Combining Clusters for the Approximate Randomization Test

Chun Pong Lau*

February 6, 2025

Abstract

This paper develops procedures to combine clusters for the approximate randomization test proposed by Canay, Romano, and Shaikh (2017a). Their test can be used to conduct inference with a small number of clusters and imposes weak requirements on the correlation structure. However, their test requires the target parameter to be identified within each cluster. A leading example where this requirement fails to hold is when a variable has no variation within clusters. For instance, this happens in difference-in-differences designs because the treatment variable equals zero in the control clusters. Under this scenario, combining control and treated clusters can solve the identification problem, and the test remains valid. However, there is an arbitrariness in how the clusters are combined. In this paper, I develop computationally efficient procedures to combine clusters when this identification requirement does not hold. Clusters are combined to maximize local asymptotic power. The simulation study and empirical application show that the procedures to combine clusters perform well in various settings.

Keywords: Cluster-robust inference, approximate randomization test, local asymptotic power, difference-in-differences.

*Kenneth C. Griffin Department of Economics, The University of Chicago, ccplau@uchicago.edu.
I am grateful to Alex Torgovitsky, Max Tabord-Meehan, and Azeem Shaikh for their continued guidance and support throughout the project. I also thank Stéphane Bonhomme, Yong Cai, Chris Hansen, Xinran Li, Panos Toulis, and Myungkou Shin for helpful feedback and discussions, as well as participants of the Econometrics Advising Group for valuable comments. All errors are my own.

1 Introduction

It is common to have clustered data in economics where observations within each cluster are correlated with each other. The recent approximate randomization test proposed by Canay, Romano, and Shaikh (2017a, CRS) imposes weak assumptions on the dependence structure and can be used to conduct inference when there is a small number of clusters. However, a main requirement in CRS is that the target parameter has to be identified within each cluster. If this requirement is not satisfied, researchers can combine clusters in a way that ensures the target parameter can be identified within each combined cluster in order to apply CRS. Any method of combining clusters that solves the identification issue leads to a valid test. However, this creates ambiguity on how clusters should be combined. This paper addresses this issue and provides guidance on how to combine clusters.

The issue of not being able to identify the target parameter within each cluster is common in empirical settings. Difference-in-differences (DID) is a leading example of this identification issue within clusters. This is primarily because the treatment indicator equals zero in the control clusters. Some examples include Bloom et al. (2013) for firm-level treatment, Burde and Linden (2013) for village-level treatment, Dincecco and Katz (2016) for country-level treatment, Goodman (2016) for state-level treatment, and Alsan and Goldin (2019) for municipality-level treatment. In addition, if each cluster contains only one unit with multiple periods, then the identification issue arises in all clusters when cluster-by-cluster regressions are used. This is because there are more coefficients than observations per cluster when time fixed effects are included.

This paper makes two contributions to the literature. First, I analyze the local asymptotic power of CRS. Canay et al. (2017a) showed that their test controls size but did not study local asymptotic power. Analyzing the local asymptotic power of CRS is non-trivial because the number of clusters is fixed and does not go to infinity. Hence, standard large-sample analysis of local power for randomization tests, such as Hoeffding (1952), cannot be applied.

Second, I develop computationally efficient procedures to combine clusters by maximizing local asymptotic power. The procedures depend on the chosen significance level. I show that the procedure can be written as a binary linear program when the significance level is "small." For other significance levels, I develop a heuristic to combine clusters. The procedures can be solved quickly and can help eliminate randomness in deciding how to combine the clusters. In the simulation study and empirical exercise, I

find that the procedures perform well in various settings.

Similar to CRS, this paper is related to the literature on inference with a small number of clusters. Bertrand et al. (2004) pointed out the issue with standard inference when there is a small number of clusters. Different tests have been proposed in this literature apart from CRS. Some recent methods include Bester et al. (2011), Ibragimov and Müller (2010, 2016), and Hagemann (2022). The Wild cluster bootstrap popularized by Cameron et al. (2008) has been shown to be valid under strong homogeneity conditions when the number of clusters is fixed by Canay et al. (2021). The Wild cluster bootstrap has also been studied in other settings, such as MacKinnon and Webb (2017), Djogbenou et al. (2019), and Webb (2023). Cai et al. (2023) provide a users-guide for CRS. Cai (2024) studies finite sample properties of the sign test. Recently, Cao et al. (2022) proposed a data-driven procedure to partition observations into clusters and conduct valid inference. Their procedure focuses on spatially indexed data and assumes there exists a dissimilarity measure between observations. This paper is different in that I take the clusters as given and focus on combining the clusters when the target parameter cannot be identified within each cluster. This paper follows the above papers and uses the model-based approach in clustering. See Abadie et al. (2022) for a design-based analysis. For surveys on the literature of conducting inference with a fixed number of clusters, see, for instance, Cameron and Miller (2015), Conley et al. (2018), and MacKinnon et al. (2023).

The remainder of this paper is organized as follows. Section 2 reviews the approximate randomization test proposed by CRS and outlines two motivating examples for combining clusters. Section 3 analyzes the local asymptotic power of CRS in order to develop the necessary tools for combining clusters. Section 4 presents the data-driven procedures to combine clusters. Sections 5 and 6 present results from Monte Carlo simulations and an empirical application, respectively. Section 7 concludes. All proofs can be found in the appendix.

2 The test and motivating examples

2.1 Setup

Consider the following linear regression model of clustered data. Clusters are indexed by $j \in \mathcal{J} \equiv \{1, \ldots, q\}$ and units in the jth cluster are indexed by $i \in \mathcal{I}_{n,j} \equiv \{1, \ldots, n_j\}$:

$$Y_{i,j} = X_{i,j}'\beta + U_{i,j}, \tag{1}$$

where $Y_{i,j}$ is an outcome variable, $X_{i,j} \in \mathbb{R}^{d_x}$ is a vector of covariates, $U_{i,j} \in \mathbb{R}$ is a residual, and $\beta \in \mathbb{R}^{d_x}$ is an unknown parameter. The total number of observations is $n \equiv \sum_{j \in \mathcal{J}} n_j$.

The goal is to test the hypothesis:

$$H_0 : c'\beta = \lambda \quad \text{vs.} \quad H_1 : c'\beta \neq \lambda, \tag{2}$$

for a given vector $c \in \mathbb{R}^{d_x} \backslash \{0_{d_x}\}$ (where 0_{d_x} is a vector of d_x zeros) and $\lambda \in \mathbb{R}$ at level $\alpha \in (0, 1)$.

2.2 Review of the approximate randomization test

This section briefly reviews the approximate randomization test by CRS in the context of the model in Section 2.1. The following five-step procedure is taken from Cai et al. (2023).

Algorithm 2.1 (Algorithm 2.1 from Cai et al. (2023))**.**

Step 1: For each $j \in \mathcal{J}$, regress $Y_{i,j}$ on $X_{i,j}$ using only the observations from cluster j. Denote $\widehat{\beta}_{n,j}$ as the corresponding estimator of β for each $j \in \mathcal{J}$.

Step 2: For each $j \in \mathcal{J}$, construct the test statistic

$$T_n \equiv \left| \frac{1}{q} \sum_{j=1}^{q} \widehat{S}_{n,j} \right|, \tag{3}$$

where $\widehat{S}_{n,j} \equiv \sqrt{n_j}(c'\widehat{\beta}_{n,j} - \lambda)$.

Step 3: Define $\mathbb{G} \equiv \{1, -1\}^q$. For each $g \equiv (g_1, \ldots, g_q) \in \mathbb{G}$, define

$$T_n(g) \equiv \left| \frac{1}{q} \sum_{j=1}^{q} g_j \widehat{S}_{n,j} \right|. \tag{4}$$

Step 4: Compute the $(1-\alpha)$-quantile of $\{T_n(g) : g \in \mathbb{G}\}$ as

$$\widehat{\text{cv}}_n(1-\alpha) \equiv \inf \left\{ u \in \mathbb{R} : \frac{1}{|\mathbb{G}|} \sum_{g \in \mathbb{G}} \mathbb{1}[T_n(g) \leq u] \geq 1 - \alpha \right\}.$$

Step 5: Compute the test as

$$\phi_n \equiv \mathbb{1}[T_n > \widehat{\text{cv}}_n(1-\alpha)]. \tag{5}$$

CRS requires the clusters to satisfy a joint convergence in distribution requirement and the corresponding limiting random variables to be invariant to sign changes. With these weak assumptions, the test is valid for a fixed number of clusters as $n \longrightarrow \infty$ under the null hypothesis. See Section 3 in Canay et al. (2017a) and Section 4 in Cai et al. (2023) for more details. The assumptions are stated as follows.

Assumption 2.2. *Let $\{\widehat{\beta}_{n,j} : j \in \mathcal{J}\}$ be the estimators of β obtained from cluster-by-cluster regressions in Step 1 of Algorithm 2.1. Assume that:*

1. $\frac{\sqrt{n_j}}{\sqrt{n}} \longrightarrow \xi_j > 0$ *for any* $j \in \mathcal{J}$.
2. $\begin{pmatrix} \sqrt{n_1}(\widehat{\beta}_{n,1} - \beta) \\ \vdots \\ \sqrt{n_q}(\widehat{\beta}_{n,q} - \beta) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} S_1 \\ \vdots \\ S_q \end{pmatrix}$, *where* $S_j \sim N(0, \Sigma_j)$ *for each* $j \in \mathcal{J}$ *and* Σ_j *is positive definite.*
3. $S_j \perp\!\!\!\perp S_k$ *for any* $j, k \in \mathcal{J}$ *with* $j \neq k$.

Assumption 2.2.1 requires that none of the clusters have a negligible size. Assumption 2.2.2 requires that $\widehat{\beta}_{n,j}$ obtained from cluster-by-cluster regression is asymptotically normal. It is satisfied when there is weak dependence within clusters so that an appropriate central limit theorem applies. Note that Assumption 3.1 of Canay et al. (2017a) imposes a weaker version in that it requires the limiting random variable to be symmetric, not necessarily normal. Assumption 2.2.3 is a standard assumption that requires the clusters to be independent of each other.

Assumption 2.2 with $\mathcal{J}$ being the set of clusters is appropriate when the target parameter can be identified within each cluster. It is not immediately satisfied when there is an identification problem within clusters. However, suppose the clusters can be combined into a coarser level ω so that the target parameter is identified in each combined cluster. Then, Assumption 2.2 holds when $\mathcal{J}$ is replaced by ω. In addition, CRS controls size when $\mathcal{J}$ is replaced by ω.

Remark 2.3. The test statistic defined in (3) is unstudentized. When the target parameter in the hypothesis is a scalar, studentizing the test statistic or not does not affect the results of the test. For details, see Section 3 of Cai et al. (2023).

Remark 2.4. CRS also offers a slightly different version of the test that is randomized to ensure that the rejection probability equals the significance level α under the null. Algorithm 2.1 uses a nonrandomized version of the test so that there is no randomness in reporting the result. See Canay et al. (2017a) for simulations comparing the randomized and nonrandomized versions of the test. They showed that using the nonrandomized

version leads to a rejection probability slightly less than α.

2.3 Motivating examples: the need to combine clusters

Recall that Step 1 of Algorithm 2.1 requires the researcher to obtain $\widehat{\beta}_{n,j}$ using cluster-by-cluster regressions. This is not possible when the target parameter cannot be identified within each cluster. This section outlines two common empirical settings with this identification issue.

Example 2.5 (Clustered regression)**.** Consider the following simplified version of the linear model (1):

$$Y_{i,j} = \beta_0 + \beta_1 D_j + U_{i,j},$$

where $D_j \in \{0,1\}$ is a cluster-level treatment indicator and $\beta_0, \beta_1, U_{i,j} \in \mathbb{R}$. Let $\mathcal{J} = \mathcal{J}_0 \cup \mathcal{J}_1$ where $\mathcal{J}_0$ is the set of control clusters and $\mathcal{J}_1$ is the set of treated clusters, so that $D_j = \mathbb{1}[j \in \mathcal{J}_1]$ for all $j \in \mathcal{J}$. While it is possible to estimate (β_0, β_1) using all observations, it is not possible to do so using cluster-by-cluster regressions due to collinearity.

As suggested by CRS, researchers need to combine treated and control clusters in order to apply Algorithm 2.1. Note that any combination of the clusters does not affect the validity of CRS because the test relies on $n \longrightarrow \infty$, while holding the number of clusters q fixed. Under the scenario where $|\mathcal{J}_1| = |\mathcal{J}_0|$, one way is to pair one treated cluster with one control cluster in this example. △

Example 2.6 (Difference-in-differences)**.** Consider the following static two-way fixed effects specification of the DID model:

$$Y_{j,t} = \alpha_j + \phi_t + D_{j,t}\beta + U_{j,t},$$

where clusters are indexed by $j \in \mathcal{J} \equiv \{1, \ldots, q\}$, time is indexed by $t \in \mathcal{T} \equiv \{1, \ldots, T\}$, α_j is the cluster fixed effect, ϕ_t is the time fixed effect, and $D_{j,t} \in \{0,1\}$ is the treatment indicator. Similar to the last example, let $\mathcal{J} = \mathcal{J}_0 \cup \mathcal{J}_1$ where $D_{j,t} = 1$ for $j \in \mathcal{J}_1$ and $t > t_0$ for some $1 < t_0 < T$ and $D_{j,t} = 0$ otherwise.

Although $D_{j,t}$ has variation across $t \in \mathcal{T}$ within clusters for $j \in \mathcal{J}_1$, the other clusters are such that $D_{j,t} = 0$ for all $j \in \mathcal{J}_0$ and $t \in \mathcal{T}$. In addition, it is not possible to identify all parameters using cluster-by-cluster regression because there are more parameters than observations when the cluster and time fixed effects are included. Nevertheless, the identification problem can be solved by combining treated and control clusters. △

3 Local asymptotic power analysis

This section analyzes the local asymptotic power of CRS in order to develop the tools for combining the clusters in Section 4.

