0.83	0.81	0.89	0.89	0.77	0.76	0.87
	SL	0.006	0.21	0.03	0.05	0.02	0.04	0.04	0.06	0.03	0.05	0.06
	CCE	0.006	0.21	0.04	0.87	0.72	0.71	0.85	0.88	0.75	0.73	0.86
	IM	0.013	0.74	0.02	0.74	0.52	0.53	0.75	0.78	0.57	0.59	0.78
	CRS	0.014	0.74	0.06	0.79	0.61	0.61	0.81	0.78	0.59	0.60	0.80
	OPTI	0.006	0.55	0.05	0.78	0.63	0.63	0.79	0.79	0.63	0.63	0.79
EMPIRICAL	LOCA	0.022	0.77	0.09	0.88	0.83	0.83	0.89	0.87	0.82	0.82	0.87
	SK	0.022	0.77	0.04	0.87	0.81	0.82	0.88	0.88	0.82	0.82	0.88
	SL	0.022	0.77	0.03	0.45	0.17	0.18	0.42	0.50	0.21	0.22	0.47
	CCE	0.022	0.77	0.03	0.84	0.77	0.76	0.85	0.87	0.81	0.81	0.86
	IM	0.074	8.01	0.02	0.89	0.81	0.80	0.89	0.92	0.84	0.84	0.91
	CRS	0.077	7.89	0.05	0.88	0.81	0.81	0.88	0.89	0.82	0.82	0.89
	OPTI	0.023	0.93	0.03	0.84	0.77	0.77	0.83	0.86	0.80	0.80	0.85

Notes: All other settings are deviations from *BENCHMARK*. The true parameter is 0. MAD is the median absolute deviation. Alternatives for power calculation depend on the variation of the full-sample IV estimator and are thus different across settings. Size-adjusted power is calculated by adjusted p-value threshold such that the hypothesis is rejected in 5% of the replications.

5.2.3. Clustering

In this subsection we look at the performance of our clustering algorithm. The results are shown in Table 3, where we look at the benchmark cases with L standing for 820 locations and S for 205 locations. For example, OLS-L is exactly the *BENCHMARK* case in the OLS settings. The maximum of numbers of clusters considered is $\bar{G} = 6$ for the small-sample case (OLS-S) and 10 for the two large-sample cases (OLS-L and IV-L).

For the small-sample case, the optimal number of clusters (G^*) is always chosen at 6 by the size-power tradeoff. CCE being over-sized indicates that the procedure often underestimates the spatial correlation and thus the size, so that the size-power tradeoff tends to pick a large number of clusters to achieve a high power, resulting in over-rejection. This is also true for the large-sample case of OLS.

6. Empirical Application: The Logic of Insurgent Electoral Violence

In this section, we apply our inference procedure based on learned clusters to an empirical example which is likely to present spatial dependence. We revisit the effect of insurgent

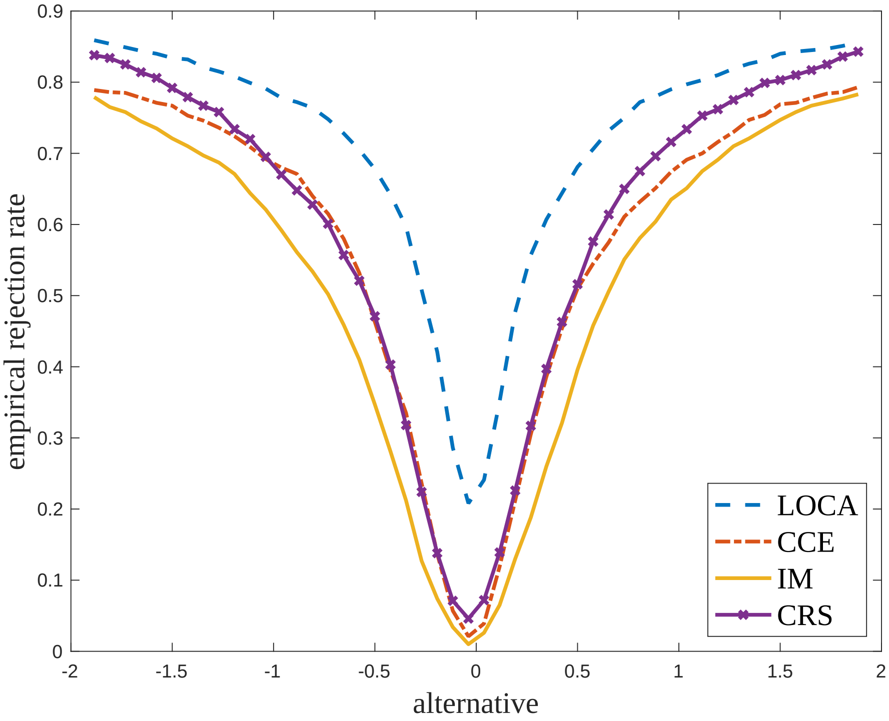

Fig 2. *Power curves in BENCHMARK of IV with K-medoids clustering.*

Table 3
Clustering

				G^*			Size (fixed G)			Estimated Size			Estimated Power			
Setting	Method	Size	Power	q25	q50	q75	G=2	G=6	G=10	q25	q50	q75	q25	q50	q75	Optimal
OLS-L	CCE	0.15	0.22	9	10	10	0.07	0.12	0.15	0.05	0.05	0.05	0.13	0.18	0.25	0.61
	IM	0.10	0.26	9	10	10	0.05	0.07	0.09	0.04	0.04	0.05	0.12	0.16	0.21	0.08
	CRS	0.08	0.24	8	9	10	0.00	0.05	0.09	0.05	0.05	0.05	0.13	0.18	0.24	0.31
OLS-S	CCE	0.19	0.24	6	6	6	0.06	0.20	-	0.05	0.05	0.05	0.12	0.16	0.25	0.87
	IM	0.08	0.15	6	6	6	0.07	0.09	-	0.02	0.03	0.04	0.09	0.12	0.18	0.09
	CRS	0.07	0.15	6	6	6	0.00	0.07	-	0.02	0.03	0.04	0.07	0.10	0.14	0.04
IV-L	CCE	0.02	0.14	10	10	10	0.03	0.02	0.02	0.05	0.05	0.05	0.10	0.16	0.30	0.54
	IM	0.02	0.27	8	10	10	0.03	0.02	0.02	0.04	0.04	0.05	0.10	0.14	0.25	0.08
	CRS	0.06	0.32	8	9	10	0.00	0.04	0.06	0.05	0.05	0.05	0.11	0.15	0.27	0.38
IV-S	CCE	0.03	0.12	6	6	6	0.03	0.04	-	0.05	0.05	0.05	0.12	0.19	0.28	0.87
	IM	0.02	0.30	6	6	6	0.03	0.02	-	0.02	0.03	0.04	0.10	0.15	0.21	0.08
	CRS	0.05	0.34	6	6	6	0.00	0.05	-	0.02	0.03	0.04	0.08	0.13	0.21	0.05

Notes: L is the large-sample case with $N = 820$ and S is with $N = 205$. For all columns, q25,q50, and q75 are 25-, 50-, and 75 quantiles of quantities across replications. G^* is the number of clusters chosen by the criterion on size-power tradeoff. The last column "Optimal" is the frequency of a particular method achieving the highest simulated power among all three methods in the setting.

attacks on voter turnouts studied in Condra et al. (2018) [7]. In Condra et al. (2018) [7], the authors study the impact of direct morning attacks on voter turnouts in the first and second rounds of the 2014 election in Afghanistan. To do so, the authors estimate the following linear model both by OLS and IV:

$$Y_{d,e} = \alpha + \beta_1 Attacks_{d,e} + \beta_2 X_{d,e} + \epsilon_{d,e}, \tag{6.1}$$

where $Y_{d,e}$ is the turnout in district d and election round e, $Attacks_{d,e}$ is the number of morning attacks in district d and election round e. The covariates in $X_{d,e}$ include election round fixed effects, voting hour wind conditions, population, a measure of precipitation and ambient temperature and the average of predawn and morning wind conditions during the preelection period. In order to overcome possible endogeneity, the authors use early morning wind conditions as an instrument for $Attacks_{d,e}$.

The number of districts D is 205 and the number of observations used in the regression is 410 (considering the two rounds). For inference purposes the authors use cluster standard errors (CCE) at the district level which are robust to within-district correlation (LOCA).

In this section we use our learned cluster-based methodology to perform inference on β_1 allowing for between-district spatial dependence. Specifically, after selecting the groups based on some distance measure, we perform inference considering all three cluster-based testing procedures CCE, IM, and CRS. Those procedures are implemented in the same way as described in Section 5 and described in detail in Appendix A.

6.1. Selection of groups/clusters based on geographic distance

The first step is to produce groups based on some observable distance measure. We think that both, the instrument and shocks that affect voting, might be spatially correlated across districts that are geographically close. We use data on latitude and longitude coordinates to calculate a measure of geographical distance d_{ij} between district i and district j.

Then, we group observations into clusters according to our distance measure. To group the data we use k-medoids (KM). Since we are working with 205 cross-sectional units, we allow for a maximum of six groups and a minimum of 25 cross-sectional units in each group. The procedure generates a sequence of group structures $\{\mathcal{C}^{(2)}, \dots, \mathcal{C}^{(6)}\}$.

Tables 4 reports the IV and first-stage estimates of β_1 for the six subgroups generated using KM. Column labeled full sample reproduces the results in Condra et al. (2018) [7] using all the observations in the sample. Condra et al. (2018) [7] found a significant negative effect (-0.145) of violence over voter turnout. Columns 3 to 8 in Table 4 display the IV and first stage estimates for each of the six subgroups generated by KM. The row labeled IV displays the second stage estimate of β_1 using the full sample and the six sub-samples generated by the KM algorithms. The row labeled First Stage displays the first stage estimates using early morning wind conditions as an instrument. Rows labeled s.e. and t-stat display cluster standard errors by location (we have two rounds of voting for each location) and t-statistic for the full sample and the six subsamples.

From Table 4, we can see that there is a clear deterioration in both the strength of the instrument in the first stage and the IV estimate in the second stage, when we look at the subgroups. For example, in the full sample, the t-statistic associated with the instrument in

TABLE 4
Impact of Early Morning Attacks on Voter Turnout during the 2014 Election

		Within 6 subgroups using k-medoids					
	Full sample	(1)	(2)	(3)	(4)	(5)	(6)
IV	-0.145	-0.180	0.054	0.021	0.085	-1.498	0.063
s.e.	0.061	2.384	0.134	0.073	0.071	2.590	0.617
t-stat	2.385	0.075	0.404	0.288	1.196	0.578	0.102
First Stage	0.281	-0.020	0.272	0.420	0.432	0.132	0.098
s.e.	0.086	0.214	0.232	0.490	0.180	0.231	0.257
t-stat	3.252	0.092	1.173	0.858	2.402	0.573	0.379

Note: The table reports IV and first stage estimates of β_1 and their associated standard errors and t-statistic for the full sample and for six groups generated using k-medoids based on geographic distance. Column labeled "Full sample" reproduces the results in Condra et al. (2018) [7]. Columns 3–8 display the IV and first stage estimates for each of the 6 sub-groups generated by k-medoids.

the first stage is 3.252 whereas in the subsamples the t-statistic is less than 1 in 4 of the six subgroups selected by the KM . In the same line, the IV estimate in the second stage is not significant in any of the subgroups selected using KM.