3.1 Notations

Let $\delta \in \mathbb{R}$ be a local parameter such that

$$\lambda = c'\beta + \frac{\delta}{\sqrt{n}}. \tag{6}$$

The object of interest is the limiting rejection probability of CRS under the local alternative as described in (6), i.e.,

$$\pi(\delta, \alpha) \equiv \lim_{n\to\infty} \mathbb{P}_\delta[T_n > \widehat{\text{cv}}_n(1-\alpha)], \tag{7}$$

where the critical value $\widehat{\text{cv}}_n(1-\alpha)$ is the $(1-\alpha)$-quantile of $\{T_n(g) : g \in \mathbb{G}\}$ as defined in Step 4 of Algorithm 2.1, and $\mathbb{P}_\delta$ is used to emphasize the dependence of the distribution of the data on the local parameter δ.

As mentioned in the introduction, the asymptotic framework here is $n \longrightarrow \infty$, while holding the number of clusters q fixed. With q being fixed, $|\mathbb{G}|$ is also fixed. Hence, the "randomization distribution" may not settle down even when $n \longrightarrow \infty$. Therefore, results on the large-sample properties of the local power for randomization tests such as Hoeffding (1952) cannot be applied here. See Remark 3.5 of Canay et al. (2017a) for more discussion.

In order to compute the local asymptotic power of CRS, note that there can be at most 2^{q-1} unique values in $\{T_n(g) : g \in \mathbb{G}\}$. This follows by the definition of $T_n(g)$ in equation (4) that

$$T_n(g) = \left|\frac{1}{q}\sum_{j=1}^{q} g_j \widehat{S}_{n,j}\right| = \left|\frac{1}{q}\sum_{j=1}^{q} (-g_j) \widehat{S}_{n,j}\right| = T_n(-g), \tag{8}$$

for any $g \in \mathbb{G}$. Therefore, I focus on the set of sign changes $\mathbb{G}_U \equiv \{g \in \mathbb{G} : g_1 = 1\}$ that give the unique values of $T_n(g)$ to avoid duplications as in (8). By construction, $|\mathbb{G}_U| = 2^{q-1}$. The critical value $\widehat{\text{cv}}_n(1-\alpha)$ can also be evaluated as the $(1-\alpha)$-quantile of $\{T_n(g) : g \in \mathbb{G}_U\}$ instead of $\{T_n(g) : g \in \mathbb{G}\}$.

For notational convenience, define $\mathbb{G}_{U,-1} \equiv \mathbb{G}_U \backslash \{1_q\}$, where $1_q \equiv (1, \ldots, 1)$ is the

identity transformation. In addition, let $L \equiv |\mathbb{G}_{U,-1}| = 2^{q-1} - 1$ and $\{\overline{g}_1, \ldots, \overline{g}_L\}$ be the set of all distinct sign changes in $\mathbb{G}_{U,-1}$.

3.2 Local asymptotic power

Following the discussion in the last section, the event $\{T_n > \widehat{\text{cv}}_n(1-\alpha)\}$ can be interpreted in terms of the sign changes in $\mathbb{G}_U$. In particular, this event collects all arrangements of $\{T_n(g) : g \in \mathbb{G}_U\}$ such that T_n is one of the largest $K \equiv \lfloor \alpha |\mathbb{G}_U| \rfloor$ terms. But recall in (4) that each $T_n(g)$ is defined as first summing the random variables $\widehat{S}_{n,j}$ over $j \in \mathcal{J}$ multiplied by the sign changes, and then taking the absolute value. Therefore, computing $\{T_n > \widehat{\text{cv}}_n(1-\alpha)\}$ is a non-trivial task because all the terms in $\{T_n(g) : g \in \mathbb{G}_U\}$ are dependent on each other.

Observe that computing the probability that T_n is the kth largest term in $\{T_n(g) : g \in \mathbb{G}_U\}$ amounts to comparing T_n with $T_n(g)$ for each $g \in \mathbb{G}_{U,-1}$. In particular, each comparison involves checking whether $T_n > T_n(g)$ or not for all $g \in \mathbb{G}_{U,-1}$. The following lemma shows that these comparisons can be expressed explicitly in terms of the partial sums of $\widehat{S}_{n,j}$ depending on the sign change $g \in \mathbb{G}_{U,-1}$.

Lemma 3.1. *For any $g, h \in \mathbb{G}$, let*

$$\mathcal{J}_{\text{same}}(g,h) \equiv \{j \in \mathcal{J} : g_j = h_j\}$$

be the set of indices such that the corresponding components of sign changes in g and h are the same and

$$\mathcal{J}_{\text{diff}}(g,h) \equiv \{j \in \mathcal{J} : g_j \neq h_j\} = \{j \in \mathcal{J} : g_j = -h_j\}$$

be the set of indices such that the corresponding components of sign changes in g and h are different. Then,

$$\begin{aligned}\{T_n(h) > T_n(g)\} = &\left\{ \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} > 0, \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} > 0 \right\} \\ &\bigcup \left\{ \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} < 0, \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} < 0 \right\}.\end{aligned} \tag{9}$$

The above lemma shows that each comparison of $T_n(h) > T_n(g)$ can be written as two disjoint events. The lemma helps to simplify the computation of local asymptotic power.

Next, recall from the beginning of this subsection that computing the local asymptotic

power requires computing the probability that T_n is the kth largest term in $\{T_n(g) : g \in \mathbb{G}_U\}$ for each $k = 1, \ldots, \lfloor \alpha |\mathbb{G}_U| \rfloor$. The next lemma considers the probability that T_n is less than a subset of terms from $\{T_n(g) : g \in \mathbb{G}_{U,-1}\}$ and is larger than the remaining terms. It uses Lemma 3.1 to represent this as a comparison using partial sums.

Lemma 3.2. *Let Assumption 2.2 hold. Let $\delta \in \mathbb{R}$ be a local alternative parameter as in (6). In addition, define the following notations.*

- *For each $g \in \mathbb{G}_{U,-1}$, define the partial sums*

$$V_{\text{same}}(g, \delta) \equiv \sum_{j \in \mathcal{J}_{\text{same}}(g, 1_q)} (Z_j + \xi_j \delta),$$
$$V_{\text{diff}}(g, \delta) \equiv \sum_{j \in \mathcal{J}_{\text{diff}}(g, 1_q)} (Z_j + \xi_j \delta),$$

  where $Z_j \equiv c' S_j$ for each $j \in \mathcal{J}$. See Assumption 2.2 for the definitions of ξ_j and S_j.

- *For each $g \in \mathbb{G}_{U,-1}$ and any $\mathcal{H} \subseteq \mathbb{G}_{U,-1}$, define*

$$\mathcal{F}^{(\ell)}(g, \mathcal{H}, \delta) \equiv \begin{cases} \{\kappa_\ell V_{\text{same}}(g, \delta) > 0,\ \kappa_\ell V_{\text{diff}}(g, \delta) < 0\} & , \ g \in \mathcal{H} \\ \{\kappa_\ell V_{\text{same}}(g, \delta) > 0,\ \kappa_\ell V_{\text{diff}}(g, \delta) > 0\} & , \ g \notin \mathcal{H} \end{cases},$$

  for $\ell \in \{1, 2\}$ with $\kappa_1 = 1$ and $\kappa_2 = -1$.

- *Define $\mathbb{M} \equiv \{1, 2\}^L$ and write each $m \in \mathbb{M}$ as $(m_1, \ldots, m_L)$.*

Let $\mathcal{H}_1$ be a subset of $\mathbb{G}_{U,-1}$ of size $k \neq L$ and $\mathcal{H}_2 \equiv \mathbb{G}_{U,-1} \backslash \mathcal{H}_1$. Then,

$$\lim_{n \to \infty} \mathbb{P}_\delta[T_n(h_1) > T_n \ \forall h_1 \in \mathcal{H}_1 \text{ and } T_n > T_n(h_2) \ \forall h_2 \in \mathcal{H}_2] = \sum_{m \in \mathbb{M}} \mathbb{P}\left[\bigcap_{l=1}^{L} \mathcal{F}^{(m_l)}(\overline{g}_l, \mathcal{H}_1, \delta) \right].$$

In the above lemma, the partial sums $V_{\text{same}}(g, \delta)$ and $V_{\text{diff}}(g, \delta)$ are based on the partial sums in (9). They are similar to those in Lemma 3.1 but taking $n \longrightarrow \infty$, and are evaluated under the local alternative as in (6). Since the comparisons are about T_n and $T_n(g)$ for $g \in \mathbb{G}_{U,-1}$, the lemma sets $h = 1_q$ directly from Lemma 3.1. They describe the event that T_n is larger than or smaller than another $T_n(\overline{g}_l)$ for $l = 1, \ldots, L$.

Next, the following is a mild assumption used to control the ties of $\{T_n(g) : g \in \mathbb{G}_U\}$.

Assumption 3.3. *Let $\mathbb{W} \equiv \{w = (w_1, \ldots, w_q) \in \mathbb{R}^q : w_j \neq 0$ for at least one $1 \leq j \leq q\}$.*

For any $w \in \mathbb{W}$ *and* $w_0 \in \mathbb{R}$,

$$w_0 + \sum_{j=1}^{q} w_j Z_j \neq 0,$$

with probability 1.

See Lemmas S.5.1 to S.5.3 in Canay et al. (2017b) on how Assumption 3.3 is related to various commonly-used test statistics. As they have stated in their lemmas, the assumption is satisfied when the distribution of Z_j is absolutely continuous with respect to the Lebesgue measure for each $j \in \mathcal{J}$.

The following theorem states the main result that expresses the local asymptotic power using the above two lemmas. The main idea of the theorem is to consider different ordering of the terms in $\{T_n(g) : g \in \mathbb{G}_U\}$ and keep those such that T_n is one of the $\lfloor \alpha |\mathbb{G}_U| \rfloor$ largest terms.

Theorem 3.4. *Consider the problem of testing* (2). *Let Assumptions 2.2 and 3.3 hold,* $\alpha \in (0,1)$, $K \equiv \lfloor \alpha |\mathbb{G}_U| \rfloor$, *and* $\delta \in \mathbb{R}$ *be a local alternative parameter as in* (6). *In addition, let* $\mathbb{H}_k$ *be the collection of all distinct size* k *subsets of* $\mathbb{G}_{U,-1}$. *Then, the local asymptotic power can be written as*

$$\pi(\delta, \alpha) = \sum_{k=1}^{K} \sum_{\mathcal{H} \in \mathbb{H}_{k-1}} \sum_{m \in \mathbb{M}} \mathbb{P}\left[\bigcap_{l=1}^{L} \mathcal{F}^{(m_l)}(\overline{g}_l, \mathcal{H}, \delta) \right], \tag{10}$$

where $\mathbb{M}$ *and* $\mathcal{F}^{(m_l)}(\overline{g}_l, \mathcal{H}, \delta)$ *are the same as defined in Lemma 3.2.*

In words, equation (10) considers the ordering of the terms in $\{T_n(g) : g \in \mathbb{G}_U\}$ and keeps those such that T_n is one of the $\lfloor \alpha |\mathbb{G}_U| \rfloor$ largest terms. Hence, there are three summations in the local asymptotic power expression. The first summation (over k) sums over the events that T_n is the kth largest term. For each k, the second summation (over $\mathcal{H}$) sums over different combinations of the events where $T_n(g) > T_n$ for all $g \in \mathcal{H} \subseteq \mathbb{G}_{U,-1}$ with $|\mathcal{H}| = k-1$. The third summation (over m) considers different combinations of the events $\{\mathcal{F}^{(m_l)}(\overline{g}_l, \mathcal{H}, \delta)\}_{l=1}^{L}$ as in Lemma 3.2. The expression in Theorem 3.4 involves computing the joint probability of multivariate normal distributions. Closed-form expressions may not be always available, but the expression can be computed numerically.

When $K = 1$, the expression in (10) simplifies greatly. The reason is that with $K = 1$, $\mathbb{H}_0 = \emptyset$, and the corresponding event $\{T_n > \widehat{\text{cv}}_n(1-\alpha)\}$ means that $T_n > T_n(g)$ for all $g \in \mathbb{G}_{U,-1}$. The following corollary shows that when $K = 1$, the local asymptotic power in Theorem 3.4 simplifies to a sum of two products of univariate normal cumulative distribution functions. Since $|\mathbb{G}_U| = 2^{q-1}$, the range of significance levels α such that $\lfloor \alpha |\mathbb{G}_U| \rfloor = 1$ is $\alpha \in [\frac{1}{2^{q-1}}, \frac{1}{2^{q-2}})$.

Corollary 3.5. *Consider the problem of testing* (2). *Let Assumptions 2.2 and 3.3 hold and* $\delta \in \mathbb{R}$ *be a local alternative parameter as in* (6). *For any* $\alpha \in [\frac{1}{2^{q-1}}, \frac{1}{2^{q-2}})$, *the local asymptotic power can be written as*

$$\pi(\delta, \alpha) = \pi_{\mathrm{L}}(\delta) + \pi_{\mathrm{R}}(\delta),$$

where $\sigma_j^2 \equiv c'\Sigma_j c$ *for each* $j \in \mathcal{J}$,

$$\pi_{\mathrm{L}}(\delta) \equiv \prod_{j \in \mathcal{J}} \Phi\left(-\frac{\xi_j \delta}{\sigma_j}\right),$$
$$\pi_{\mathrm{R}}(\delta) \equiv \prod_{j \in \mathcal{J}} \left[1 - \Phi\left(-\frac{\xi_j \delta}{\sigma_j}\right)\right].$$

A few useful properties of the local asymptotic power for $K = 1$ above are summarized below.

Property 3.6. *Consider the problem of testing* (2). *Let Assumptions 2.2 and 3.3 hold,* $\delta \in \mathbb{R}$ *be a local alternative parameter as in* (6), *and* $\alpha \in [\frac{1}{2^{q-1}}, \frac{1}{2^{q-2}})$. *The following properties hold.*

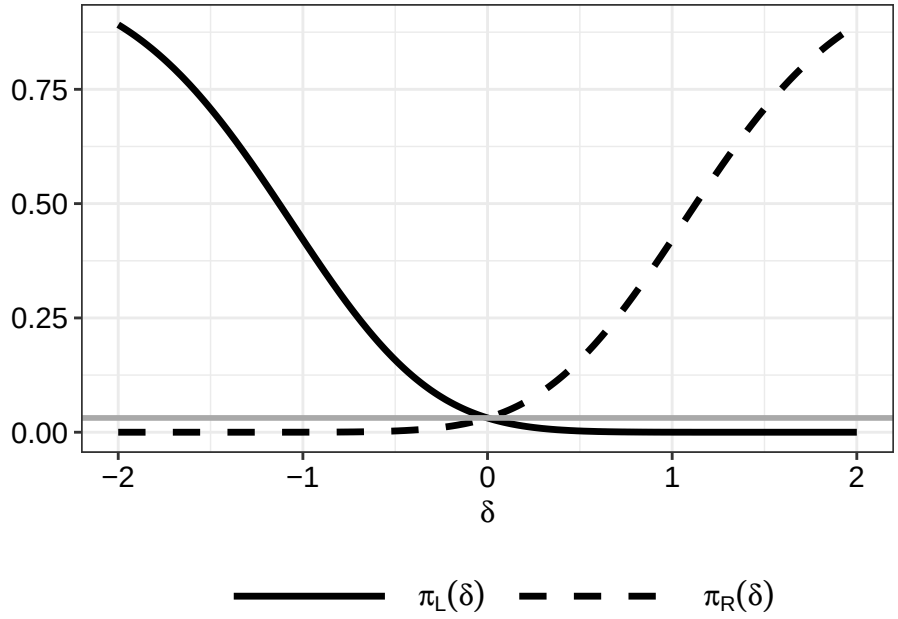
1. $\frac{\partial \pi_{\mathrm{L}}(\delta)}{\partial \delta} < 0$ *and* $\frac{\partial \pi_{\mathrm{R}}(\delta)}{\partial \delta} > 0$.
2. $\pi_{\mathrm{L}}(0) = \pi_{\mathrm{R}}(0) = \frac{1}{2^q}$.
3. (a) *If* $\delta < 0$, $\pi_{\mathrm{L}}(\delta) > \frac{1}{2^q} > \pi_{\mathrm{R}}(\delta)$.
   (b) *If* $\delta > 0$, $\pi_{\mathrm{R}}(\delta) > \frac{1}{2^q} > \pi_{\mathrm{L}}(\delta)$.

Property 3.6.1 shows that $\pi_{\mathrm{L}}(\delta)$ and $\pi_{\mathrm{R}}(\delta)$ move in opposite directions in δ. Property 3.6.2 shows that $\pi_{\mathrm{L}}(\delta)$ and $\pi_{\mathrm{R}}(\delta)$ intersect once at $\delta = 0$. Property 3.6.3 is implied by the first two properties. It states that when $\delta \neq 0$, then one of $\pi_{\mathrm{L}}(\delta)$ or $\pi_{\mathrm{R}}(\delta)$ is larger than the other. The smaller term is bounded above by $\frac{1}{2^q}$, which is usually a small value. Finally, the functions $\pi_{\mathrm{L}}(\delta)$ and $\pi_{\mathrm{R}}(\delta)$ can be interpreted as the local asymptotic power of one-sided tests.

Example 3.7. Suppose that $q = 5$, and $\frac{\xi_j}{\sigma_j} = 1$ for $j = 1, \ldots, 5$. Figure 1 plots $\pi_{\mathrm{L}}(\delta)$ and $\pi_{\mathrm{R}}(\delta)$ for $\delta \in [-2, 2]$ and demonstrates Property 3.6. The figure shows that $\pi_{\mathrm{L}}(\delta)$ is decreasing in δ and $\pi_{\mathrm{R}}(\delta)$ is increasing in δ. They intersect once at $\delta = 0$. In addition, $\pi_{\mathrm{L}}(\delta)$ is larger than $\pi_{\mathrm{R}}(\delta)$ when $\delta < 0$, and vice versa.

△

Figure 1: Illustration of the local asymptotic power function with $q = 5$.



4 Combining clusters

This section develops data-driven procedures to combine clusters based on the theory developed in Section 3. The procedures aim to find the best combination of clusters to maximize the local asymptotic power.

4.1 Assumptions and goal

Assume the clusters described by $\mathcal{J}$ can be classified into two groups such that the identification issue can be solved by combining one cluster from each group. This is formalized in the assumption below.

Assumption 4.1. *The set of clusters $\mathcal{J}$ is partitioned into nonempty sets $\mathcal{J}_0$ and $\mathcal{J}_1$ such that:*

1. $\mathcal{J}_0 \cup \mathcal{J}_1 = \mathcal{J}$.
2. $\mathcal{J}_0 \cap \mathcal{J}_1 = \emptyset$.
3. *For any $j_0 \in \mathcal{J}_0$ and $j_1 \in \mathcal{J}_1$, there is no identification issue in the pair of clusters $\{j_0, j_1\}$.*

In the case where treatments are at the cluster level, as in Examples 2.5 and 2.6, Assumption 4.1 is automatically satisfied with $\mathcal{J}_0$ being the set of control clusters and $\mathcal{J}_1$ being the set of treated clusters. For exposition purposes, I also assume that the numbers of clusters in $\mathcal{J}_0$ and $\mathcal{J}_1$ are equal. Hence, the goal of this section is to pair the

clusters. This assumption is relaxed in Section 4.4.

Assumption 4.2.

1. $|\mathcal{J}_0| = |\mathcal{J}_1| = \bar{q} \in \mathbb{N}$.
2. *Assume that the researcher is interested in forming $\bar{q}$ pairs of clusters, where each pair contains one cluster from $\mathcal{J}_0$ and one cluster from $\mathcal{J}_1$. Each cluster can only be used in one of the pairs. Denote Ω as the set of all such ways of pairing the clusters.*

Under Assumption 4.2.2, the size of Ω is $\bar{q}!$. For each method of combining clusters $\omega \in \Omega$, write $\omega \equiv \{\{1, \omega_1\}, \{2, \omega_2\}, \ldots, \{\bar{q}, \omega_{\bar{q}}\}\}$, where each pair $\{j, \omega_j\}$ represents the clusters $j \in \mathcal{J}_0$ and $\omega_j \in \mathcal{J}_1$ being grouped together. The following example illustrates the notations.

Example 4.3. Consider $q = 6$. Let $\mathcal{J}_0 = \{1, 2, 3\}$ be the control clusters and $\mathcal{J}_1 = \{4, 5, 6\}$ be the treated clusters. Then, there are $3! = 6$ different ways to pair the clusters. Here, Ω collects all different ways of pairing the clusters:

$$\begin{aligned}\Omega = \{&\{\{1,4\},\{2,5\},\{3,6\}\},\ \{\{1,5\},\{2,4\},\{3,6\}\},\ \{\{1,6\},\{2,4\},\{3,5\}\},\\ &\{\{1,4\},\{2,6\},\{3,5\}\},\ \{\{1,5\},\{2,6\},\{3,4\}\},\ \{\{1,6\},\{2,5\},\{3,4\}\}\}.\end{aligned}$$

For instance, if $\omega = \{\{1,5\},\{2,6\},\{3,4\}\}$, then $\omega_1 = 5$, $\omega_2 = 6$, and $\omega_3 = 4$ in terms of the notations in the preceding paragraph. △

As mentioned in Section 2.2, under the scenario where clusters have to be combined, the weak assumptions required on the clusters remain the same by replacing $\mathcal{J}$ in Assumption 2.2 with $\omega \in \Omega$. To emphasize the dependence on the pair of clusters $\{j, \omega_j\}$ from $\omega \in \Omega$ and that the pair of clusters are being considered instead of the individual clusters, the notations n_j, ξ_j, $\widehat{\beta}_{n,j}$, and σ_j that are indexed by j in the last two sections are updated to n_{j,ω_j}, ξ_{j,ω_j}, $\widehat{\beta}_{n,j,\omega_j}$, and σ_{j,ω_j} respectively.

With the above notations, the goal is to solve the following problem in a computationally efficient way:

$$\max_{\omega \in \Omega} \quad \pi(\omega, \delta, \alpha), \tag{11}$$

where $\pi(\omega, \delta, \alpha)$ is the local asymptotic power at significance level $\alpha \in (0, 1)$ with local alternative parameter $\delta \in \mathbb{R}$ from (6) for a given method of combining clusters $\omega \in \Omega$.

A simple approach to solve the optimization problem (11) would be to compute $\pi(\omega, \delta, \alpha)$ for each $\omega \in \Omega$. However, enumerating all possible groupings of clusters

and computing $\pi(\omega, \delta, \alpha)$ for each $\omega \in \Omega$ can be computationally intensive when $\overline{q}$ is large because there are $\overline{q}!$ different ways in the case of pairing clusters. This motivates the development of computationally efficient procedures for solving (11).

Let $\widehat{\pi}_n(\omega, \delta, \alpha)$ be the estimator of the objective function in (11). In addition, let $\widehat{\omega}_n$ be the solution to the sample analog of problem (11), i.e., using $\widehat{\pi}_n(\omega, \delta, \alpha)$ as the objective in the optimization problem. Note that Ω remains unchanged when solving the sample analog of (11). To evaluate $\widehat{\pi}_n(\omega, \delta, \alpha)$, it requires estimating parameters and variances. This means one has to estimate $\xi_{j,r}$ and $\sigma_{j,r}$ for all $j \in \mathcal{J}_1$ and $r \in \mathcal{J}_0$. The term $\xi_{j,r}$ is the ratio of cluster size. For $\sigma_{j,r}$, it requires the researcher to specify a working model on the dependence structure and apply a variance estimator. For instance, variance estimators are available for time-dependent data (such as Newey and West (1987)) or spatial data (such as Conley (1999)). Alternatively, one can also estimate variance using quasi-maximum likelihood estimation as in Cao et al. (2022). Note that even though $\sigma_{j,r}$ has to be estimated, these variance estimators are not used in Algorithm 2.1. Thus, specifying the working model and the variance estimator does not mean conducting inference using the variance estimator directly.

In establishing the validity of the data-driven procedure below, Theorem 4.7 ahead does not require the correct specification of the working model. However, the choice of the working model can affect the power performance as it affects the objective function of (11).

To justify the validity of the test that combines the clusters according to the optimal solution to the sample analog of (11), consider the following assumption that requires how the clusters are combined has a negligible impact on the test. This is similar to the condition imposed in Cao et al. (2022).

Assumption 4.4. *There exists $\widetilde{\Omega} \subseteq \Omega$ such that the following holds:*

1. $\lim_{n\to\infty} \mathbb{P}[\widehat{\omega}_n \notin \widetilde{\Omega}] = 0.$
2. $\lim_{n\to\infty} \sup_{\omega\in\widetilde{\Omega}} |\mathbb{P}[\phi_n(\widehat{\omega}_n) = 1 | \widehat{\omega}_n = \omega] - \mathbb{P}[\phi_n(\omega) = 1]| = 0.$

The above assumption can be satisfied in different settings. For example, this holds through sample-splitting where part of the data is used to obtain $\widehat{\omega}_n$, and another part is used to conduct inference. Note that Assumption 4.4 is also satisfied under some weak assumptions as follows.

Assumption 4.5. *Let $\delta \in \mathbb{R}$ be a local alternative parameter as in (6) and $\alpha \in (0, 1)$. Assume that the following statements hold:*

1. $\widehat{\pi}_n(\omega, \delta, \alpha) \xrightarrow{p} \pi(\omega, \delta, \alpha)$ *for each* $\omega \in \Omega$.
2. $\widetilde{\Omega} = \{\omega^\star\}$, *where* $\omega^\star \in \Omega$.
3. $\pi(\omega^\star, \delta, \alpha) > \pi(\omega, \delta, \alpha)$ *for each* $\omega \in \Omega \backslash \{\omega^\star\}$.

Assumption 4.5.1 is a mild condition requiring the estimator to converge for each $\omega \in \Omega$. As before, this assumption does not require correctly specifying the local asymptotic power function. Assumptions 4.5.2 and 4.5.3 require that $\omega^\star \in \Omega$ is the unique solution to (11). The following proposition shows that Assumption 4.4 is satisfied under these conditions.

Proposition 4.6. *Consider the problem of testing* (2). *Let Assumption 2.2 hold with* $\mathcal{J}$ *replaced by any* $\omega \in \Omega$, *Assumptions 4.1 and 4.2 hold. Then, Assumption 4.5 implies that Assumption 4.4 holds.*

The following theorem establishes the validity of the test that chooses $\widehat{\omega}_n \in \Omega$ by solving problem (11).

Theorem 4.7. *Consider the problem of testing* (2). *Let Assumption 2.2 hold with* $\mathcal{J}$ *replaced by any* $\omega \in \Omega$, *Assumptions 3.3, 4.1, 4.2, and 4.4 hold, and* $\alpha \in (0, 1)$. *Let* $\phi_n(\omega)$ *be the test in* (5) *that uses* $\omega \in \Omega$ *to combine the cluster. Then,*

$$\lim_{n \to \infty} \mathbb{P}[\phi_n(\widehat{\omega}_n) = 1] \leq \alpha,$$

under the null hypothesis.

The above theorem shows that the data-driven procedure controls size under some high-level conditions when the sample analog of problem (11) is used. The validity of the procedure does not require the local asymptotic power function to be correctly specified, although having an incorrect specification can lead to power loss.

4.2 Procedure for $K = 1$

This section develops a procedure to combine clusters for $K = 1$, i.e., when the significance level is chosen such that $\alpha \in [\frac{1}{2^{\bar{q}-1}}, \frac{1}{2^{\bar{q}-2}})$. The procedure utilizes the local asymptotic power derived in Corollary 3.5. Using the notations from Section 4.1, the local asymptotic power for $K = 1$ with a given $\omega \equiv \{\{1, \omega_1\}, \{2, \omega_2\}, \ldots, \{\bar{q}, \omega_{\bar{q}}\}\} \in \Omega$ and $\delta \in \mathbb{R}$ can be written as

$$\pi(\omega, \delta, \alpha) \equiv \prod_{j=1}^{\bar{q}} \Phi\left(-\frac{\xi_{j,\omega_j}\delta}{\sigma_{j,\omega_j}}\right) + \prod_{j=1}^{\bar{q}} \left[1 - \Phi\left(-\frac{\xi_{j,\omega_j}\delta}{\sigma_{j,\omega_j}}\right)\right], \tag{12}$$

for $\alpha \in [\frac{1}{2^{\bar{q}-1}}, \frac{1}{2^{\bar{q}-2}})$.