The resulting clustering as in the map is shown in Figure 3.

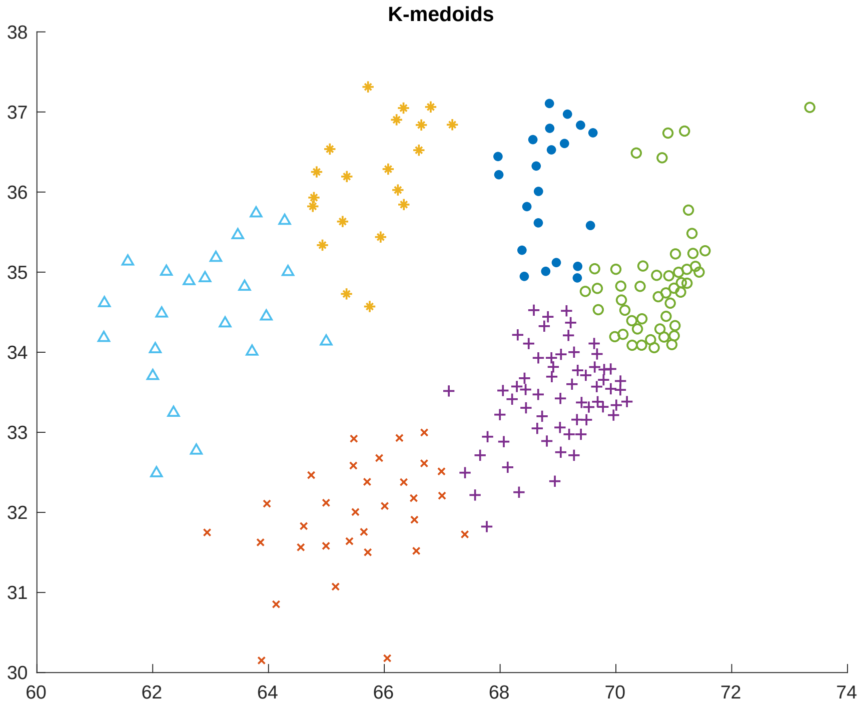

FIG 3. *KM*

Table 5
Inference Based on Selected Clusters

	t-stat	p-value	C.I.	G^*
LOCA	2.385	0.018	[-0.265, -0.025]	205
CCE	1.291	0.253	[-0.433, 0.144]	6
IM	0.954	0.384	[-0.896, 0.411]	6
CRS	-	0.564	[-1.496, 0.084]	6

Note: The table reports the results of the inference procedure based on selected clusters. Row labels indicate which procedure is used. Column labeled t-stat reports t-statistic of the IV estimate of β_1 for each of the procedures. Column p-value reports the rejection rate evaluated at the null: $\beta_1 = 0$. Column C.I reports confidence intervals of the IV estimate of β_1 for each of the procedures. Column G^* indicates the number of clusters used in each procedure.

6.2. Results

Table 5 shows the results of our inference procedure based on selected clusters. Row labels indicate which procedure is used. Row labeled LOCA uses cluster standard errors at the district level as in Condra et al. (2018) [7]. Row labeled CCE uses cluster standard errors for groups selected by the k-medoids algorithm based on geographic distance. Row labeled IM reports the IM inference procedure using groups selected by the k-medoids alghorithm based on geographic distance. Row labeled CRS reports the CRS inference procedure using groups selected by the k-medoids alghorithm based on geographic distance. Column labeled t-stat reports t-statistic of the IV estimate of β_1 for each of the procedures. Column p-value reports the rejection rate evaluated at the null: $\beta_1 = 0$. Column C.I. reports confidence intervals of the IV estimate of β_1 for each of the procedures. Column G^* denotes number of groups used in each procedure. The optimal number of cluster G^* based on the size-power trade-off is 6 in CCE, IM, and CRS. Looking at the t-stat and p-value columns, we see that when considering spatial dependence beyond within-district dependence, the effect of violence over voter turnout is no longer significant. Each of the confidence intervals for the IV estimator of β_1 constructed using our inference procedure is considerably wider than the confidence interval constructed with cluster standard errors at the district level. More importantly, when we consider between-district spatial dependence, the confidence intervals for the IV estimator of β_1 in each of our inferential methods include positive values for the estimated marginal effect of violence on voter turnout.

7. Conclusion

This paper proposes a cluster-based inferential procedure. Observations are grouped into clusters which are learned using a unsupervised learning algorithm given a dissimilarity measure. We consider a set of cluster-based inference procedure on the learned clusters. We give conditions under which our procedure asymptotically attains correct size. We illustrate the finite sample validity and apply our procedure to an empirical example.

Appendix A: Implementation

In this section we present the details of the proposed procedure we use in the simulation and applications sections. In those settings a panel data of $N \times T$ is given, along with the geographical coordinates of N locations. We consider a linear model

$$y_{it} = \theta_0 x_{it} + w_{it}'\gamma + u_{it},$$

along with a scalar instrument variable z_{it}, where $\theta_0 \in \mathbb{R}$ is the parameter of interest. For the rest part of the section, we only consider this IV model and take the OLS case as an special case where $z_{it} = x_{it}$.

Generally, our method is described by Algorithm 1 and consists of three steps: (i) learning a sequence of clustering structures $\mathscr{C}$, (ii) choosing the optimal clustering $\mathcal{C}^*$ with the optimal cluster number G^* based on some criterion of size-power tradeoff, (iii) performing cluster-based hypothesis testing using $\mathcal{C}^*$. Details are given in the following subsections.

A.1. Clustering

We only cluster on the location and always assign different time periods of a single location to the same cluster. We choose $\bar{G} = \lceil N^{1/3} \rceil$. For each $G \in \{2, \ldots, \bar{G}\}$, we run k-medoids with G clusters, using the squared Euclidean distance of the locations as the input. This step generates the sequence of clustering structures $\mathscr{C} = \{\mathcal{C}^{(2)}, \ldots, \mathcal{C}^{(\bar{G})}\}$.

A.2. Size-power tradeoff

Let $\tilde{y}_{it}$, $\tilde{x}_{it}$, and $\tilde{z}_{it}$ be the partialed-out variables from w_{it}, i.e. $\tilde{y}_{it}$ is the residual of regressing y_{it} on w_{it}, and same for $\tilde{x}_{it}$ and $\tilde{z}_{it}$. Then, we have $\widehat{\theta} - \theta_0 = ((NT)^{-1} \sum_{i,t} \tilde{z}_{it}\tilde{x}_{it})^{-1}(NT)^{-1} \sum_{i,t} \tilde{z}_{it}\widehat{u}_{it}$. Define the score by

$$s_{it} = \frac{\tilde{z}_{it}\widehat{u}_{it}}{\frac{1}{NT}\sum_{j,r} \tilde{z}_{jr}\tilde{x}_{jr}}.$$

Next, we estimate a covariance model for the scores. Namely, We consider the class of covariance functions parameterized by $b = (b_1, b_2, b_3, b_4)'$ and given by

$$\left\{ \mathrm{cov}(s_{it}, s_{jr}; b) = b_1 \exp\left(- \left(\frac{\|L_i - L_j\|_2}{b_2} + \frac{|t-r|}{b_3} \right) \right) \cos\left(b_4 \left(\frac{\|L_i - L_j\|_2}{b_2} + \frac{|t-r|}{b_3} \right) \right) \right\}_{b \in \mathbb{R}_+^4},$$

and estimate b by Gaussian QMLE. Let the resulting covariance matrix estimator be $\widehat{\Sigma} \in \mathbb{R}^{(NT)\times(NT)}$.

Fix a certain cluster-based testing procedure such as IM or CRS, and the significance level α. Given a clustering structure $\mathcal{C}$, we can now simulate the Type-I error $t_1(\mathcal{C})$ and Type-II error $t_2(\mathcal{C}, p; \theta)$ against a certain alternative $\theta_0 = \theta$, where $p \leq \alpha$ is the p-value threshold, i.e. we allow the threshold to be smaller than α in order to make the test more conservative. To simulate data, note that the vector of group-level means can be written as

$$A_{\mathcal{C}} s = \left(\frac{1}{|\mathsf{C}_g|} \sum_{(i,t)\in \mathsf{C}_g} s_{it} \right)_{\mathsf{C}_g \in \mathcal{C}},$$

where $s = (s_{it})_{i=1,\ldots,N,t=1,\ldots,T}$ is the vector of all scores. Thus, the distribution of the vector of group-level means can be approximated by $N(\mathbf{0}, A_{\mathcal{C}}\widehat{\Sigma}A_{\mathcal{C}})$. We generate 10000 observations from this distribution and calculate the rejection rates to form the estimates for Type-I and Type-II errors.

Finally, we solve

$$\begin{aligned} (\mathcal{C}^*, p^*) = \quad & \underset{\mathcal{C}\in\mathscr{C}, p\in(0,\alpha]}{\arg\min} & \frac{1}{J}\sum_{j=1}^{J} t_2(\mathcal{C}, p; \theta_j) \\ & \text{s.t.} & t_1(\mathcal{C}, p) \leq \alpha \end{aligned}$$

That is, we select the clustering structure with the highest (simulated) power among those whose (simulated) sizes are controlled. The set of interesting alternatives $\{\theta_j\}_{j=1}^{J}$ is chosen to be $\{0.25se, 0.75se, 1.25se, 1.75se, 2.25se, 2.75se\}$, where se is the White standard error of θ_0 estimator.

A.3. Group-based hypothesis testing

We consider three testing procedures: the t-statistic based test as in [14] (IM), and the randomization test as in [6] (CRS). The procedures are given here for reference.

A.3.1. IM

The IM procedure is given as follows. For some $x \in \mathbb{R}^G$, define $\bar{x} = G^{-1}\sum_{i=1}^{G} x_i$ and $\mathrm{se}(x) = \sqrt{(G-1)^{-1}\sum_{i=1}^{G}(x_i - \bar{x})^2}$. The t-statistic is defined by $t(x) = \sqrt{G}\bar{x}/\mathrm{se}(x)$. Let $\mathrm{cv}_G(\alpha)$ be the critical value of usual two-sided t-test of level α with $G-1$ degree of freedom. Let $S \in \mathbb{R}^G$ be a vector of within-group estimators for a scalar parameter, i.e., the i-th entry of S is the resulting estimator using only the data from the i-th group. At significance level α, the IM test is given by

$$\psi(S) = \mathbb{1}\{|t(S)| > \mathrm{cv}_G(\alpha)\}.$$

A.3.2. CRS

We now describe the specific version of CRS used in this paper. Consider the following setting with observed data

$$\mathscr{D}_n \sim P_n \in \mathbf{P}_n$$

where $\mathbf{P}_n$ is a set of distributions on a sample space $\mathscr{X}_n$ and $i = 1, ...n$ indexes observations. We assume that $\mathscr{D}_n = \{W_{i,n}\}_{i=1}^n$ where $W_{i,n}$ are real vector valued random variables. Consider a hypothesis testing problem:

$$H_0 : P_n \in \mathbf{P}_{n,0} \quad \text{vs} \quad H_1 : P_n \in \mathbf{P}_n \setminus \mathbf{P}_{n,0} \ \text{ at level } \ \alpha \in (0, 1).$$

The tests constructed in CRS require that the distribution of the observed data exhibits approximate symmetry. To make this precise, Let T be a real valued test statistic, such that large values of T provide evidence against H_0. Let $\mathbf{H}$ be a finite group along with an action $\mathscr{D}_n \mapsto h\mathscr{D}_n$ for all $h \in \mathbf{H}$.

Let $\mathcal{C}$ be a partition of $\{1, ..., n\}$ with G clusters $\mathcal{C} = (\mathsf{C}_1, ..., \mathsf{C}_G)$. We assume that T factors through a function S which takes the form $S_n(\mathscr{D}_n) = (S_{n,1}(\mathscr{D}_n), ..., S_{n,G}(\mathscr{D}_n))$. Therefore, by an abuse of notation, we may write $T(\mathscr{D}_n) = T(S_n(\mathscr{D}_n))$. Further, each component g of S_n depends only on $\{W_{i,n}\}_{i\in\mathsf{C}_g}$. That is, $S_{n,g}(\mathscr{D}_n) = S_{n,g}(\{W_{i,n}\})_{i\in\mathsf{C}_g}$. We assume that the action of $\mathbf{H}$ respects S_n in the sense that the action can be equivalently expressed $h\mathscr{D}_n = hS_n(\mathscr{D}_n)$.

Let $M = |\mathbf{H}|$ and let $T^1(\mathscr{D}_n) \leqslant T^2(\mathscr{D}_n) \leqslant \cdots \leqslant T^M(\mathscr{D}_n)$ denote the order statistics of the orbit of $\{T(h\mathscr{D}_n) : h \in \mathbf{H}\}$. Let $k = M(1-\alpha)$, $M^+(X) = \left|\left\{1 \leqslant j \leqslant M : T^j(\mathscr{D}_n) > T^k\mathscr{D}_n\right\}\right|$ and $M^0(X) = \left|\left\{1 \leqslant j \leqslant M : T^j(X) = T^k(\mathscr{D}_n)\right\}\right|$. Let $a(\mathscr{D}_n) = \frac{M\alpha - M^+(\mathscr{D}_n)}{M^0(\mathscr{D}_n)}$. A randomization test is given by:

$$\text{Reject if } \phi(\mathscr{D}_n) = 1, \quad \phi(\mathscr{D}_n) = \begin{cases} 1 & T(\mathscr{D}_n) > T^k(\mathscr{D}_n) \\ a(X) & T(\mathscr{D}_n) = T^k(\mathscr{D}_n) \\ 0 & T(\mathscr{D}_n) < T^k(\mathscr{D}_n) \end{cases}$$

In CRS it was shown that if (i) $S_n(\mathscr{D}_n) \xrightarrow{d} S$ under P_n, (ii) $hS \overset{d}{=} S$ for all h, (iii) for distinct h, h', either $T \circ h = T \circ h'$ or $\mathrm{P}(T(hS) \neq T(h'S)) = 1$, (iv) T is contintuous and the action of h is continuous for each h, then

$$\mathrm{E}_{P_n}\left[\phi(S_n(X))\right] \to \alpha.$$

The result of CRS shows randomization inference under asymptotic approximate symmetry conditions is asymptotic valid. Our procedure is motivated by the fact that in many applications (including regression and IV data), a partition of the data into clusters of observations which are mostly independent leads to valid approximate inference.

A leading example of functions S_n includes the case when $\mathscr{D}_n = (y_i, x_i, w_i)_{i=1}^n$ is regression data satisfying $y_i = \beta_0 x_i + w_i'\gamma_0 + \varepsilon_i$ with scalar regressor x_i, and

$$S_n(\mathscr{D}_n) = \left(\sqrt{n}(\widehat{\beta}_1 - \beta_0), \ldots, \sqrt{n}(\widehat{\beta}_G - \beta_0)\right)$$

as the $G \times 1$ vector of OLS estimates of a parameter of interest using the data in each of the G subsets of the data. Another leading case is for instrumental variables data $\mathscr{D}_n = (y_i, x_i, w_i, z_i)_{i=1}^n$ and again $S_n(\mathscr{D}_n) = \left(\sqrt{n}(\widehat{\beta}_1 - \beta_0), \ldots, \sqrt{n}(\widehat{\beta}_G - \beta_0)\right)$ as the $G \times 1$ vector of IV estimates. In this paper, we will mainly be interested in the case $\mathbf{H} = \{\pm 1\}^G$ with action

$$hS_n(\mathscr{D}_n) = (h_1\sqrt{n}(\widehat{\beta}_1 - \beta_0), ..., h_G\sqrt{n}(\widehat{\beta}_G - \beta_0)).$$

In those examples, the test statistic T can be simple average of the elements in $S_n(\mathcal{D}_n)$ or $hS_n(\mathcal{D}_n)$. Under common modeling and regularity assumption, $\mathbf{H}$ can be used for the commonly encountered test $H_0 : \beta_0 = 0$.

Appendix B: Proofs of Results in Section 3

B.1. Supporting Results for Proposition 1

Here we show the metric embedding propositions and their Proofs.

Theorem 3 (Assoud)**.** *Let (X, d) be an arbitrary (non necessarily finite) metric space such that $K = dim_{2\times}(\mathsf{X}) < \infty$. Let $\varepsilon \in (0, 1)$. Then there exists an L-bi-Lipschitz map $(\mathsf{X}, d^\varepsilon) \to \mathbb{R}^r$ for some L, r which depend only on ε and K.*

Ahlfors regularity of X implies constant doubling. In this context, we have the following proposition.

Proposition 5. *Suppose that X satisfies C, δ-finite-Ahlfors regularity. Then the doubling dimension $dim_{2\times}(\mathsf{X}_n)$ is bounded by $dim_{2\times} \leqslant \delta \log_2(3C^2)$.*

This notion of regularity implies that, coursely, X have the same dimension in all locations. This can be generalized slightly as the following example illustrates.

Proposition 6. *Suppose that X can be decomposed as a finite disjoint union $\mathsf{X} = \mathsf{X}^{(1)} \sqcup ... \sqcup \mathsf{X}^{(m)}$. Suppose that separately, $\mathsf{X}^{(j)}$ satisfy C_j, δ_j-finite-Ahlfors regularity. Then $dim_{2\times}(\mathsf{X}_n)$ is bounded by $dim_{2\times}(\mathsf{X}) \leqslant 2\sum_{j=1}^m \delta_j \log_2(3C_j^2)$.*

Finite bounds on Assoud's theorem are given in [27]. See also [31] and Assoud's original dissertation [2].

Proposition 7. *Let (X, d) be C, δ-regular. Then $(\mathsf{X}, d^{3/4})$ has an L-bi-Lipschitz map into $\mathbb{R}^r$ where $r \leqslant 833\delta \log(3C^2)$ and the Lipschitz constant L depends only on C, δ.*

Proof. The proof follows from Assoud's Theorem as well as the fact that X has constant doubling by the previous proposition. The proof is completed by chasing the constants in [27]. Preliminaries in [27] notation (note that δ_i is a distinct new quantity here): $\tau = \frac{\varepsilon^\theta}{32(\log K)^\theta}$, θ set to 1/3. $\frac{r}{\delta_i} = 4\tau^{\frac{-3}{1-\varepsilon}} \left(\frac{4\varepsilon}{c^*\gamma}\right)^{\frac{-1}{1-\varepsilon}}$.

Bounds on c^*:

Case I

$$c^* = \frac{\frac{\tau}{1-\tau}}{\left(\frac{64 \log K}{\tau^{1+1/(1-\varepsilon)}}\right)^{1-\varepsilon} \left(\frac{\log K}{\varepsilon}\right)^{1+\theta}} \leqslant \frac{\tau}{1-\tau} \leqslant \frac{1}{16}$$

Case II

$$c^* \leqslant 1$$

Need :

$$e\left(\frac{\varepsilon}{\log K}\right)^{\frac{1}{2}c\log K}\left(K^{4+2\log_2(r/\delta_i)}+1\right) \leqslant 1$$

Sufficient is that

$$16(4+2\log_2(r/\delta_i))\frac{1}{\log_2\log K+\log\frac{1}{\varepsilon}} \leqslant c$$

Next bound $\log_2(r/\delta_i)$.

$$\log_2(r/\delta_i) \leqslant 20+4\left(\log_2\left(\frac{1}{\varepsilon}\right)+\log_2(c^*)+\log_2\log(K)\right)$$

Sufficient that

$$16\left[4+2\left(20+4\left(\log_2\left(\frac{1}{\varepsilon}\right)+\log_2(c^*)+\log_2\log(K)\right)\right)\right]\frac{1}{\log_2\log K+\log_2\frac{1}{\varepsilon}} \leqslant c$$

Sufficient that

$$704+128\left(1+\log_2(c^*)\right) \leqslant c$$

Sufficient that $c = 833$ to deal with integer problems. □

B.2. Proof of Proposition 1

In this section we show the proof of mixing CLT over arrays of uniformly Ahlfors domains.

Proof. In the case that X_n embed isometrically into $\mathbb{R}^\nu$, the conditions imply those required for JP Corollary 1, which has the same conclusion as the conclusion of this proposition. (Condition **C2**(v) in the text is stronger than what is required for JP, who only ask that $\bar{\alpha}_{1,\infty}(m) = O(m^{-\nu-\mu})$ rather than $\bar{\alpha}_{1,\infty}(m) = O(m^{-\nu-\frac{4}{3}\mu})$.