To develop a computationally feasible approach for solving optimization problem (11) with objective (12), recall from Assumption 4.2 that each cluster in $j \in \mathcal{J}_0$ is matched with a cluster in $r \in \mathcal{J}_1$. Moreover, each cluster from $\mathcal{J}_0$ and $\mathcal{J}_1$ is matched only once. As a result, there are only $\bar{q}$ possible terms $\sigma_{j,r}$ and $\xi_{j,r}$ for each $j \in \mathcal{J}_0$. Hence, there are $\bar{q}^2$ possible terms to compute in total. This is less than $\bar{q}!$ when $\bar{q} \geq 4$. See the example below.

Example 4.8. Consider $q = 8$. Let $\mathcal{J}_0 = \{1, 2, 3, 4\}$ be the control clusters and $\mathcal{J}_1 = \{5, 6, 7, 8\}$ be the treated clusters. For each $j \in \mathcal{J}_0$, there are four distinct parameters $\sigma_{j,r}$ to precompute over $r \in \mathcal{J}_1$. Hence, there are $4^2 = 16$ terms to precompute. Similarly, $\xi_{j,r}$ also has 16 possible combinations. △

Let $\Psi_{j,r} \equiv \Phi\left(-\frac{\xi_{j,r}\delta}{\sigma_{j,r}}\right)$ for each $j \in \mathcal{J}_0$ and $r \in \mathcal{J}_1$. There are $\bar{q}^2$ possible combinations of this term by the above reasoning. Following the above discussion, these terms can also be precomputed in advance. Let $z_{j,r} \in \{0, 1\}$ be an indicator variable that equals 1 if clusters $j \in \mathcal{J}_0$ and $r \in \mathcal{J}_1$ are paired together, and equals 0 otherwise. Then, the optimization problem (11) with objective (12) can be formulated as the following integer program:

$$\begin{aligned}
\max_{z_{j,r}} \quad & \prod_{j=1}^{\bar{q}} \sum_{r=1}^{\bar{q}} \Psi_{j,r} z_{j,r} + \prod_{j=1}^{\bar{q}} \sum_{r=1}^{\bar{q}} \left(1 - \Psi_{j,r}\right) z_{j,r} \\
\text{s.t.} \quad & \sum_{j=1}^{\bar{q}} z_{j,r} = 1 \qquad \text{for each } j = 1, \ldots, \bar{q} \\
& \sum_{r=1}^{\bar{q}} z_{j,r} = 1 \qquad \text{for each } r = 1, \ldots, \bar{q} \\
& z_{j,r} \in \{0, 1\} \qquad \text{for each } j = 1, \ldots, \bar{q} \text{ and } r = 1, \ldots, \bar{q}.
\end{aligned} \tag{13}$$

While the optimization problem (13) has transformed the optimization problem (11) with objective (12) into an integer program with linear constraints, the objective function is nonlinear in $z_{j,r}$. To obtain a computationally feasible program, recall from Property 3.6.3 that if $\delta < 0$, then

$$\prod_{j=1}^{\bar{q}} \sum_{r=1}^{\bar{q}} \Psi_{j,r} z_{j,r} > \frac{1}{2^{\bar{q}}} > \prod_{j=1}^{\bar{q}} \sum_{r=1}^{\bar{q}} \left(1 - \Psi_{j,r}\right) z_{j,r}, \tag{14}$$

where the relationship is reversed for $\delta > 0$. Since $z_{j,r}$ is a binary variable, taking

logarithm on inequality (14) gives

$$\sum_{j=1}^{\bar{q}}\sum_{r=1}^{\bar{q}} z_{j,r}\log\Psi_{j,r} > -\bar{q}\log 2 > \sum_{j=1}^{\bar{q}}\sum_{r=1}^{\bar{q}} z_{j,r}\log(1-\Psi_{j,r}). \tag{15}$$

Note that $\log\Psi_{j,r}$ and $\log(1-\Psi_{j,r})$ have been pre-computed in advance. Hence, the terms on the left-hand side and the right-hand side of (15) are linear in the unknowns $z_{j,r}$.

Next, inequality (14) implies that the RHS is in the small interval $[0, \frac{1}{2^{\bar{q}}}]$. The interval $[0, \frac{1}{2^{\bar{q}}}]$ can be partitioned into A subintervals $[\epsilon_0, \epsilon_1], [\epsilon_1, \epsilon_2], \ldots, [\epsilon_{A-1}, \epsilon_A]$ where $0 < \epsilon_0 < \epsilon_1 < \cdots < \epsilon_A = \frac{1}{2^{\bar{q}}}$. Since the log of RHS of (14) is considered in (15), the number ϵ_0 can be chosen to be $\min_{j,r\in\{1,\ldots,\bar{q}\}}(1-\Psi_{j,r})^{\bar{q}}$ or a very small number above 0 in order to have a finite value for $\log\epsilon_0$.

If $\delta < 0$, the optimization problem (13) can be approximated as solving multiple integer linear programs where each program maximizes the left-hand side of (15) and restricts the right-hand side of (15) in a small interval $[\log\epsilon_{a-1}, \log\epsilon_a]$ for each $a = 1, \ldots, A$:

$$\begin{aligned}
\max_{z_{j,r}} \quad & \sum_{j=1}^{\bar{q}}\sum_{r=1}^{\bar{q}} z_{j,r}\log\Psi_{j,r} \\
\text{s.t.} \quad & \sum_{j=1}^{\bar{q}} z_{j,r} = 1 && \text{for each } j = 1, \ldots, \bar{q} \\
& \sum_{r=1}^{\bar{q}} z_{j,r} = 1 && \text{for each } r = 1, \ldots, \bar{q} \\
& \sum_{j=1}^{\bar{q}}\sum_{r=1}^{\bar{q}} z_{j,r}\log\left(1-\Psi_{j,r}\right) \in [\log\epsilon_{a-1}, \log\epsilon_a] \\
& z_{j,r} \in \{0,1\} && \text{for each } j = 1, \ldots, \bar{q} \text{ and } r = 1, \ldots, \bar{q}.
\end{aligned} \tag{16}$$

Since (16) is merely a linear program with binary variables, solving A of them is fast even for large values of A. After solving the optimization problem (16) for each $a = 1, \ldots, A$, the solution can be chosen to be the one that gives the largest local asymptotic power. In practice, (16) is replaced with its sample analog, i.e., by replacing $\Psi_{j,r}$ with $\widehat{\Psi}_{j,r}$ for each $j \in \mathcal{J}_1$ and $r \in \mathcal{J}_0$. This is done by estimating $\xi_{j,r}$ and $\sigma_{j,r}$ as discussed in Section 4.1.

If $\delta > 0$, the procedure is the same except the 0-1 linear integer program in (16)

becomes the following program:

$$\begin{aligned}
\max_{z_{j,r}} \quad & \sum_{j=1}^{\bar{q}} \sum_{r=1}^{\bar{q}} z_{j,r} \log\left(1 - \Psi_{j,r}\right) \\
\text{s.t.} \quad & \sum_{j=1}^{\bar{q}} z_{j,r} = 1 && \text{for each } j = 1, \ldots, \bar{q} \\
& \sum_{r=1}^{\bar{q}} z_{j,r} = 1 && \text{for each } r = 1, \ldots, \bar{q} \\
& \sum_{j=1}^{\bar{q}} \sum_{r=1}^{\bar{q}} z_{j,r} \log \Psi_{j,r} \in [\log \epsilon_{a-1}, \log \epsilon_a] \\
& z_{j,r} \in \{0, 1\} && \text{for each } j = 1, \ldots, \bar{q} \text{ and } r = 1, \ldots, \bar{q}.
\end{aligned} \tag{17}$$

The procedure described in this section is summarized as follows.

Algorithm 4.9.

Step 1: Choose the local alternative parameter $\delta \in \mathbb{R}$. Estimate and precompute $\Psi_{j,r}$ for each $j, r = 1, \ldots, \bar{q}$. There are $\bar{q}^2$ terms to precompute.

Step 2: Form A subintervals of $[0, \frac{1}{2^{\bar{q}}}]$ as $[\epsilon_0, \epsilon_1], [\epsilon_1, \epsilon_2], \ldots, [\epsilon_{A-1}, \epsilon_A]$ where $0 < \epsilon_0 < \epsilon_1 < \cdots < \epsilon_A = \frac{1}{2^{\bar{q}}}$ and ϵ_0 is $\min_{j,r \in \{1,\ldots,\bar{q}\}} (1 - \Psi_{j,r})^{\bar{q}}$ or a smaller positive number.

Step 3: If $\delta < 0$, solve the sample analog of (16). If $\delta > 0$, solve the sample analog of (17). Solve the program for each subinterval $[\log \epsilon_{a-1}, \log \epsilon_a]$ where $a = 1, \ldots, A$. Let $\omega^{(a)}$ be the solution and $\pi^{(a)}$ be the local asymptotic power for the ath problem. Set $\pi^{(a)} = -\infty$ if the ath program is infeasible.

Step 4: Return $\omega^{(a^\star)}$ as the solution such that $a^\star = \arg\max_{a=1,\ldots,A} \pi^{(a)}$.

Step 5: Conduct inference using CRS as in Algorithm 2.1 with $\omega^{(a^\star)}$ as the clusters.

Algorithm 4.9 requires the researcher to choose a value of A in Step 2. The following proposition establishes that when A is large enough, then the approximation in Algorithm 4.9 yields the same solution as in the original optimization problem (13). Hence, by partitioning $[0, \frac{1}{2^{\bar{q}}}]$ into fine enough intervals, the solution from the approximation step coincides with the solution from (13).

Proposition 4.10. *Let $\pi^\star$ be the optimal value to optimization problem (13). There exists $\epsilon_0 > 0$ and a positive integer A_0 such that if $[\epsilon_0, \frac{1}{2^{\bar{q}}}]$ is partitioned to A_0 equally-spaced intervals, then the optimal value that corresponds to $a^\star$ in Step 4 of Algorithm 4.9 is $\pi^\star$.*

The above proposition assumes the researcher partitions the interval $[0, \frac{1}{2^{\overline{q}}}]$ into equally-spaced intervals. This assumption is for convenience purposes only as it seemed to be a natural way to run Algorithm 4.9.

This section ends with three remarks on the above data-driven procedure for combining clusters.

Remark 4.11 (Log-linearization of the objective function.)**.** Instead of solving A 0-1 integer linear programs as in Algorithm 4.9, one can approximate the objective function of (13) by taking the sum of the left-hand side and the right-hand side of (15) as the objective function. This gives the following 0-1 integer linear program:

$$\begin{aligned} \max_{z_{j,r}} \quad & \sum_{j=1}^{\overline{q}} \sum_{r=1}^{\overline{q}} z_{j,r} \log \Psi_{j,r} + \sum_{j=1}^{\overline{q}} \sum_{r=1}^{\overline{q}} z_{j,r} \log(1 - \Psi_{j,r}) \\ \text{s.t.} \quad & \sum_{j=1}^{\overline{q}} z_{j,r} = 1 \qquad \text{for each } j = 1, \ldots, \overline{q} \\ & \sum_{r=1}^{\overline{q}} z_{j,r} = 1 \qquad \text{for each } r = 1, \ldots, \overline{q} \\ & z_{j,r} \in \{0, 1\} \qquad \text{for each } j = 1, \ldots, \overline{q} \text{ and } r = 1, \ldots, \overline{q}. \end{aligned} \tag{18}$$

Optimization problem (18) has the advantage of only requiring to solve one optimization problem when compared to (16). However, the objective function of (18) may not be a good approximation of (13) because it weights the two summations equally. In fact, simulations show this can perform worse than randomly picking a grouping of clusters in each draw of data.

Remark 4.12 (Homogeneity of clusters)**.** When the control clusters are homogeneous such that $\Psi_{j,r} = \Psi_j$ across all $r \in \mathcal{J}_1$ for each $j \in \mathcal{J}_0$, how the clusters are combined is not going to affect the local asymptotic power. The same is also true if the treated clusters are homogeneous such that $\Psi_{j,r} = \Psi_r$ across all $j \in \mathcal{J}_0$ for each $r \in \mathcal{J}_1$.

Remark 4.13 (Step 3 of Algorithm 4.9)**.** If $\delta < 0$, it is possible that for some $a = 1, \ldots, A$, optimization problem (16) is infeasible. This happens when there are no combinations of $z_{j,r}$ such that the following constraints hold:

$$\sum_{j=1}^{\overline{q}} \sum_{r=1}^{\overline{q}} z_{j,r} \log\left(1 - \Psi_{j,r}\right) \in [\log \epsilon_{a-1}, \log \epsilon_a], \tag{19}$$

for a particular value of a. However, there must exist at least one $a' = 1, \ldots, A$ such

that (19) holds as long as ϵ_0 is small enough. This is because the summation in (19) is bounded above by $-\bar{q}\log 2$ by construction in (15). If $\delta > 0$, the same discussion applies to the constraint

$$\sum_{j=1}^{\bar{q}}\sum_{r=1}^{\bar{q}} z_{j,r}\log\Psi_{j,r} \in [\log\epsilon_{a-1}, \log\epsilon_a]$$

in optimization problem (17).

4.3 Procedure for $K > 1$

The previous section focuses on developing an algorithm that uses the local asymptotic power of $K = 1$ as the objective function. In practice, researchers may be interested in conducting a test that involves $K > 1$. The reason for having a simple 0-1 integer linear program in the previous section is because the local asymptotic power for $K = 1$ has a simple expression as stated in Corollary 3.5. The other cases require evaluating the expression in Theorem 3.4 that involves more complicated expressions but can be evaluated using numerical methods.

Similar to the discussion in Section 4.2 on the estimation of variances, evaluating the rejection probability in Theorem 3.4 requires the researcher to specify a working model on the dependence structure and apply a variance estimator. Recall that Theorem 4.7 does not impose any requirement on K. Hence, the choice of the working model and the variance estimator does not affect the validity of the procedure, but can affect the power performance.