In the case that X_n do not embed isometrically, let L be the bi-Lipschitz constant from the maps $(\mathsf{X}_n, d_n^{1-1/4}) \to \tilde{\mathsf{X}}_n \subseteq \mathbb{R}^{\nu_1}$ where ν_1 can be taken $\nu_1 \leqslant \lceil 833\delta\log(3C^2)\rceil$. The process $\{\{Z_{i,n}\}_{i=1}^n\}_{n=1}^\infty$ indexed on X_n yields a process $\{\{\tilde{Z}_{i,n}\}_{i=1}^n\}_{n=1}^\infty$ indexed on $\tilde{\mathsf{X}}_n$. It is sufficient to check the conditions of Corollary 1 in Jenis and Prucha (2009) for this new process. We apply the same set array of constants $c_{i,n}$ to $\tilde{Z}_{i,n}$. Assumption 1 in JP is satisfied by the fact that distances are at least ρ_0 in X_n for some $\rho_0 > 0$ by Ahflors regulartity. Note that L depends only on C, δ and in particular, does not change with n. Then $\forall i, j$, $\tilde{d}_n(i,j) \geqslant L^{-1}N^{-1/2}\rho_0 > 0$ which also does not depend on n. Condition **C2**(ii) is identical to Equation 3 in JP.

The next conditions in JP are mixing conditions. To verify these, let $\tilde{\alpha}_{k,l,n}(r)$ and $\bar{\tilde{\alpha}}_{k,l,n}(r)$ denote the corresponding mixing coefficients for $\tilde{Z}_{i,n}$ over $\tilde{\mathsf{X}}_n$. Note that $\tilde{d}(U,V) \geqslant r \Rightarrow$

$d(U,V) \geqslant L^{-1}\nu_1^{-1/2}r^{\frac{3}{4}}$. Let $c = L^{-1}\nu_1^{-1/2}$. Then $\bar{\bar{\alpha}}_{k,l}(r) \leqslant \bar{\alpha}_{k,l}(cr^{3/4})$. To verify Equation 4 in JP, it is sufficient to show that

$$\sum_{m=1}^{\infty} \bar{\bar{\alpha}}_{1,1}(m)m^{\nu_1 \times \frac{\mu+2}{\mu}-1} < \infty.$$

Note that $\bar{\bar{\alpha}}_{1,1}(m)$ is nonincreasing and defined for nonnegative real m and $m^{\nu_1 \times \frac{\mu+2}{\mu}-1}$ is a polynomial in m. Thus, the above summation is bounded up to a constant by the corresponding integral

$$\int_{m=0}^{\infty} \bar{\bar{\alpha}}_{1,1}(m)m^{\nu \times \frac{\mu+2}{\mu}-1}dm$$

Substituting $\bar{\bar{\alpha}}_{1,1}(m) \leqslant \bar{\alpha}_{1,1}(cm^{3/4})$ 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

$$\int_{m'=0}^{\infty} \bar{\alpha}_{1,1}(m'){m'}^{\frac{4}{3}(\nu_1 \times \frac{\mu+2}{\mu}-1)}{m'}^{1/3}dm' < \infty.$$

This integral is in turn bounded by a constant times the summation

$$\sum_{m'=1}^{\infty} \bar{\alpha}_{1,1}(m'){m'}^{\frac{4}{3}\nu_1 \times \frac{\mu+2}{\mu}-1} = \sum_{m'=1}^{\infty} \bar{\alpha}_{1,1}(m'){m'}^{\nu \times \frac{\mu+2}{\mu}-1} < \infty$$

which is assumed to be finite under Condition **C2**(iii), thus verifying Equation 4 in JP. Using similar arguments, Condition **C2**(iv) implies Assumption 4(2) in JP. Next, Condition **C2**(v) implies that

$$\bar{\bar{\alpha}}_{1,\infty}(m) \leqslant \bar{\alpha}_{1,\infty}(cm^{3/4}) = O((cm^{3/4})^{\frac{4}{3}(\nu_1-\mu)}) = O(m^{\nu_1-\mu})$$

thus verifying Assumption 4(3) in JP. Finally, Condition **C2**(vi) implies Assumption 5 in JP. This verifies the assumptions of Corollary 1 in JP. The conclusion of Corollary 1 in JP is identical to the conclusion for this result. □

B.3. Proof of Proposition 2

In this section we show that small boundaries give vanishing covariance.

Proof. We prove this in the case that $|\mathcal{C}| = 2$ is which case for $\mathsf{C} \subseteq \mathsf{X}$, we take $\mathsf{D} = \mathsf{X} \setminus \mathsf{C}$. The general case follows by applying the same arguments, and by retracing boundaries when necessary. By arguments in BCH (previously also given in Jenis and Prucha, (2009) and Bolthausen (1982)), $\sigma_{n,\mathsf{C}}^{-1}\sigma_{n,\mathsf{D}}^{-1}n = O(1)$ and

$$\left|\operatorname{cov}\left(\sigma_{n,\mathsf{C}}^{-1}\sum_{i\in\mathsf{C}} Z_j, \sigma_{n,\mathsf{D}}^{-1}\sum_{j\in\mathsf{D}} Z_j\right)\right| \leqslant \sigma_{n,\mathsf{C}}^{-1}\sigma_{n,\mathsf{D}}^{-1}\sum_{(i,j)\in\mathsf{C}\times\mathsf{D}} \bar{\alpha}_{1,1}(\lceil d(i,j)\rceil)^{\mu/(2+\mu)}$$

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

Note that

$$|\{(i,j) \in \mathsf{C} \times \mathsf{D} : k-1 \leqslant d(i,j) < k\}| \leqslant |\{(i,j) \in \mathsf{X} \times \mathsf{X} : k-1 \leqslant d(i,j) < k\}| \leqslant nCk^{\delta}$$

by Condition **C1**. By Condition **C3** it also follows that

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

Then the original covariance is bounded by

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

The first term satisfies $\sigma_{n,\mathsf{C}}^{-1}\sigma_{n,\mathsf{D}}^{-1}\sum_{k=1}^{r} o(n)\bar{\alpha}_{1,1}(k) \to 0$ by $\sum_{k=1}^{r} o(n)\bar{\alpha}_{1,1}(k) \leqslant \sum_{k=1}^{\infty} o(n)\bar{\alpha}_{1,1}(k) < \infty$ and $\sigma_{n,\mathsf{C}}^{-1}\sigma_{n,\mathsf{D}}^{-1}o(n) \to 0$. The second term satisfies $\sigma_{n,\mathsf{C}}^{-1}\sigma_{n,\mathsf{D}}^{-1}\sum_{m=r}^{\infty} nCk^{\delta}\bar{\alpha}_{1,1}(k) \to 0$ by $n\sigma_{n,\mathsf{C}}^{-1}\sigma_{n,\mathsf{D}}^{-1} = O(1)$, $\sum_{m=1}^{\infty} nCk^{\delta}\bar{\alpha}_{1,1}(k) < \infty$ and $r \to \infty \implies \sum_{m=r}^{\infty} Ck^{\delta}\bar{\alpha}_{1,1}(k) \to 0$. This concludes the proof. □

B.4. Proof of Proposition 3

Proof. We prove in the case where $\mathcal{C}_n = \{\mathsf{C}_n, \mathsf{D}_n\}$. General cases with $G > 2$ follow by applying the same arguments. We suppress n for notation simplicity, i.e., $\mathsf{C} = \mathsf{C}_n$.

Step 1. We first show

$$\begin{pmatrix} \Omega_{\mathsf{C}}^{-1/2} & 0 \\ 0 & \Omega_{\mathsf{D}}^{-1/2} \end{pmatrix} s_n \xrightarrow{d} N(0, I_{2K})$$

where $s_n = ((|\mathsf{C}|^{-1/2}\sum_{i\in\mathsf{C}} w_i u_i)', (|\mathsf{D}|^{-1/2}\sum_{i\in\mathsf{D}} w_i u_i)')'$.

Let $v_n = \mathrm{var}(s_n)$. We want to show $v_n^{-1/2}s_n \xrightarrow{d} N(0, I_{2K})$ by the Cramér-Wold device. For $\lambda \in \mathbb{R}^{2K}$, let

$$z_i = \begin{pmatrix} \frac{1}{\sqrt{|\mathsf{C}|}} w_i u_i \mathbb{1}\{i \in \mathsf{C}\} \\ \frac{1}{\sqrt{|\mathsf{D}|}} w_i u_i \mathbb{1}\{i \in \mathsf{D}\} \end{pmatrix}.$$

Then, it is equivalent to show $\sum_{i\in\mathsf{C}\cup\mathsf{D}} \lambda' v_n^{1/2} z_i \xrightarrow{d} N(0,1)$ for any λ with $\|\lambda\| = 1$. To apply Proposition 1, note that **C2**(i), (iii), (iv), and (v) are satisfied by **C2***(i), (iii), (iv), and (v), so it suffices to verify **C2**(ii) & (vi). To see **C2**(ii), note

$$\begin{aligned} \left|\frac{\lambda' v_n^{-1/2} z_i}{c_{i,n}}\right| &\leq \|\lambda\| \cdot \|v_n^{-1/2}\|_F \cdot \left\|\frac{z_i}{c_{i,n}}\right\| \\ &\leq M \left\|\frac{w_i u_i}{c_{i,n}\min\{\sqrt{|\mathsf{C}|}, \sqrt{|\mathsf{D}|}\}}\right\| \\ &\leq \frac{M}{\delta}\|n^{-1/2} w_i u_i\|, \end{aligned} \tag{B.1}$$

where $\sup_n \|v_n^{-1/2}\|_F < M$ for some $M < \infty$ by **C2*(vii)** and $\liminf_{n\to\infty} \min_{\mathsf{C}\in\mathcal{C}_n} |\mathsf{C}|/n > \delta$ for some δ by **C3**. Thus, we have

$$\begin{aligned}
& \lim_{k\to\infty} \sup_n \sup_{i\in\mathsf{X}_n} \mathrm{E}[|\lambda' v_n^{-1/2} z_i/c_{i,n}|^{2+\mu} \mathbb{1}\{|\lambda' v_n^{-1/2} z_i/c_{i,n}| > k\}] \\
\leq \quad & \left(\frac{M}{\delta}\right)^{2+\mu} \lim_{k\to\infty} \sup_n \sup_{i\in\mathsf{X}_n} \mathrm{E}[\|n^{-1/2} w_i u_i/c_{i,n}\|^{2+\mu} \mathbb{1}\{\|n^{-1/2} w_i u_i/c_{i,n}\| > \delta k/M\}] \\
\leq \quad & \left(\frac{M}{\delta}\right)^{2+\mu} \lim_{k\to\infty} \sup_n \sup_{i\in\mathsf{X}_n} \mathrm{E}[\|n^{-1/2} w_i u_i/c_{i,n}\|^{2+\mu} \mathbb{1}\{\|n^{-1/2} w_i u_i/c_{i,n}\| > k\}] \\
= \quad & 0,
\end{aligned}$$