Algorithm 4.9 is found to be effective in computing the optimal grouping of clusters for $K > 1$ in simulations. Hence, it can be used to obtain an initial solution. When $\bar{q}$ is large and $K > 1$, evaluating the local asymptotic power for each $\omega \in \Omega$ can be costly. The following algorithm provides a heuristic to proceed from the initial solution. The main idea is to start from the initial point provided by Algorithm 4.9. Then, one proceeds to see if it is possible to improve upon the existing solution using local search. This algorithm below is motivated by the 2-opt algorithm for the traveling salesman problem (see, for instance, Flood (1956), Croes (1958), and Lin and Kernighan (1973)).

Algorithm 4.14.

Step 1: Obtain an initial solution $\omega^{(1)} \equiv \{\{1, \omega_1^{(1)}\}, \{2, \omega_2^{(1)}\}, \ldots, \{\bar{q}, \omega_{\bar{q}}^{(1)}\}\}$ by running Algorithm 4.9. Let the corresponding local asymptotic power of the initial solution be $\pi^{(1)}$.

Step 2: Swap the assignment of every two pairs of clusters in $\omega^{(1)}$ and compute the local

asymptotic power after the swap. This means picking two distinct pairs of clusters in $\omega^{(1)}$, e.g., $\{i, \omega_i^{(1)}\}$ and $\{j, \omega_j^{(1)}\}$, swap their assignment as $\{i, \omega_j^{(1)}\}$ and $\{j, \omega_i^{(1)}\}$, keep the other assignments the same and recompute the local asymptotic power.

Step 3: If a better solution is found, i.e., there exists $\omega^{(2)}$ with local asymptotic power $\pi^{(2)}$ such that $\pi^{(2)} > \pi^{(1)}$, repeat Step 2 with $\omega^{(1)}$ replaced by $\omega^{(2)}$. Stop if there is no solution that results in a higher local asymptotic power.

Step 4: Let $\omega^\star$ be the solution to the problem. Conduct inference using CRS as in Algorithm 2.1 using $\omega^\star$ as the clusters.

Algorithm 4.14 presents a procedure by conducting pairwise swaps. This can be easily extended for swapping more groupings.

4.4 Extension

This section briefly discusses the case with an unequal number of clusters in $\mathcal{J}_0$ and $\mathcal{J}_1$ for Algorithm 4.9. Assume without loss of generality that $|\mathcal{J}_0| < |\mathcal{J}_1|$.

Following the notations in Section 4.1, let Ω be the collection of all methods of combining clusters. In addition, let $\mathcal{J}_0 = \{1, 2, \ldots, \overline{q}\}$. Hence, the goal is to form $\overline{q}$ groups of clusters, where each group has one cluster from $\mathcal{J}_0$, and at least one cluster from $\mathcal{J}_1$. Moreover, all clusters from $\mathcal{J}_1$ have to be used. As before, each cluster can only be used once in the grouping. Therefore, each $\omega \in \Omega$ can be written as $\{\{1, \omega_1\}, \{2, \omega_2\}, \ldots, \{\overline{q}, \omega_{\overline{q}}\}\}$ as before, with the exception that ω_j can represent more than one cluster from $\mathcal{J}_1$ here for each $j = 1, \ldots, \overline{q}$. Let $\mathcal{M}$ be the set of all possible ways of partitioning clusters in $\mathcal{J}_1$ into $\overline{q}$ groups, so that $\omega_j \in \mathcal{M}$ for each $j = 1, \ldots, \overline{q}$, and that each $m \in \mathcal{M}$ is nonempty and satisfies $m \subset \mathcal{J}_1$. For example, if $\mathcal{J}_0 = \{1, 2\}$ and $\mathcal{J}_1 = \{3, 4, 5\}$ and one wishes to form two groups of clusters, the set $\mathcal{M}$ can be written as $\mathcal{M} = \{\{3\}, \{4\}, \{5\}, \{3, 4\}, \{3, 5\}, \{4, 5\}\}$.

Similar to Section 4.2, let $z_{j,m} \in \{0, 1\}$ be a binary variable that equals 1 when cluster $j \in \mathcal{J}_0$ is combined with $m \in \mathcal{M}$. Define $\Psi_{j,m} \equiv \Psi\left(-\frac{\xi_{j,m}\delta}{\sigma_{j,m}}\right)$, where $\xi_{j,m}$ is the ratio of cluster size and $\sigma_{j,m}^2$ is the asymptotic variance in the group of cluster $\{j, m\}$, where $j \in \mathcal{J}_0$ and $m \in \mathcal{M}$.

For each $a = 1, \ldots, A$ and the corresponding interval $[\log \epsilon_{a-1}, \log \epsilon_a]$, the optimization problem analogous to problem (16) for $\delta < 0$ in the scenario of $|\mathcal{J}_0| < |\mathcal{J}_1|$ is as

follows:

$$
\begin{aligned}
\max_{z_{j,m}} \quad & \sum_{j=1}^{\bar{q}} \sum_{m \in \mathcal{M}} z_{j,m} \log \Psi_{j,m} \\
\text{s.t.} \quad & \sum_{m \in \mathcal{M}} z_{j,m} = 1 && \text{for each } j = 1, \ldots, \bar{q} \\
& \sum_{j=1}^{\bar{q}} z_{j,m} = 1 && \text{for each } m \in \mathcal{M} \\
& \sum_{j=1}^{\bar{q}} \sum_{m \in \mathcal{M}} z_{j,m} \log \left(1 - \Psi_{j,m}\right) \in [\log \epsilon_{a-1}, \log \epsilon_a] \\
& z_{j,m} \in \{0, 1\} && \text{for each } j = 1, \ldots, \bar{q} \text{ and } m \in \mathcal{M}.
\end{aligned}
$$

Similarly, if $\delta > 0$, the optimization problem becomes

$$
\begin{aligned}
\max_{z_{j,m}} \quad & \sum_{j=1}^{\bar{q}} \sum_{m \in \mathcal{M}} z_{j,m} \log \left(1 - \Psi_{j,m}\right) \\
\text{s.t.} \quad & \sum_{m \in \mathcal{M}} z_{j,m} = 1 && \text{for each } j = 1, \ldots, \bar{q} \\
& \sum_{j=1}^{\bar{q}} z_{j,m} = 1 && \text{for each } m \in \mathcal{M} \\
& \sum_{j=1}^{\bar{q}} \sum_{m \in \mathcal{M}} z_{j,m} \log \Psi_{j,m} \in [\log \epsilon_{a-1}, \log \epsilon_a] \\
& z_{j,m} \in \{0, 1\} && \text{for each } j = 1, \ldots, \bar{q} \text{ and } m \in \mathcal{M}.
\end{aligned}
$$

The above optimization problems are binary integer programs as Section 4.2, and are fast to solve.

5 Monte Carlo simulations

5.1 Data generating process

This section explores the finite-sample performance of the data-driven procedures to combine clusters in Section 4. Here, I consider the data generating process (DGP) based on the simulation design in Hagemann (2022), which is a version of the simulation design in Conley and Taber (2011).

Let there be $q = 12$ clusters, which consists of 6 treated clusters represented by $\mathcal{J}_1 = \{1, 2, \ldots, 6\}$ and 6 control clusters represented by $\mathcal{J}_0 = \{7, 8, \ldots, 12\}$. In the following, $j \in \mathcal{J} \equiv \mathcal{J}_1 \cup \mathcal{J}_0$ indices all the clusters and $t \in \mathcal{T} \equiv \{1, 2, \ldots, T\}$ indices time. Let $D_{t,j} \equiv \mathbb{1}[j \in \mathcal{J}_1 \text{ and } t > t_0]$ for $j \in \mathcal{J}$ and $t \in \mathcal{T}$ and $I_t \equiv \mathbb{1}[t > t_0]$ for $t \in \mathcal{T}$. Each simulated data is generated from the following model:

$$\begin{aligned} Y_{t,j} &= \theta_0 I_t + \beta D_{t,j} + \gamma_1 X_{1,t,j} + \gamma_2 X_{2,t,j} + \gamma_3 X_{3,t,j} + \xi_j + U_{t,j}, \\ U_{t,j} &= \rho U_{t-1,j} + V_{t,j}, \\ X_{1,t,j} &= \gamma_4 I_t D_{j,t} + W_{t,j}, \end{aligned}$$

with $\theta_0 = \gamma_1 = \gamma_2 = \gamma_3 = \xi_j = 1$, $\rho = 0.5$, $\gamma_4 = 0.8$, $t_0 = 10$, and $T = 20$ as in Hagemann (2022). I consider three different DGPs below. DGP 1 generates the random variables in the same way as Hagemann (2022). DGPs 2 and 3 have different variance structures and assume there is the same number of heterogenous treated and control clusters. In all the specifications, the parameter $h \in \{1, 2, 3, 4\}$ governs the amount of heterogeneity in data, where a larger value of h represents a higher degree of heterogeneity. The goal is to test the two-sided hypothesis $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$. I consider two significance levels $\alpha = 0.05$ and $\alpha = 0.1$ in order to study the finite-sample properties of Algorithms 4.9 and 4.14 separately. The specification for each $j \in \mathcal{J}$ are as follows.

DGP 1. $X_{2,t,j}, X_{3,t,j}, V_{t,j}, W_{t,j} \overset{\text{i.i.d.}}{\sim} N(0, \sigma_j^2)$, with

$$\sigma_j = \begin{cases} 20 & , \quad j \geq 13 - h \\ 1 & , \quad \text{otherwise} \end{cases}.$$

DGP 2. Same as DGP 1, but

$$\sigma_j = \begin{cases} 5 + 3(j \bmod 6) & , \quad (j \bmod 6) \leq h - 1 \\ 1 & , \quad \text{otherwise} \end{cases}.$$

DGP 3. Same as DGP 1, but

$$\sigma_j = \begin{cases} 2.5^{1+(j \bmod 6)} & , \quad (j \bmod 6) \leq h - 1 \\ 1 & , \quad \text{otherwise} \end{cases}.$$

For each simulation design, I estimate the rejection rate obtained from the procedure

in Section 4 and use CRS to conduct inference for each of the $6! = 720$ different methods of combining clusters. Apart from CRS, I also apply other recent methods for conducting inference with a fixed number of clusters that were mentioned in Section 1. The methods and their abbreviations are summarized in Table 1. Some details of considering the two versions of CRS in Table 1 are as follows. First, **CRS-Data** is the data-driven procedure that uses optimization problem (16) for $\delta < 0$ and problem (17) for $\delta > 0$ when $\alpha = 0.05$. For these two optimization problems, I partition the interval $[0, \frac{1}{2^5}]$ into $A = 200$ subintervals. When $\alpha = 0.1$, the heuristic in Algorithm 4.14 is used. One of the two programs (16) and (17) is used to obtain an initial solution, depending on the sign of δ. Next, **CRS-Random** chooses one of the 720 ways of combining clusters randomly in each draw. This aims to represent the scenario where the researcher chooses a random grouping of clusters that can solve the identification problem for each data set. For each specification, I use 20,000 Monte Carlo simulations. The data-driven procedure **CRS-Data** uses the local alternative parameter $\delta = 2\sqrt{qT}$ under $\beta \geq 0$ and $\delta = -2\sqrt{qT}$ under $\beta < 0$. I consider $\beta \in \{-3, -2, \ldots, 3\}$ in the simulation exercises.

Table 1: Abbreviations for the methods.

Various versions of CRS	
CRS-Data	The data-driven procedure as in Algorithm 4.9 for $\alpha = 0.05$ and Algorithm 4.14 for $\alpha = 0.1$.
CRS-Random	Randomly choose one way of combining clusters in each draw of data.
Other methods	
BCH	The test in Bester et al. (2011) that uses the cluster covariance matrix estimator.
H	The adjusted permutation test by Hagemann (2022).
IM	The t-test in Ibragimov and Müller (2016).
WCB	The Wild cluster bootstrap (Cameron et al., 2008).

5.2 Simulation results

Figure 2 presents the rejection rate curves for various DGPs, heterogeneity parameters h, and methods for conducting inference at significance level $\alpha = 0.05$. Each column corresponds to one of the three DGPs, with different amounts of heterogeneity. In each figure, the grey region is formed by overlapping 720 different rejection rate curves based on each of the different ways of combining clusters, and conducting inference using CRS.

The other colored lines correspond to various methods as indicated in the legend with the abbreviations specified in Table 1.

For DGP 1 (column 1 of Figure 2), the treated clusters are homogeneous for the different values of h because the heterogeneity in variance only affects the control clusters. Hence, how the clusters are combined should not affect the population local asymptotic power as discussed in Remark 4.12. Indeed, both **CRS-Data** and **CRS-Random** give similar performance in finite-samples. On the other hand, BCH and WCB over-rejects under the null when $h = 4$.

For the other DGPs (columns 2 and 3 of Figure 2), there is heterogeneity in both the treated and control clusters. Hence, there is more variation in the CRS rejection rates depending on how the clusters are combined. CRS controls size in all designs. In addition, **CRS-Data** (as indicated by the solid black lines) always belong to the top regions formed by the grey curves. This shows that Algorithm 4.9 can lead to close to oracle performance in the simulations. Besides, **CRS-Data** also performs better than **CRS-Random** (as indicated by the solid blue lines). This shows that the data-driven procedures using (16) or (17) perform better than choosing a random grouping of clusters. On the other hand, WCB can over-reject in designs with more heterogeneity.

Next, I examine the performance of Algorithm 4.14 by considering $\alpha = 0.1$ for the same set of simulations. The results are summarized in Figure 3. The observations for $\alpha = 0.1$ are similar to those when $\alpha = 0.05$ as in Figure 2. CRS controls size in all designs. **CRS-Data** continues to be able to lead to a close-to-oracle grouping of clusters as the solid black lines belong to the top parts of the grey regions. In addition, **CRS-Data** compares favorably to other methods in various DGPs and heterogeneity parameters.

To conclude, the simulation results in this section show that the data-driven procedures to combine clusters perform well under various settings.

6 Empirical application

6.1 Setup

Dincecco and Katz (2016) study the impact of fiscal and administrative powers of states on economic performance, using data from 11 European countries. Table 3 of Dincecco and Katz (2016) regresses real per capita GDP growth on two binary variables representing fiscal centralization and limited government, together with other variables and fixed effects. Their results are clustered at the country level.

Figure 2: Rejection rate curves for the simulations at significance level $\alpha = 0.05$.

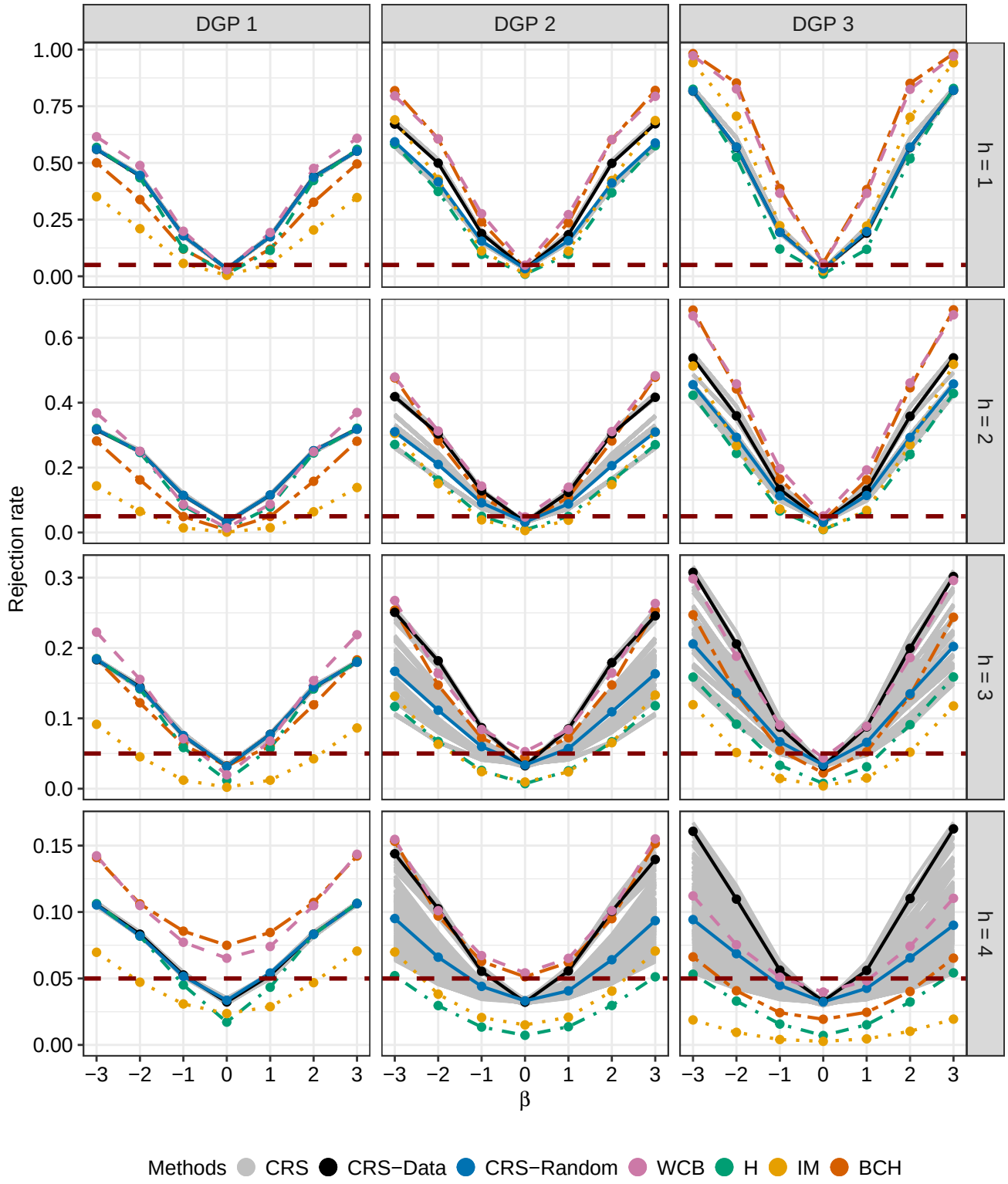
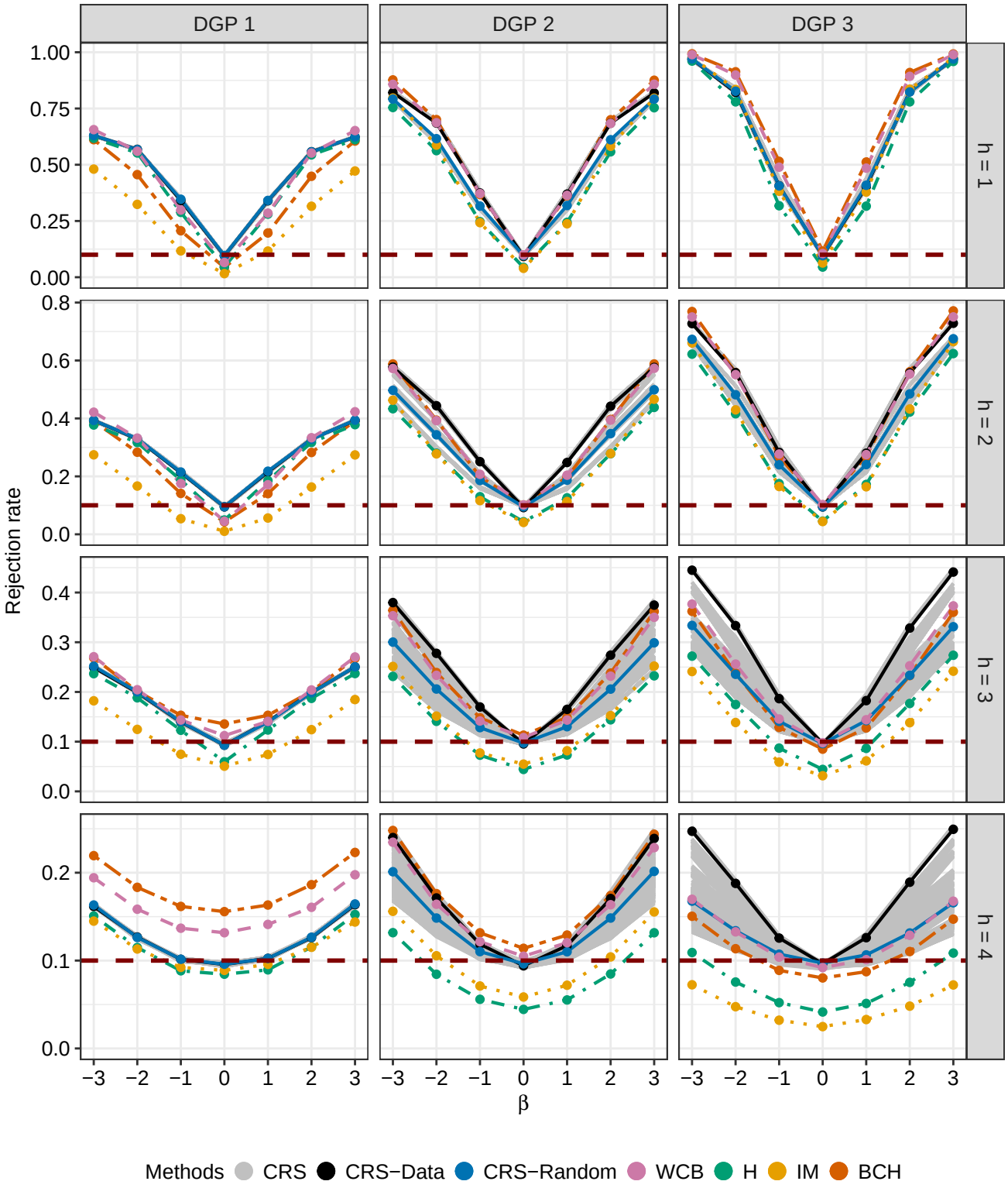


Figure 3: Rejection rate curves for the simulations at significance level $\alpha = 0.1$.



The baseline regression they consider is:

$$Y_{t,j} = \beta_0 + \beta_1 C_{t,j} + \beta_2 L_{t,j} + \mu_j + U_{t,j}, \tag{20}$$

where j indices country, t indices year, $Y_{t,j}$ is the (logarithm) annual growth rate of real per capita GDP in country j between years $t-1$ and t, $C_{t,j} \in \{0,1\}$ equals 1 for fiscal centralization in country j and year t, $L_{t,j} \in \{0,1\}$ equals 1 for limited government in country j and year t, and μ_j is the country fixed-effects. The goal is to conduct inference on $\widehat{\beta}_1$ and $\widehat{\beta}_2$.

I focus on columns (1) to (3) of Table 3 in Dincecco and Katz (2016) in order to focus on the issue that there exist clusters in which the binary variables $C_{t,j}$ or $L_{t,j}$ do not have variation over time. The specifications for the three columns are as follows.

Column (1). As in (20).

Column (2). As in column (1) and with year fixed effects.

Column (3). As in column (2) and with country-specfic time trends.

In this situation of having only 11 clusters, CRS can be used to conduct valid inference. However, there are clusters with no variation in the main variables of interest. There are three countries with $C_{t,j} = 1$ in all periods (Belgium, England, and Piedmont). There are two countries with $L_{t,j} = 1$ in all periods (Belgium and Denmark). As a result, researchers would have to combine clusters in order to ensure that all parameters can be identified in each combined cluster. In addition, the data is at the aggregate level in that there is one observation for each year in each cluster. Therefore, researchers would also have to combine clusters in the presence of time fixed effects. In this section, I form five groups of clusters across all specifications.

In order to study the performance of the data-driven procedures in Section 4, this section performs a calibrated simulation exercise. I first estimate the model using the actual data. Then, I simulate data by using various distributions of the error terms. In each specification, let $\widehat{\beta}_1$ and $\widehat{\beta}_2$ be the estimates of β_1 and β_2 obtained from data respectively. I take them as the true values of β_1 and β_2 in the calibrated simulation procedure. In addition, I perform inference on the coefficients of the two binary variables fiscal centralization $C_{t,j}$ and limited government $L_{t,j}$ separately. Hence, in conducting calibrated simulation exercises for the coefficient on $C_{t,j}$, I set $\beta_1 = \widehat{\beta}_1$ under the null, $\beta_1 = \widehat{\beta}_1 + \Delta$ under the alternative, and keep $\beta_2 = \widehat{\beta}_2$. Similarly, for the coefficient on $L_{t,j}$, I set $\beta_2 = \widehat{\beta}_2$ under the null, $\beta_2 = \widehat{\beta}_2 + \Delta$ under the alternative, and keep $\beta_1 = \widehat{\beta}_1$. The significance level is $\alpha = 0.1$ for all simulations. There are 9,660 different ways to combine

the clusters in this calibrated simulation exercise. I use 2,000 Monte Carlo simulations for each column and specification. I consider $\Delta \in \{-3, -2, \ldots, 3\}$.

In the calibrated simulation exercise, I first use the residuals $\{U_{t,j}\}$ to fit an AR(1) model

$$U_{t,j} = \rho_j U_{t,j-1} + \epsilon_{t,j},$$

where $\epsilon_{t,j} \sim N(0, v_j^2)$ for each country $j \in \mathcal{J}$. Using the estimated parameters $\{\widehat{\rho}_j\}_{j\in\mathcal{J}}$ and $\{\widehat{v}_j\}_{j\in\mathcal{J}}$, I consider the following specifications of the error terms in conducting the calibrated simulation exercise.

Specification (1). Use the original $\widehat{v}_j$ for each country $j \in \mathcal{J}$.

Specification (2). Same as specification 1, but replace $\widehat{v}_j$ by $10\widehat{v}_j$ for $j \leq 4$.

Specification (3). Same as specification 1, but replace $\widehat{v}_j$ by $10\widehat{v}_j$ for $j \leq 4$ and $\widehat{v}_j$ by $5\widehat{v}_j$ for $5 \leq j \leq 8$.

More details of the calibrated simulation exercise can be found in Appendix C.

6.2 Results

Figures 4 and 5 report the results for inference on the coefficients of the variables $C_{t,j}$ and $L_{t,j}$ respectively. Each figure reports the rejection rates against Δ, where Δ was defined two paragraphs ago. $\Delta = 0$ corresponds to the results under the null. Based on the findings in Section 5 that CRS has favorable performance across various designs, I focus on examining the performance of the data-driven approach here. See Table 1 for the definitions of **CRS-Data** and **CRS-Random**. Similar to Section 5, the grey regions are formed by the 9,660 rejection rate curves that are based on the different ways of forming five groups out of 11 clusters. As in the previous section, the data-driven procedure **CRS-Data** uses the local alternative parameter $\delta = 2\sqrt{qT}$ under $\Delta \geq 0$ and $\delta = -2\sqrt{qT}$ under $\Delta < 0$, where T is the number of periods in the calibrated simulation exercise.

The simulation results show that **CRS-Data** controls size and chooses powerful grouping of clusters in various designs. It also continues to perform better than choosing a random grouping of clusters.

Figure 4: Rejection rate curves for inference on the coefficient for fiscal centralization $C_{t,j}$.