by **C2***(ii). To verify **C2**(vi), note

$$\mathrm{var}\left(\sum_{i=1}^n \lambda' v_n^{-1/2} z_i\right) = \lambda' v_n^{-1/2} \mathrm{var}(s_n)(v_n^{-1/2})'\lambda = 1,$$

so

$$\inf_n n|\mathsf{X}_n|^{-1}\left(\max_{i\in\mathsf{X}_n} c_{i,n}^{-2}\right) \mathrm{var}\left(\sum_{i=1}^n \lambda' v_n^{-1/2} z_i\right) = \inf_n n|\mathsf{X}_n|^{-1}\left(\max_{i\in\mathsf{X}_n} c_{i,n}^{-2}\right) > 0,$$

by **C2***(vi). Therefore, Proposition 1 applies and we have

$$v_n^{-1/2} s_n \xrightarrow{d} N(0,1) \tag{B.2}$$

For some $k \in \{1,\ldots,K\}$, let $w_i^{(k)}$ be the k-th entry of w_i. Following the same argument as in the proof of Proposition 2, we have

$$\begin{aligned}
\left|\mathrm{cov}\left(\frac{1}{\sqrt{|\mathsf{C}|}}\sum_{i\in\mathsf{C}} w_i^{(k)} u_i, \frac{1}{\sqrt{|\mathsf{D}|}}\sum_{i\in\mathsf{D}} w_i^{(l)} u_i\right)\right| \quad & \leq \quad \frac{n}{\sqrt{|\mathsf{C}|\cdot|\mathsf{D}|}} \cdot \frac{1}{n} \sum_{(i,j)\in\mathsf{C}\times\mathsf{D}} \bar{\alpha}_{1,1}(\lceil d(i,j)\rceil^{\mu/(2+\mu)} \\
& \leq \quad \frac{1}{\delta}\cdot o(1) \\
& \to \quad 0.
\end{aligned} \tag{B.3}$$

The second inequality is by **C3** and δ is the uniform lower bound of $|\mathsf{C}|/n$.

Since $(\cdot)^{-1/2}$ is uniformly continuous on $\{A \in \mathbb{R}^{2K} : A = A', \lambda_{\min}(A) > \delta\}$ for some $\delta > 0$, we have

$$\begin{aligned}
& v_n^{-1/2} - \begin{pmatrix} \Omega_{\mathsf{C}}^{-1/2} & 0 \\ 0 & \Omega_{\mathsf{D}}^{-1/2} \end{pmatrix} \\
= \quad & \begin{pmatrix} \Omega_{\mathsf{C}} & \mathrm{cov}\left(\frac{\sum_{i\in\mathsf{C}} w_i u_i}{\sqrt{|\mathsf{C}|}}, \left(\frac{\sum_{i\in\mathsf{D}} w_i u_i}{\sqrt{|\mathsf{D}|}}\right)'\right) \\ \mathrm{cov}\left(\left(\frac{\sum_{i\in\mathsf{C}} w_i u_i}{\sqrt{|\mathsf{C}|}}\right)', \frac{\sum_{i\in\mathsf{D}} w_i u_i}{\sqrt{|\mathsf{D}|}}\right) & \Omega_{\mathsf{D}} \end{pmatrix}^{-1/2} \\
& - \begin{pmatrix} \Omega_{\mathsf{C}} & 0 \\ 0 & \Omega_{\mathsf{D}} \end{pmatrix}^{-1/2} \\
= \quad & o(1),
\end{aligned} \tag{B.4}$$

by **C2***(vii).

Combining (B.2), (B.3), and (B.4), we have

$$\begin{aligned}
\begin{pmatrix} \Omega_{\mathsf{C}}^{-1/2} & 0 \\ 0 & \Omega_{\mathsf{D}}^{-1/2} \end{pmatrix} s_n &= (v_n^{-1/2} + o(1)) s_n \\
&= v_n^{-1/2} s_n + o(1) v_n^{1/2} v_n^{-1/2} s_n \\
&= v_n^{-1/2} s_n + o(1) O(1) O_p(1) \\
&\xrightarrow{d} N(0, I_{2K}).
\end{aligned}$$

The third equality is by **C2***(vii).

Step 2. Let $q_{i,\mathsf{C}} = w_i x_i' - \mathrm{E}[|\mathsf{C}|^{-1} \sum_{i \in \mathsf{C}} w_i x_i']$ and $\widehat{Q}_{n,\mathsf{C}} = |\mathsf{C}|^{-1} \sum_{i \in \mathsf{C}} w_i x_i'$. By **C2***(viii) and the same reasoning as in **Step 1**,

$$\begin{pmatrix} \mathrm{var}\left(\frac{\sum_{i \in \mathsf{C}} q_i}{\sqrt{|\mathsf{C}|}} \right) & 0 \\ 0 & \mathrm{var}\left(\frac{\sum_{i \in \mathsf{D}} q_i}{\sqrt{|\mathsf{D}|}} \right) \end{pmatrix}^{-1/2} \begin{pmatrix} \frac{1}{\sqrt{|\mathsf{C}|}} \sum_{i \in \mathsf{C}} q_i \\ \frac{1}{\sqrt{|\mathsf{D}|}} \sum_{i \in \mathsf{D}} q_i \end{pmatrix} \xrightarrow{d} N(0, I_{2K}).$$

Therefore,

$$\begin{aligned}
\widehat{Q}_n - Q_n &= \begin{pmatrix} \frac{1}{\sqrt{|\mathsf{C}|}} & 0 \\ 0 & \frac{1}{\sqrt{|\mathsf{D}|}} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{|\mathsf{C}|}} \sum_{i \in \mathsf{C}} q_i \\ \frac{1}{\sqrt{|\mathsf{D}|}} \sum_{i \in \mathsf{D}} q_i \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{\sqrt{|\mathsf{C}|}} & 0 \\ 0 & \frac{1}{\sqrt{|\mathsf{D}|}} \end{pmatrix} \begin{pmatrix} \mathrm{var}\left(\frac{\sum_{i \in \mathsf{C}} q_i}{\sqrt{|\mathsf{C}|}} \right) & 0 \\ 0 & \mathrm{var}\left(\frac{\sum_{i \in \mathsf{D}} q_i}{\sqrt{|\mathsf{D}|}} \right) \end{pmatrix}^{1/2} \\
&\quad \cdot \begin{pmatrix} \mathrm{var}\left(\frac{\sum_{i \in \mathsf{C}} q_i}{\sqrt{|\mathsf{C}|}} \right) & 0 \\ 0 & \mathrm{var}\left(\frac{\sum_{i \in \mathsf{D}} q_i}{\sqrt{|\mathsf{D}|}} \right) \end{pmatrix}^{-1/2} \begin{pmatrix} \frac{1}{\sqrt{|\mathsf{C}|}} \sum_{i \in \mathsf{C}} q_i \\ \frac{1}{\sqrt{|\mathsf{D}|}} \sum_{i \in \mathsf{D}} q_i \end{pmatrix} \\
&= o(1) O(1) O_p(1) \\
&= o_p(1).
\end{aligned}$$

By uniform continuity of $(\cdot)^{-1}$ on $\{A \in \mathbb{R}^{2K} : A = A', \lambda_{\min}(A) > \delta\}$ for some $\delta > 0$, we have

$$\begin{pmatrix} \widehat{Q}_{\mathsf{C}}^{-1} & 0 \\ 0 & \widehat{Q}_{\mathsf{D}}^{-1} \end{pmatrix} - \begin{pmatrix} Q_{\mathsf{C}}^{-1} & 0 \\ 0 & Q_{\mathsf{D}}^{-1} \end{pmatrix} = o_p(1).$$

Step 3. By Corollary F4 in [30] ,

$$V_n^{-1/2} \begin{pmatrix} Q_{\mathsf{C}}^{-1} & 0 \\ 0 & Q_{\mathsf{D}}^{-1} \end{pmatrix} s_n = \left[V_n^{-1/2} \begin{pmatrix} Q_{\mathsf{C}}^{-1} & 0 \\ 0 & Q_{\mathsf{D}}^{-1} \end{pmatrix} v_n^{1/2} \right] \cdot \left(v_n^{-1/2} s_n \right) \xrightarrow{d} N(0, I_{2K}).$$

Then, combining previous steps, we have

$$V_n^{-1/2} S_n = V_n^{-1/2} \begin{pmatrix} \widehat{Q}_{\mathsf{C}}^{-1} & 0 \\ 0 & \widehat{Q}_{\mathsf{D}}^{-1} \end{pmatrix} s_n$$

$$
\begin{aligned}
&= V_n^{-1/2}\left(\begin{pmatrix} Q_{\mathsf{C}}^{-1} & 0 \\ 0 & Q_{\mathsf{D}}^{-1} \end{pmatrix} + o_p(1)\right) s_n \\
&= V_n^{-1/2}\begin{pmatrix} Q_{\mathsf{C}}^{-1} & 0 \\ 0 & Q_{\mathsf{D}}^{-1} \end{pmatrix} s_n + V^{-1/2}\cdot o_p(1)\cdot s_n \\
&= V_n^{-1/2}\begin{pmatrix} Q_{\mathsf{C}}^{-1} & 0 \\ 0 & Q_{\mathsf{D}}^{-1} \end{pmatrix} s_n + V^{-1/2}\cdot o_p(1)\cdot v_n^{1/2}\left(v_n^{-1/2} s_n\right) \\
&= V_n^{-1/2}\begin{pmatrix} Q_{\mathsf{C}}^{-1} & 0 \\ 0 & Q_{\mathsf{D}}^{-1} \end{pmatrix} s_n + O(1)o_p(1)O(1)O_p(1) \\
&\xrightarrow{d} N(0, I_{2K}).
\end{aligned}
$$

□

Appendix C: Proof of Propositions 4

In this section we verify the balanced and small boundary conditions for k-medoids.