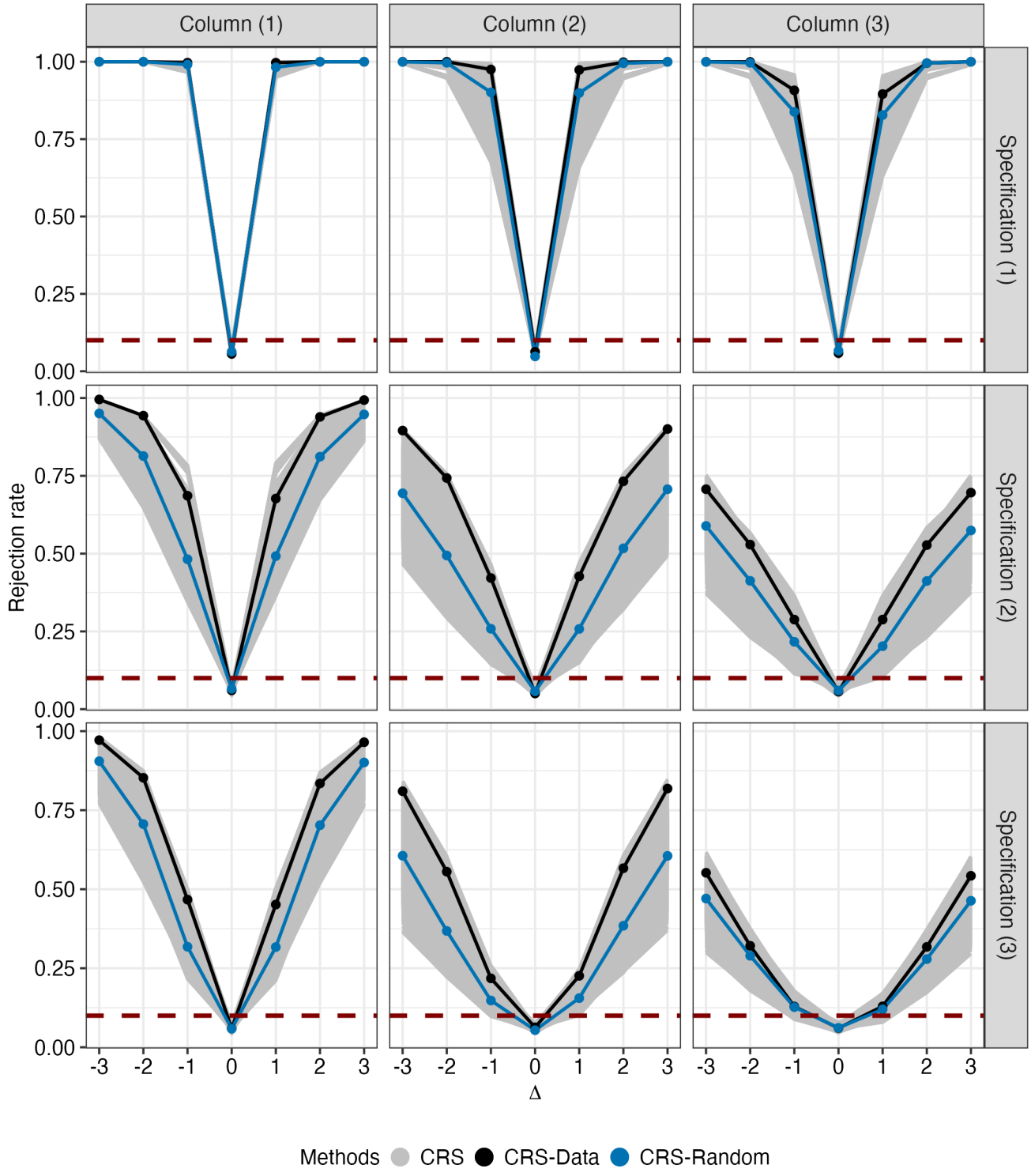
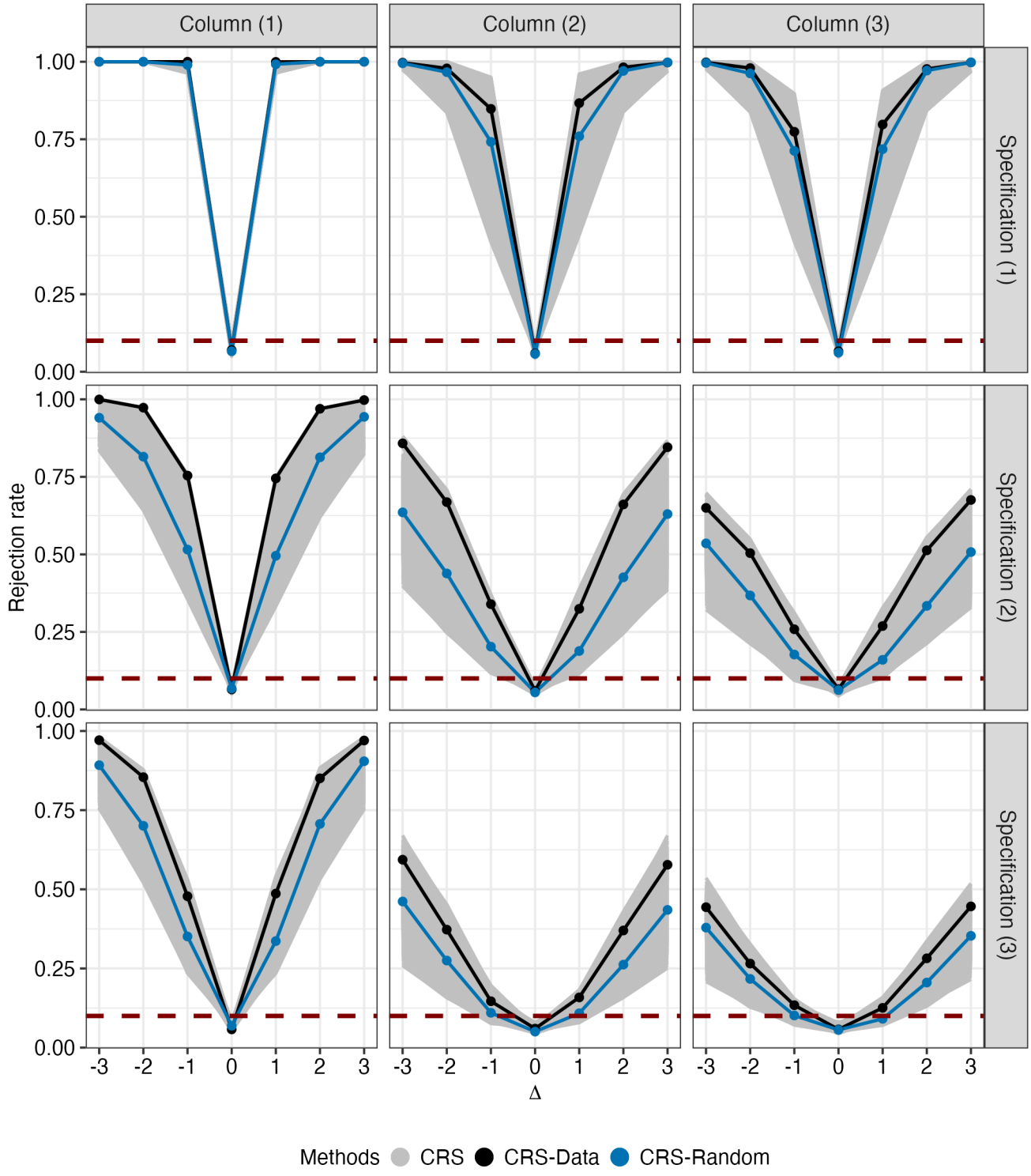


Figure 5: Rejection rate curves for inference on the coefficient for limited government $L_{t,j}$.



7 Conclusion

The approximate randomization test in Canay et al. (2017a) imposes weak assumptions on clusters and can be used to conduct valid inference when there is a small number of clusters. The test requires researchers to perform cluster-by-cluster regressions. When the target parameters cannot be identified within each cluster, researchers would have to combine clusters to perform the test.

In this paper, I first analyzed the local asymptotic power of CRS. Using the local asymptotic power as a criterion, I developed computationally efficient algorithms that can be used to guide how to combine clusters when there is an identification issue within cluster. Monte Carlo simulations and an empirical application show that the data-driven procedure performs well.

Appendix

The appendix contains all the proofs for the results in the main text and details of the calibrated simulation exercise in Section 6.

A Supplemental lemma

The lemma below is similar to Lemma 3.1, except that it considers the equality of the two terms.

Lemma A.1. *For any $g, h \in \mathbb{G}$, let $\mathcal{J}_{\text{same}}(g,h)$ and $\mathcal{J}_{\text{diff}}(g,h)$ be as defined in Lemma 3.1. Then,*

$$\{T_n(h) = T_n(g)\} = \left\{ \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} = 0 \right\} \bigcup \left\{ \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} \neq 0, \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} = 0 \right\}.$$

Proof of Lemma A.1. To begin with, write $\widetilde{T}_n(g) \equiv \sum_{j=1}^{q} g_j \widehat{S}_{n,j}$ so that $T_n(g) = |\widetilde{T}_n(g)|$ for any $g \in \mathbb{G}$. Note that the event $\{T_n(h) = T_n(g)\}$ can be written as the following nine unions of events depending on the signs of $\widetilde{T}_n(h)$ and $\widetilde{T}_n(g)$:

$$\begin{aligned}
\{T_n(h) = T_n(g)\} &= \{|\widetilde{T}_n(h)| = |\widetilde{T}_n(g)|\} \\
&= \{\widetilde{T}_n(h) = \widetilde{T}_n(g), \widetilde{T}_n(h) > 0, \widetilde{T}_n(g) > 0\} && \text{(A.1)} \\
&\cup \{\widetilde{T}_n(h) = \widetilde{T}_n(g), \widetilde{T}_n(h) = 0, \widetilde{T}_n(g) > 0\} && \text{(A.2)} \\
&\cup \{-\widetilde{T}_n(h) = \widetilde{T}_n(g), \widetilde{T}_n(h) < 0, \widetilde{T}_n(g) > 0\} && \text{(A.3)} \\
&\cup \{\widetilde{T}_n(h) = \widetilde{T}_n(g), \widetilde{T}_n(h) > 0, \widetilde{T}_n(g) = 0\} && \text{(A.4)} \\
&\cup \{\widetilde{T}_n(h) = \widetilde{T}_n(g), \widetilde{T}_n(h) = 0, \widetilde{T}_n(g) = 0\} && \text{(A.5)} \\
&\cup \{-\widetilde{T}_n(h) = \widetilde{T}_n(g), \widetilde{T}_n(h) < 0, \widetilde{T}_n(g) = 0\} && \text{(A.6)} \\
&\cup \{\widetilde{T}_n(h) = -\widetilde{T}_n(g), \widetilde{T}_n(h) > 0, \widetilde{T}_n(g) < 0\} && \text{(A.7)} \\
&\cup \{\widetilde{T}_n(h) = -\widetilde{T}_n(g), \widetilde{T}_n(h) = 0, \widetilde{T}_n(g) < 0\} && \text{(A.8)} \\
&\cup \{-\widetilde{T}_n(h) = -\widetilde{T}_n(g), \widetilde{T}_n(h) < 0, \widetilde{T}_n(g) < 0\}. && \text{(A.9)}
\end{aligned}$$

The events (A.2), (A.4), (A.6) and (A.8) are empty because each of these four events requires $\widetilde{T}_n(g) = 0$ and $\widetilde{T}_n(g) \neq 0$ at the same time.

Using the definitions of the sets $\mathcal{J}_{\text{same}}(g,h)$ and $\mathcal{J}_{\text{diff}}(g,h)$, $\widetilde{T}_n(h)$ can be written as

$$\widetilde{T}_n(h) = \sum_{j=1}^{q} h_j \widehat{S}_{n,j} = \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} + \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j}, \tag{A.10}$$

and $\widetilde{T}_n(g)$ can be written in terms of the sign changes in h as

$$\begin{aligned} \widetilde{T}_n(g) &= \sum_{j=1}^{q} g_j \widehat{S}_{n,j} \\ &= \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} g_j \widehat{S}_{n,j} + \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} g_j \widehat{S}_{n,j} \\ &= \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} - \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j}. \end{aligned} \tag{A.11}$$

The next step is to simplify the remaining events.

- (A.1) can be rewritten as follows:

$$\begin{aligned} &\{\widetilde{T}_n(h) = \widetilde{T}_n(g), \widetilde{T}_n(h) > 0, \widetilde{T}_n(g) > 0\} \\ &\quad = \{\widetilde{T}_n(h) = \widetilde{T}_n(g),\ \widetilde{T}_n(g) > 0\} \\ &\quad = \left\{ \sum_{j=1}^{q} h_j \widehat{S}_{n,j} = \sum_{j=1}^{q} g_j \widehat{S}_{n,j},\ \sum_{j=1}^{q} g_j \widehat{S}_{n,j} > 0 \right\} \\ &\quad = \left\{ \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} + \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} = \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} - \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j}, \right. \\ &\qquad \left. \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} - \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} > 0 \right\} \\ &\quad = \left\{ \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} = 0,\ \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} > \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} \right\} \\ &\quad = \left\{ \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} = 0,\ \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} > 0 \right\}. \end{aligned} \tag{A.12}$$

- (A.3) can be rewritten as follows:

$$\begin{aligned} &\{-\widetilde{T}_n(h) = \widetilde{T}_n(g), \widetilde{T}_n(h) < 0, \widetilde{T}_n(g) > 0\} \\ &\quad = \{-\widetilde{T}_n(h) = \widetilde{T}_n(g),\ \widetilde{T}_n(g) > 0\} \end{aligned}$$

$$= \left\{ -\sum_{j=1}^{q} h_j \widehat{S}_{n,j} = \sum_{j=1}^{q} g_j \widehat{S}_{n,j},\ \sum_{j=1}^{q} g_j \widehat{S}_{n,j} > 0 \right\}$$

$$= \left\{ -\sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} - \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} = \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} - \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j}, \right.$$
$$\left. \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} - \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} > 0 \right\}$$

$$= \left\{ \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} = 0,\ \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} > \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} \right\}$$

$$= \left\{ \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} = 0,\ \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} < 0 \right\}. \tag{A.13}$$

- (A.5) can be rewritten as follows:

$$\begin{aligned}
&\{\widetilde{T}_n(h) = \widetilde{T}_n(g), \widetilde{T}_n(h) = 0, \widetilde{T}_n(g) = 0\} \\
&\quad = \{\widetilde{T}_n(h) = 0,\ \widetilde{T}_n(g) = 0\} \\
&\quad = \left\{ \sum_{j=1}^{q} h_j \widehat{S}_{n,j} = 0,\ \sum_{j=1}^{q} g_j \widehat{S}_{n,j} = 0 \right\} \\
&\quad = \left\{ \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} + \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} = 0, \right. \\
&\qquad \left. \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} - \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} = 0 \right\} \\
&\quad = \left\{ \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} = 0,\ \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} = 0 \right\}.
\end{aligned} \tag{A.14}$$