Proof. Let $\bar{r}_n = \log n$. When possible, from this point on, n is excluded from notation. Consider two points $x, y \in \mathsf{X}$. Let $M = \{z : |d(x,z) - d(z,y)| \leqslant \bar{r}\}$. Let M_0 be an $\bar{r}^2$-net of M. Suppose for sake of contradiction that $|M| \neq o(n)$. Then $|M_0| \neq o(n/\bar{r}^{2\delta})$. Let $A \subseteq [\frac{1}{2}, \frac{9}{10}]$ satisfy $|a - a'| \geqslant \bar{r}^3/d(x,y)$ for each $a, a' \in A$. Take $|A| \geqslant \frac{4}{10}\frac{d(x,y)}{2}/\bar{r}^3$. For $a \in A$, let M_a' consist of interpolants z_a such that $|d(x, z_a) - ad(x,z)| \leqslant K$ and $|d(z_a, z) - (1-a)d(x,z)| \leqslant K$ for each $z \in M_0$. For $a \in A$, let M_a consist of interpolants z_a such that $|d(x,z_a) - ad(x,z)| \leqslant K$ and $|d(z_a,z) - (1-a)d(x,z)| \leqslant K$ for each $z \in M_0$. Then by trigonometry, $\cup_{a \in A} M_a$ is a $3\bar{r}$-separated for n sufficiently large and contains $|A| \times |M_0|$ elements. This can be checked in more detail, by constructing the line segments $\overline{\iota(x)\iota(z)}$ and $\overline{\iota(x)\iota(z')}$ where ι is the coarse isometry to Euclidean space E. Then there are points u, u' which belong to the above constructed line segments with distances $d_{\mathsf{E}}(\iota(z_a), u)$, $d_{\mathsf{E}}(\iota(z_a'), u')$ bounded. u, u' are then shown sufficiently separated to yield the claim. As a result,

$$
\left| \bigcup_{a \in A, z_a \in M_a} \mathbf{B}_{\bar{r}}(z_a) \right| \geqslant |A| \times |M_0| C^{-1} \bar{r}^{\delta} \geqslant \frac{4}{20}\frac{d(x,y)}{\bar{r}^3}|M_0| C^{-1}\bar{r}^{\delta}.
$$

But by $|M_0| \neq o(n/\bar{r}^{2\delta})$, it follows that the above quantity must be larger than n infinitely often provided $d(x,y) \geqslant \bar{r}^{4+\delta}$ for n sufficiently large. This is impossible. Therefore, the small boundaries result is shown once it is shown that k-medoids terminates with medoids $x_1, ..., x_k$ such that $d(x_k, x_l) \geqslant (\log n)^{4+\delta}$ for n sufficiently large. Again for contradiction, suppose there is a sequence $\ell_n = o(1)$ such that for infinitely many n, there are two clusters $\mathsf{C}_1, \mathsf{C}_2$ with medoids x_1, x_2 satisfying $d(x_1, x_2) < \ell_n n^{1/\delta}$. By the pigeonhole principal, there must be a cluster C_3 with n/G members. Then $diam(\mathsf{C}_3)$ must be at least $C^{-1}(n/G)^{1/\delta}$. Let x_3 be the corresponding medoid. Then there must be $x_3' \in \mathsf{C}_3$ such that $d(x_3, x_3') \geqslant \frac{1}{4}C^{-1}(n/G)^{1/\delta}$ and $d(x_3', x_k) \geqslant \frac{1}{4}C^{-1}(n/G)^{1/\delta}$ for any other medoid x_k. Then consider the update in the partitioned medoid algorithm given by $x_2 \leftarrow x_3'$. This update is cost reducing for n sufficiently large. To see this, note that for elements, $x \in \mathbf{B}_{\frac{1}{4}C^{-1}(n/G)^{1/\delta}}(x_3')$ the total cost reduction from being reassigned from a medoid centered around x_2 to a medoid centered around x_3' is at least $|\mathbf{B}_{\frac{1}{4}C^{-1}(n/G)^{1/\delta}}(x_3')|\frac{1}{4}C^{-1}(n/G)^{1/\delta} \geqslant \frac{1}{4}C^{-1}(n/G)^{1/\delta}C^{-1}(\frac{1}{4}C^{-1}(n/G)^{1/\delta})^{\delta}$. The total cost increase from reassigning elements in C_2 to C_1 is at most $\ell_n n^{1/\delta}|\mathsf{C}_2| \leqslant \ell_n n^{1/\delta} n$. 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 $d(x_1, x_2) < \ell_n n^{1/\delta}$ giving the desired contradiction. Finally, note that $d(x_k, x_l) \geqslant \ell_n n^{1/\delta}$ for all medoids x_k, x_l, some ℓ_n bounded uniformly away from 0, and for n sufficiently large implies the balanced clusters condition after applying Ahflors regularity. □

Appendix D: Proof of Theorem 1

D.1. *High level proposition for Theorem 1*

Assumption H1. (Almost sure representation) There exists $\{\tilde{S}_{n,\mathcal{C}}, \tilde{S}^*_{n,\mathcal{C}}\}_{n\geq 1, \mathcal{C}\in\mathscr{C}_n}$ and U defined on a common probability space with $\mathbb{P}$, such that $\forall n \geq 1$ and $\forall \mathcal{C} \in \mathscr{C}_n$, $\tilde{S}_{n,\mathcal{C}} =_d S_{n,\mathcal{C}}$, $\tilde{S}^*_{n,\mathcal{C}} =_d S^*_{n,\mathcal{C}}$, $\sup_{\mathcal{C}\in\mathscr{C}_n} \|\tilde{S}_{n,\mathcal{C}} - \tilde{S}^*_{n,\mathcal{C}}\| \to 0$ with probability one, and U is uniformly distributed on $[0, 1]$ and independent of other random elements. In addition, for each n and $\mathcal{C}$, $\tilde{S}^*_{n,\mathcal{C}} = (\tau_{n,\mathsf{C}_1} r_1, \dots, \tau_{n,\mathsf{C}_G} r_G)$ where $\mathcal{C} = (\mathsf{C}_1, \dots, \mathsf{C}_G)$, $\tilde{R}^*_G = (r_1, \dots, r_G)$, and $\tilde{R}^*_G$ is not a function of n.

Assumption H2. (Normality) For each n and $\mathcal{C}$, $S^*_{n,\mathcal{C}} \sim N(0, V)$, with V is diagonal.

Proposition 8. *Suppose* **H1**, **H2**, *and* **R2** *hold. Then,*

$$\sup_{\mathcal{C}\in\mathscr{C}_n} (\mathrm{E}_{P_n}[\psi(S_{n,\mathcal{C}})] - \alpha)_+ \to 0,$$

where ψ is the result of the IM procedure.

Proof. Let $\{\tilde{S}_{n,\mathcal{C}}, \tilde{S}^*_{n,\mathcal{C}}\}$ be defined as in **H1**. By Theorem 1 of [14], **H2** and **R2**(i) imply

$$\mathrm{E}_{P_n}[\psi(S^*_{n,\mathcal{C}})] \leq \alpha,$$

for each n and $\mathcal{C}$. Note that

$$\begin{aligned}(\mathrm{E}_{P_n}[\psi(S_{n,\mathcal{C}})] - \alpha)_+ &\leq |\mathrm{E}_{P_n}[\psi(S_{n,\mathcal{C}})] - \mathrm{E}_{P_n}[\psi(S^*_{n,\mathcal{C}})]| + (\mathrm{E}_{P_n}[\psi(S^*_{n,\mathcal{C}})] - \alpha)_+ \\ &= |\mathrm{E}_{\mathbb{P}}[\psi(\tilde{S}_{n,\mathcal{C}})] - \mathrm{E}_{\mathbb{P}}[\psi(\tilde{S}^*_{n,\mathcal{C}})]|,\end{aligned}$$

so it suffices to show

$$\sup_{\mathcal{C}\in\mathscr{C}_n} |\mathrm{E}_{\mathbb{P}}[\psi(\tilde{S}_{n,\mathcal{C}})] - \mathrm{E}_{\mathbb{P}}[\psi(\tilde{S}^*_{n,\mathcal{C}})]| \to 0.$$

Let $\mathrm{E}_{n,\mathcal{C}} = \{\psi(\tilde{S}_{n,\mathcal{C}}) = \psi(\tilde{S}^*_{n,\mathcal{C}})\}$, i.e. $\mathrm{E}_{n,\mathcal{C}}$ is the set where $|t(\tilde{S}_{n,\mathcal{C}})|$ and $|t(\tilde{S}^*_{n,\mathcal{C}})|$ are on the same side of $\mathrm{cv}_G(\alpha)$. We follow the same strategy as in the proof of Proposition 3. Then, it suffices to show that there exists Ω with $\mathbb{P}(\Omega) = 1$ such that for each $\omega \in \Omega$, there exists N_ω such that $\forall n \geq N_\omega$ and $\mathcal{C} \in \mathscr{C}_n$, $\omega \in \mathrm{E}_{n,\mathcal{C}}$.

By **H1**, **R2**(ii)&(iii), there exists Ω with $\mathbb{P}(\Omega) = 1$ such that for each $\omega \in \Omega$,

(i) $\sup_{\mathcal{C}\in\mathscr{C}_n} \|\tilde{S}_{n,\mathcal{C}} - \tilde{S}^*_{n,\mathcal{C}}\| \to 0$;
(ii) $\limsup_{n\to\infty} \sup_{\mathcal{C}\in\mathscr{C}_n} |\bar{\tilde{S}}^*_{n,\mathcal{C}}| \le M_\omega$;
(iii) $\liminf_{n\to\infty} \inf_{\mathcal{C}\in\mathscr{C}_n} \mathrm{se}(\tilde{S}^*_{n,\mathcal{C}}) \ge \underline{\eta}_\omega > 0$;
(iv) $\limsup_{n\to\infty} \sup_{\mathcal{C}\in\mathscr{C}_n} \mathrm{se}(\tilde{S}^*_{n,\mathcal{C}}) \le \overline{\eta}_\omega$;
(v) $\liminf_{n\to\infty} \inf_{\mathcal{C}\in\mathscr{C}_n} ||t(\tilde{S}^*_{n,\mathcal{C}})| - \mathrm{cv}_G(\alpha)| > \eta_\omega$, where $G = |\mathcal{C}|$.

For notation simplicity, let $s = \tilde{S}_{n,\mathcal{C}}$ and $s^* = \tilde{S}^*_{n,\mathcal{C}}$. By $\sup_{\mathcal{C}\in\mathscr{C}_n} |s - s^*| \to 0$ and uniform continuity of $\mathrm{se}(\cdot)$, there exists N_1 such that $\forall n \ge N_1$, $|(se)(s) - \mathrm{se}(s^*)| < \underline{\eta}_\omega/2$, so by **R2**(ii),

$$\mathrm{se}(s)\mathrm{se}(s^*) < \frac{\underline{\eta}_\omega^2}{2}. \tag{D.1}$$

By uniform continuity of $\bar{\cdot}$ and $\mathrm{se}(\cdot)$, there exists N_2 such that $\forall n \ge N_2$, we have

$$|\bar{s} - \bar{s}^*| < \frac{\underline{\eta}_\omega^2 \eta_\omega}{4(\overline{\eta}_\omega + M_\omega)} \tag{D.2}$$

and

$$|\mathrm{se}(s^*) - \mathrm{se}(s)| < \frac{\underline{\eta}_\omega^2 \eta_\omega}{4(\overline{\eta}_\omega + M_\omega)}. \tag{D.3}$$

Thus, $\forall n \ge N_\omega = \max\{N_1, N_2\}$, $\forall \mathcal{C} \in \mathscr{C}_n$,

$$\begin{aligned}
|t(s) - t(s^*)| &= \left|\frac{\bar{s}\cdot\mathrm{se}(s^*) - \bar{s}^*\cdot\mathrm{se}(s)}{\mathrm{se}(s)\mathrm{se}(s^*)}\right| \\
&\le \frac{|\bar{s}\cdot\mathrm{se}(s^*) - \bar{s}^*\cdot\mathrm{se}(s^*)| + |\bar{s}^*\cdot\mathrm{se}(s^*) - \bar{s}^*\cdot\mathrm{se}(s)|}{\mathrm{se}(s)\mathrm{se}(s^*)} \\
&\le \frac{|\bar{s} - \bar{s}^*|\cdot\mathrm{se}(s^*) + |\bar{s}^*|\cdot|\mathrm{se}(s^*) - \mathrm{se}(s)|}{\mathrm{se}(s)\mathrm{se}(s^*)} \\
&< \eta_\omega.
\end{aligned}$$

Thus, for n large enough, $|t(s)|$ and $|t(s^*)|$ are on the same side of the critical value. This concludes the proof.

□

D.2. Proof of Theorem 1

Proof. See the proof of Theorem 2 for verifying **H1**. This also implies **H2**. Therefore, Proposition 8 applies. □

Appendix E: Proof of Theorem 2

E.1. Supporting lemma

Lemma 1. *(Uniform Fatou's Lemma) Let $\{f_{n,\mathcal{C}}\}_{n\ge1,\mathcal{C}\in\mathscr{C}_n}$ be a set of measurable functions that are uniformly bounded, i.e. $\exists M$ s.t. $|f_{n,\mathcal{C}}| \le M$, $\forall n \ge 1, \forall \mathcal{C} \in \mathscr{C}_n$. Assume $\mathscr{C}_n$ is a finite set for each n. Then,*

$$\limsup_{n\to\infty} \sup_{\mathcal{C}\in\mathscr{C}_n} \int f_{n,\mathcal{C}} \le \int \limsup_{n\to\infty} \sup_{\mathcal{C}\in\mathscr{C}_n} f_{n,\mathcal{C}}.$$

Proof. Define $g_n = \sup_{m \geq n} \sup_{\mathcal{C} \in \mathscr{C}_m} f_{m,\mathcal{C}}$. Note that g_n is measurable, non-increasing, and bounded from below, so g_n converges pointwise and

$$\lim_{n \to \infty} g_n = \limsup_{n \to \infty} \sup_{\mathcal{C} \in \mathscr{C}_n} f_{n,\mathcal{C}}.$$

Also, the monotone convergence theorem implies

$$\lim_{n \to \infty} \int g_n = \int \lim_{n \to \infty} g_n = \int \limsup_{n \to \infty} \sup_{\mathcal{C} \in \mathscr{C}_n} f_{n,\mathcal{C}}.$$

Note $f_{n,\mathcal{C}} \leq g_n$ for each $\mathcal{C}$, so $\int f_{n,\mathcal{C}} \leq \int g_n$ for each $\mathcal{C}$ and thus

$$\sup_{\mathcal{C} \in \mathscr{C}_n} \int f_{n,\mathcal{C}} \leq \int g_n.$$

Combining results, we have

$$\limsup_{n \to \infty} \sup_{\mathcal{C} \in \mathscr{C}_n} \int f_{n,\mathcal{C}} \leq \limsup_{n \to \infty} \int g_n = \int \limsup_{n \to \infty} \sup_{\mathcal{C} \in \mathscr{C}_n} f_{n,\mathcal{C}}.$$

□

E.2. High level proposition for Theorem 2

This section presents a high level proposition and its assumptions for CRS with learned clustering. Suppose the vector of group-level estimator is $S_{n,\mathcal{C}}$ with clustering $\mathcal{C}$. The set of invariance transformation is $\mathcal{G}_\mathcal{C}$ and the test statistic function is T. Note that this allows for a sequence of clusterings with increasing number of groups, which is more general than the settings considered in this paper.

Assumption H3. (Invariance transformation) Let $\mathcal{G}_\mathcal{C}$ be the set of transformations associated with clustering structure $\mathcal{C}$, and T be the test statistic function.

(i) $\forall n \geq 1, \forall \mathcal{C} \in \mathscr{C}_n, \forall g \in \mathcal{G}_\mathcal{C}$, we have $g(S^*_{n,\mathcal{C}}) =_d S^*_{n,\mathcal{C}}$.
(ii) Let $\Omega_0 = \{T(gS^*_{n,\mathcal{C}}) \neq T(g'S^*_{n,\mathcal{C}}), \forall n \geq 1, \forall \mathcal{C} \in \mathscr{C}_n, \forall g, g' \in \mathcal{G}_\mathcal{C}, g \neq g'\}$, then $\mathrm{P}(\Omega_0) = 1$.
(iii) $\forall n \geq 1, \forall \mathcal{C} \in \mathscr{C}_n, \forall g \in \mathcal{G}_\mathcal{C}$, $T \circ g$ is uniformly continuous.
(iv) $\{g : \forall n \geq 1, \forall \mathcal{C} \in \mathscr{C}_n, g \in \mathcal{G}_\mathcal{C}\}$ is finite.

Proposition 9. *Suppose* **H1**, **H3**, *and* **R1** *hold. Then,*

$$\sup_{\mathcal{C} \in \mathscr{C}_n} |\mathrm{E}_{P_n}[\phi(S_{n,\mathcal{C}})] - \alpha| \to 0,$$

where ϕ is the result of the CRS procedure with transformation sets $\{\mathcal{G}_\mathcal{C}\}$ and test statistic function T.

Proof. Let $\{\tilde{S}_{n,\mathcal{C}}, \tilde{S}^*_{n,\mathcal{C}}\}$ and U be defined as in **H1**.

By **H3**(i) and Theorem 2.1 in [6], $\mathrm{E}_{P_n}[\phi(S^*_{n,\mathcal{C}})] = \alpha$. Since $\forall n \geq 1$ and $\forall \mathcal{C} \in \mathscr{C}_n$, $\tilde{S}_{n,\mathcal{C}} =_d S_{n,\mathcal{C}}$, $\tilde{S}^*_{n,\mathcal{C}} =_d S^*_{n,\mathcal{C}}$, we have $\mathrm{E}_{\mathbb{P}}[\phi(\tilde{S}_{n,\mathcal{C}}, U)] = \mathrm{E}_{P_n}[\phi(S_{n,\mathcal{C}})]$ and $\mathrm{E}_{\mathbb{P}}[\phi(\tilde{S}^*_{n,\mathcal{C}}, U)] = \mathrm{E}_{P_n}[\phi(S^*_{n,\mathcal{C}})]$. Thus,

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

Thus, it suffices to show $\sup_{\mathcal{C} \in \mathscr{C}_n} |\mathrm{E}_{\mathbb{P}}[\phi(\tilde{S}_{n,\mathcal{C}}, U) - \phi(\tilde{S}^*_{n,\mathcal{C}}, U)]| \to 0$.

Let $\mathrm{E}_{n,\mathcal{C}}$ be the event where the orderings of $\{T(g\tilde{S}_{n,\mathcal{C}}) : g \in \mathcal{G}_{\mathcal{C}}\}$ and $\{T(g\tilde{S}^*_{n,\mathcal{C}}) : g \in \mathcal{G}_{\mathcal{C}}\}$ correspond to the same transformations $g^{(1)}, \dots, g^{(|\mathcal{G}_{\mathcal{C}}|)}$. Then,

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

Thus, the goal is to show $\sup_{\mathcal{C} \in \mathscr{C}_n} \mathrm{E}_{\mathbb{P}}[\mathbb{1}_{\mathrm{E}^c_{n,\mathcal{C}}}] \to 0$. By **H1**, **H3**(ii), and **R1**(ii), there exist Ω with $\mathbb{P}(\Omega) = 1$ such that for each $\omega \in \Omega$,

(i) $\sup_{\mathcal{C} \in \mathscr{C}_n} \|\tilde{S}_{n,\mathcal{C}} - \tilde{S}^*_{n,\mathcal{C}}\| \to 0$;
(ii) $\inf_n \inf_{\mathcal{C} \in \mathscr{C}_n} \inf_{g \neq g'} |T(g(\tilde{S}^*_{n,\mathcal{C}})) - T(g'(\tilde{S}^*_{n,\mathcal{C}}))| > \delta_\omega > 0$;
(iii) $\forall n \geq 1, \forall \mathcal{C} \in \mathscr{C}_n, \forall g, g' \in \mathcal{G}_{\mathcal{C}}, g \neq g'$, we have

$$T(g\tilde{S}^*_{n,\mathcal{C}}) \neq T(g'\tilde{S}^*_{n,\mathcal{C}}).$$

Now fix $\omega \in \Omega$. For each $n \geq 1$ and $\mathcal{C} \in \mathscr{C}_n$, let $g^{(1)}_{n,\mathcal{C}}, \dots, g^{(|\mathcal{G}_{\mathcal{C}}|)}_{n,\mathcal{C}}$ be such that

$$T(g^{(1)}_{n,\mathcal{C}}(\omega)\tilde{S}^*_{n,\mathcal{C}}(\omega)) < \cdots < T(g^{(|\mathcal{G}_{\mathcal{C}}|)}_{n,\mathcal{C}}(\omega)\tilde{S}^*_{n,\mathcal{C}}(\omega)).$$

For some $g \in \mathcal{G}_{\mathcal{C}}$, by uniform continuity of $T \circ g$ as in **H3**(iii), there exists $\varepsilon_\omega(g) > 0$ such that $\|\tilde{S}_{n,\mathcal{C}}(\omega) - \tilde{S}^*_{n,\mathcal{C}}(\omega)\| < \varepsilon_\omega(g)$ implies $|T(g\tilde{S}^*_{n,\mathcal{C}}(\omega)) - T(g\tilde{S}^*_{n,\mathcal{C}}(\omega))| < \delta_\omega/2$, for δ_ω defined in **H3**(iii). Let $\varepsilon_\omega = \min\{\varepsilon_\omega(g) : \forall n \geq 1, \forall \mathcal{C} \in \mathscr{C}_n, \forall g \in \mathcal{G}_{\mathcal{C}}\}$, which is well-defined by **H3**(iv). Thus, $\|\tilde{S}_{n,\mathcal{C}}(\omega) - \tilde{S}^*_{n,\mathcal{C}}(\omega)\| < \varepsilon_\omega$ implies $|T(g\tilde{S}^*_{n,\mathcal{C}}(\omega)) - T(g\tilde{S}^*_{n,\mathcal{C}}(\omega))| < \delta_\omega/2$ for any $g \in \mathcal{G}_{\mathcal{C}}$.

Since $\omega \in \Omega_1$, there exists N_ω such that $\forall n \geq N_\omega$, we have $\sup_{\mathcal{C} \in \mathscr{C}_m} \|\tilde{S}_{n,\mathcal{C}}(\omega) - \tilde{S}^*_{n,\mathcal{C}}(\omega)\| < \varepsilon_\omega$. Therefore, $\forall n \geq N_\omega, \forall \mathcal{C} \in \mathscr{C}_n, \forall j = 1, \dots, |\mathcal{G}_{\mathcal{C}}| - 1$,

$$\begin{aligned}
T(g^{(j+1)}_{n,\mathcal{C}}(\omega)\tilde{S}_{n,\mathcal{C}}(\omega)) - T(g^{(j)}_{n,\mathcal{C}}(\omega)\tilde{S}_{n,\mathcal{C}}(\omega)) =& T(g^{(j+1)}_{n,\mathcal{C}}(\omega)\tilde{S}_{n,\mathcal{C}}(\omega)) - T(g^{(j+1)}_{n,\mathcal{C}}(\omega)\tilde{S}^*_{n,\mathcal{C}}(\omega)) \\
& + T(g^{(j+1)}_{n,\mathcal{C}}(\omega)\tilde{S}^*_{n,\mathcal{C}}(\omega)) - T(g^{(j)}_{n,\mathcal{C}}(\omega)\tilde{S}^*_{n,\mathcal{C}}(\omega)) \\
& + T(g^{(j)}_{n,\mathcal{C}}(\omega)\tilde{S}^*_{n,\mathcal{C}}(\omega)) - T(g^{(j)}_{n,\mathcal{C}}(\omega)\tilde{S}_{n,\mathcal{C}}(\omega)) \\
>& 0,
\end{aligned}$$

since the first and the third terms are smaller than $\delta_\omega/2$ in absolute value and the second term is greater than δ_ω. That implies $\omega \in \cap_{\mathcal{C} \in \mathscr{C}_n} \mathrm{E}_{n,\mathcal{C}}$, $\forall n \geq N_\omega$.

By Lemma 1,

$$\begin{aligned}\limsup_{n\to\infty}\sup_{\mathcal{C}\in\mathscr{C}_n}\mathrm{E}_{\mathbb{P}}[\mathbb{1}_{\mathrm{E}^c_{n,\mathcal{C}}}] &\le \mathrm{E}_{\mathbb{P}}\left[\limsup_{n\to\infty}\sup_{\mathcal{C}\in\mathscr{C}_n}\mathbb{1}_{\mathrm{E}^c_{n,\mathcal{C}}}\right]\\ &= \mathrm{E}_{\mathbb{P}}\left[\limsup_{n\to\infty}\mathbb{1}\{\omega\in\cup_{\mathcal{C}\in\mathscr{C}_n}\mathrm{E}^c_{n,\mathcal{C}}\}\right]\\ &= \mathrm{E}_{\mathbb{P}}\left[\limsup_{n\to\infty}\mathbb{1}\{\omega\in(\cap_{\mathcal{C}\in\mathscr{C}_n}\mathrm{E}_{n,\mathcal{C}})^c\}\right]\\ &= 0,\end{aligned}$$

so we obtain $\sup_{\mathcal{C}\in\mathscr{C}_n}\mathrm{E}_{\mathbb{P}}[\mathbb{1}_{\mathrm{E}^c_{n,\mathcal{C}}}]\to 0$ as $n\to\infty$. □

E.3. Proof of Theorem 2

Proof. To apply Proposition 9, we verify **H1** and **H3**. Note that for each n, $\mathscr{C}_n = \{\mathcal{C}_n^{(2)},\ldots,\mathcal{C}_n^{(\bar{G})}\}$ only has $\bar{G}-1$ clusterings. For each n, G corresponds to a unique $\mathcal{C}\in\mathscr{C}_n$. Thus, we let $S_{n,G}=S_{n,\mathcal{C}}$ for $G=|\mathcal{C}|$.

H1 Recall $S^*_{n,G}\sim N(0,\mathsf{Diag}(\tau_1^2,\ldots,\tau_G^2))$ with $\tau_g^2=\mathrm{var}(\sqrt{|\mathsf{C}_g|}(\widehat{\theta}_{n,g}-\theta_0))$ for some $\mathsf{C}_g\in\mathcal{C}$. Define the normalized version of $S_{n,G}$ and $S^*_{n,G}$ as

$$R_{n,G}=\left(\frac{\sqrt{|\mathsf{C}_g|}(\widehat{\theta}_{n,g}-\theta_0)}{\tau_g}\right)_{\mathsf{C}\in\mathcal{C}_n^{(G)}}=\mathsf{Diag}(\tau_1,\ldots,\tau_G)^{-1}S_{n,G}$$

and $R^*_G\sim N(0,I_G)$, where G is the number of groups in a certain clustering. Note that R^*_G is not a function of n. Then, Theorem 3 implies

$$R_{n,G}\xrightarrow{d}R^*_G,$$

as $n\to\infty$ for any fixed G. By the almost-sure representation theorem (see Theorem 2.19 in [34] for example), there exist a common probability measure $\mathbb{P}$, a scalar random variable U, and $\{\{\tilde{R}_{n,G}\}_{n\ge1},\tilde{R}^*_G\}_{G=1}^{\bar{G}}$ such that

(i) $\tilde{R}_{n,G}=_d R_{n,G}$, $\tilde{R}^*_G=_d R^*_G$,
(ii) $\|\tilde{R}_{n,G}-\tilde{R}^*_G\|\to 0$ with probability one for each G,
(iii) U is uniformly distributed and independent of everything else.

For each n and $\mathcal{C}\in\mathscr{C}_n$ with $|\mathcal{C}|=G$, let $\tilde{S}_{n,G}=\mathsf{Diag}(\tau_1,\ldots,\tau_G)\tilde{R}_{n,G}$ and $\tilde{S}^*_{n,G}=\mathsf{Diag}(\tau_1,\ldots,\tau_G)\tilde{R}^*_G$, i.e., $\tilde{S}_{n,G}$ and $\tilde{S}^*_{n,G}$ are non-normalized version of $\tilde{R}_{n,G}$ and $\tilde{R}^*_G$, respectively. Then,

$$\sup_{\mathcal{C}\in\mathscr{C}_n}|\tilde{S}_{n,G}-\tilde{S}^*_{n,G}|\le M\sup_{\mathcal{C}\in\mathscr{C}_n}|\tilde{R}_{n,G}-\tilde{R}^*_G|\to 0,$$

for some $M<\infty$ by **R1**(i).

H3 (i) Since $S^*_{n,\mathcal{C}}$ is a vector of zero-mean and independent normal random variables, flipping the signs of a set of the entries does not change the distribution of $S^*_{n,\mathcal{C}}$.

(ii) For some n, $\mathcal{C}$ with $|\mathcal{C}| = G$, and $g, g' \in \mathcal{G}_\mathcal{C}$ with $g \neq g'$, $V = \{v \in \mathbb{R}^G | T(g - g')v = 0\}$ is a linear subspace with dimensionality lower than G, so $\mathrm{P}(S^*_{n,\mathcal{C}} \in V) = 0$ by Gaussianity. Since we have only countable combinations of $(n, \mathcal{C}, g, g')$, there exists such Ω_2.

(iii) $T \circ g$ is a linear transformation on Euclidean spaces and thus uniformly continuous.

(iv) Note that for different clusterings $\mathcal{C}$ with the same number of groups G, the sets of transformations are the same. Thus, for some fixed $N > \bar{G}$,

$$\{g : \forall n \geq 1, \forall \mathcal{C} \in \mathscr{C}_n, g \in \mathcal{G}_\mathcal{C}\} = \cup_{\mathcal{C} \in \mathscr{C}_N} \mathcal{G}_\mathcal{C}$$

is finite, since both $\mathscr{C}_N$ and $\mathcal{G}_\mathcal{C}$ are finite.

□

References

[1] Arellano, M. (1987). Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics*, 49(4):431–434.

[2] Assoud, P. (1977). *Espaces Metriques, Plongements, Facteurs.* Doctoral Dissertation, Universite de Paris XI, 91405 Orsay France.

[3] Barrios, T., Diamond, R., Imbens, G. W., and Koleśar, M. (2012). Clustering, spatial correlations, and randomization inference. *Journal of the American Statistical Association*, 107(498):578–591.

[4] 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.

[5] Bester, C. A., Conley, T. G., Hansen, C. B., and Vogelsang, T. J. (2008). Fixed-b asymptotics for spatially dependent robust nonparametric covariance matrix estimators. Mimeo.

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

[7] 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.

[8] Conley, T. G. (1996). *Econometric Modelling of Cross-Sectional Dependence.* Ph.D. Dissertation, University of Chicago.

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

[10] Fama, E. F. and MacBeth, J. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, 81:607–636.

[11] Foerster, A. T., Sarte, P.-D. G., and Watson, M. W. (2011). Sectoral versus Aggregate Shocks: A Structural Factor Analysis of Industrial Production. *Journal of Political Economy*, 119(1):1–38.

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

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

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

[15] Jenish, N. and Prucha, I. (2007). Central limit theorems and uniform laws of large numbers for arrays of random fields. Mimeo.

[16] 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.

[17] Jenish, N. and Prucha, I. R. (2012). On spatial processes and asymptotic inference under near-epoch dependence. *Journal of Econometrics*, 170(1):178 – 190.

[18] Kelejian, H. H. and Prucha, I. (1999). A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review*, 40:509–533.

[19] 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.

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

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

[22] 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.

[23] Lee, L.-f. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial econometric models. *Econometrica*, 72:1899–1926.

[24] Lee, L.-f. (2007a). Gmm and 2sls estimation of mixed regressive, spatial autoregressive models. *Journal of Econometrics*, 137:489–514.

[25] Lee, L.-f. (2007b). Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. *Journal of Econometrics*, 140:333–374.

[26] Leung, M. (2019). Dependence-Robust Inference Using Randomized Subsampling. *SSRN Electronic Journal.*

[27] Naor, A. and Neiman, O. (2012). Assouads theorem with dimension independent of the snowflaking. *Rev. Mat. Iberoam.*, 28(4):11231142.

[28] Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708.

[29] Ozdagli, A. K. and Weber, M. (2017). Monetary Policy Through Production Networks: Evidence from the Stock Market. *SSRN Electronic Journal.*

[30] Pötscher, B. M. and Prucha, I. R. (1997). *Dynamic Nonlinear Econometric Models.* Springer Berlin Heidelberg, Berlin, Heidelberg.

[31] Semmes, S. (1999). Bilipschitz embeddings of metric spaces into euclidean spaces. *Publicacions Matemtiques*, 43(2):571–653.

[32] Song, K. (2016). Ordering-Free Inference from Locally Dependent Data.

[33] 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.

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