- (A.7) can be rewritten as follows:

$$\begin{aligned}
&\{\widetilde{T}_n(h) = -\widetilde{T}_n(g), \widetilde{T}_n(h) > 0, \widetilde{T}_n(g) < 0\} \\
&\quad = \{\widetilde{T}_n(h) = -\widetilde{T}_n(g),\ \widetilde{T}_n(g) < 0\} \\
&\quad = \left\{ \sum_{j=1}^{q} h_j \widehat{S}_{n,j} = -\sum_{j=1}^{q} g_j \widehat{S}_{n,j},\ \sum_{j=1}^{q} g_j \widehat{S}_{n,j} < 0 \right\} \\
&\quad = \left\{ \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} + \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} = -\sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} + \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j}, \right.
\end{aligned}$$

$$\left. \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} - \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} < 0 \right\}$$
$$= \left\{ \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} = 0, \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} < \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} \right\}$$
$$= \left\{ \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} = 0, \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} > 0 \right\}. \tag{A.15}$$

- (A.9) can be rewritten as follows:

$$\begin{aligned}
&\{-\widetilde{T}_n(h) = -\widetilde{T}_n(g), \widetilde{T}_n(h) < 0, \widetilde{T}_n(g) < 0\} \\
&\quad = \{-\widetilde{T}_n(h) = -\widetilde{T}_n(g),\ \widetilde{T}_n(g) < 0\} \\
&\quad = \left\{ -\sum_{j=1}^{q} h_j \widehat{S}_{n,j} = -\sum_{j=1}^{q} g_j \widehat{S}_{n,j},\ \sum_{j=1}^{q} g_j \widehat{S}_{n,j} < 0 \right\} \\
&\quad = \left\{ -\sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} - \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} = -\sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} + \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j}, \right. \\
&\qquad \left. \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} - \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} < 0 \right\} \\
&\quad = \left\{ \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} = 0, \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} < \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} \right\} \\
&\quad = \left\{ \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} = 0, \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} < 0 \right\}.
\end{aligned} \tag{A.16}$$

Let $A \equiv \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j}$ and $B \equiv \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j}$. Then, the union of events (A.12) to (A.16) become

$$\begin{aligned}
&\{A > 0, B = 0\} \cup \{A = 0, B < 0\} \cup \{A = 0, B = 0\} \cup \{A = 0, B > 0\} \cup \{A < 0, B = 0\} \\
&\quad = (\{A = 0, B < 0\} \cup \{A = 0, B = 0\} \cup \{A = 0, B > 0\}) \\
&\qquad \cup (\{B = 0, A > 0\} \cup \{B = 0, A < 0\}) \\
&\quad = \{A = 0\} \cup \{A \neq 0, B = 0\},
\end{aligned}$$

which yields the desired result. □

The following lemma describes the limiting distribution under the local alternative. This result is used in multiple proofs so it is stated as a lemma below.

Lemma A.2. *Let Assumption 2.2 hold and let $\delta \in \mathbb{R}$ be the local alternative parameter as in (6). Then,*

$$\begin{pmatrix} \widehat{S}_{n,1} \\ \vdots \\ \widehat{S}_{n,q} \end{pmatrix} \xrightarrow{d} \begin{pmatrix} Z_1 \\ \vdots \\ Z_q \end{pmatrix} + \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_q \end{pmatrix} \delta,$$

where $Z_j \equiv c'S_j \sim N(0, \sigma_j^2)$ and $\sigma_j^2 \equiv c'\Sigma_j c$ for all $j \in \mathcal{J}$. In addition, $Z_j \perp\!\!\!\perp Z_k$ for any $j \neq k$.

Proof of Lemma A.2. Under the local alternative that $c'\beta = \lambda + \frac{\delta}{\sqrt{n}}$, the following holds for each $j \in \mathcal{J}$:

$$\widehat{S}_{n,j} = \sqrt{n_j}(c'\widehat{\beta}_{n,j} - \lambda) = \sqrt{n_j}\left(c'\widehat{\beta}_{n,j} - c'\beta + \frac{\delta}{\sqrt{n}}\right) = \sqrt{n_j}c'(\widehat{\beta}_{n,j} - \beta) + \frac{\sqrt{n_j}\delta}{\sqrt{n}}.$$

Therefore,

$$\begin{aligned} \begin{pmatrix} \widehat{S}_{n,1} \\ \vdots \\ \widehat{S}_{n,q} \end{pmatrix} &= \begin{pmatrix} \sqrt{n_1}c'(\widehat{\beta}_{n,1} - \beta) \\ \vdots \\ \sqrt{n_q}c'(\widehat{\beta}_{n,q} - \beta) \end{pmatrix} + \begin{pmatrix} \frac{\sqrt{n_1}}{\sqrt{n}} \\ \vdots \\ \frac{\sqrt{n_q}}{\sqrt{n}} \end{pmatrix} \delta \\ &\xrightarrow{d} \begin{pmatrix} c'S_1 \\ \vdots \\ c'S_q \end{pmatrix} + \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_q \end{pmatrix} \delta \\ &= \begin{pmatrix} Z_1 \\ \vdots \\ Z_q \end{pmatrix} + \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_q \end{pmatrix} \delta, \end{aligned}$$

by Slutsky's theorem, Assumption 2.2 and by defining Z_j as in the statement of the current lemma. In addition, $Z_j \perp\!\!\!\perp Z_k$ for any $j \neq k$ follows from $S_j \perp\!\!\!\perp S_k$ for any $j \neq k$ in Assumption 2.2. □

B Proofs to the main text

Proof of Lemma 3.1. To begin with, write $\widetilde{T}_n(g) \equiv \sum_{j=1}^{q} g_j \widehat{S}_{n,j}$ so that $T_n(g) = |\widetilde{T}_n(g)|$. Then, the event $\{T_n(h) > T_n(g)\}$ can be written as the following four unions of events depending on the signs of $\widetilde{T}_n(h)$ and $\widetilde{T}_n(g)$:

$$\begin{aligned}
\{T_n(h) > T_n(g)\} &= \{|\widetilde{T}_n(h)| > |\widetilde{T}_n(g)|\} \\
&= \{\widetilde{T}_n(h) > \widetilde{T}_n(g), \widetilde{T}_n(h) \geq 0, \widetilde{T}_n(g) \geq 0\} && \text{(B.1)} \\
&\quad \cup \{\widetilde{T}_n(h) > -\widetilde{T}_n(g), \widetilde{T}_n(h) \geq 0, \widetilde{T}_n(g) < 0\} && \text{(B.2)} \\
&\quad \cup \{-\widetilde{T}_n(h) > \widetilde{T}_n(g), \widetilde{T}_n(h) < 0, \widetilde{T}_n(g) \geq 0\} && \text{(B.3)} \\
&\quad \cup \{-\widetilde{T}_n(h) > -\widetilde{T}_n(g), \widetilde{T}_n(h) < 0, \widetilde{T}_n(g) < 0\}. && \text{(B.4)}
\end{aligned}$$

The next step is to simplify each of the four events in (B.1) to (B.4). Note that each of the four events above contains the intersection of three inequalities. The second inequality is implied by the first and third inequalities in each event. In addition, $\widetilde{T}_n(h)$ and $\widetilde{T}_n(g)$ can be written as in (A.10) and (A.11).

- (B.1) can be rewritten as follows:

$$\begin{aligned}
&\{\widetilde{T}_n(h) > \widetilde{T}_n(g), \widetilde{T}_n(h) \geq 0, \widetilde{T}_n(g) \geq 0\} \\
&\quad = \{\widetilde{T}_n(h) > \widetilde{T}_n(g),\ \widetilde{T}_n(g) \geq 0\} \\
&\quad = \left\{ \sum_{j=1}^{q} h_j \widehat{S}_{n,j} > \sum_{j=1}^{q} g_j \widehat{S}_{n,j},\ \sum_{j=1}^{q} g_j \widehat{S}_{n,j} \geq 0 \right\} \\
&\quad = \left\{ \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} + \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} > \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} - \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j}, \right. \\
&\qquad \left. \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} - \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} \geq 0 \right\} \\
&\quad = \left\{ \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} > 0,\ \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} \geq \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} \right\}. && \text{(B.5)}
\end{aligned}$$

- (B.2) can be rewritten as follows:

$$\begin{aligned}
&\{\widetilde{T}_n(h) > -\widetilde{T}_n(g), \widetilde{T}_n(h) \geq 0, \widetilde{T}_n(g) < 0\} \\
&\quad = \{\widetilde{T}_n(h) > -\widetilde{T}_n(g),\ \widetilde{T}_n(g) < 0\}
\end{aligned}$$

$$= \left\{ \sum_{j=1}^{q} h_j \widehat{S}_{n,j} > -\sum_{j=1}^{q} g_j \widehat{S}_{n,j},\ \sum_{j=1}^{q} g_j \widehat{S}_{n,j} < 0 \right\}$$

$$= \left\{ \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} + \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} > -\sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} + \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j}, \right.$$
$$\left. \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} - \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} < 0 \right\}$$

$$= \left\{ \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} > 0,\ \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} < \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} \right\}. \tag{B.6}$$

- (B.3) can be rewritten as follows:

$$\{-\widetilde{T}_n(h) > \widetilde{T}_n(g), \widetilde{T}_n(h) < 0, \widetilde{T}_n(g) \geq 0\}$$
$$= \{-\widetilde{T}_n(h) > \widetilde{T}_n(g),\ \widetilde{T}_n(g) \geq 0\}$$
$$= \left\{ -\sum_{j=1}^{q} h_j \widehat{S}_{n,j} > \sum_{j=1}^{q} g_j \widehat{S}_{n,j},\ \sum_{j=1}^{q} g_j \widehat{S}_{n,j} \geq 0 \right\}$$
$$= \left\{ -\sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} - \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} > \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} - \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j}, \right.$$
$$\left. \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} - \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} \geq 0 \right\}$$
$$= \left\{ \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} < 0,\ \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} \geq \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} \right\}. \tag{B.7}$$

- (B.4) can be rewritten as follows:

$$\{-\widetilde{T}_n(h) > -\widetilde{T}_n(g), \widetilde{T}_n(h) < 0, \widetilde{T}_n(g) < 0\}$$
$$= \{-\widetilde{T}_n(h) > -\widetilde{T}_n(g),\ \widetilde{T}_n(g) < 0\}$$
$$= \left\{ -\sum_{j=1}^{q} h_j \widehat{S}_{n,j} > -\sum_{j=1}^{q} g_j \widehat{S}_{n,j},\ \sum_{j=1}^{q} g_j \widehat{S}_{n,j} < 0 \right\}$$
$$= \left\{ -\sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} - \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} > -\sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} + \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j}, \right.$$
$$\left. \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} - \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} < 0 \right\}$$

$$= \left\{ \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} < 0, \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j} < \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j} \right\}. \tag{B.8}$$

Let $A \equiv \sum_{j \in \mathcal{J}_{\text{same}}(g,h)} h_j \widehat{S}_{n,j}$ and $B \equiv \sum_{j \in \mathcal{J}_{\text{diff}}(g,h)} h_j \widehat{S}_{n,j}$. Then, the union of events (B.5) to (B.8) become

$$\begin{aligned}
&\{B > 0, A \geq B\} \cup \{A > 0, A < B\} \cup \{A < 0, A \geq B\} \cup \{B < 0, A < B\} \\
&\quad = (\{A > 0, B > 0, A \geq B\} \cup \{A > 0, B > 0, A < B\}) \\
&\qquad \cup (\{A < 0, B < 0, A \geq B\} \cup \{A < 0, B < 0, A < B\}) \\
&\quad = \{A > 0, B > 0\} \cup \{A < 0, B < 0\},
\end{aligned}$$

which yields the desired result. □

Proof of Lemma 3.2. To begin with, define

$$\mathcal{E}_n(\mathcal{H}_1) \equiv \{T_n > T_n(h_2) \text{ for all } h_2 \in \mathbb{G}_{U,-1} \backslash \mathcal{H}_1 \text{ and } T_n(h_1) > T_n \text{ for all } h_1 \in \mathcal{H}_1\},$$

for any $\mathcal{H}_1 \subseteq \mathbb{G}_{U,-1}$. By definition, the event $\mathcal{E}_n(\mathcal{H}_1)$ can be written as the intersection of $\{T_n > T_n(g)\}$ for $g \in \mathbb{G}_{U,-1} \backslash \mathcal{H}_1$ and $\{T_n(g) > T_n\}$ for $g \in \mathcal{H}_1$. Recall that $T_n = T_n(1_q)$, where $1_q \equiv (1, \ldots, 1)$ is the identity transformation. Using Lemma 3.1, these events can be rewritten as

$$\{T_n > T_n(g)\} = \mathcal{A}_n^{(1)}(g) \cup \mathcal{A}_n^{(2)}(g), \tag{B.9}$$

and

$$\{T_n(g) > T_n\} = \mathcal{B}_n^{(1)}(g) \cup \mathcal{B}_n^{(2)}(g), \tag{B.10}$$

where

$$\begin{aligned}
\mathcal{A}_n^{(1)}(g) &\equiv \left\{ \sum_{j \in \mathcal{J}_{\text{same}}(g,1_q)} \widehat{S}_{n,j} > 0, \sum_{j \in \mathcal{J}_{\text{diff}}(g,1_q)} \widehat{S}_{n,j} > 0 \right\}, \\
\mathcal{A}_n^{(2)}(g) &\equiv \left\{ \sum_{j \in \mathcal{J}_{\text{same}}(g,1_q)} \widehat{S}_{n,j} < 0, \sum_{j \in \mathcal{J}_{\text{diff}}(g,1_q)} \widehat{S}_{n,j} < 0 \right\}, \\
\mathcal{B}_n^{(1)}(g) &\equiv \left\{ \sum_{j \in \mathcal{J}_{\text{same}}(g,1_q)} \widehat{S}_{n,j} > 0, \sum_{j \in \mathcal{J}_{\text{diff}}(g,1_q)} \widehat{S}_{n,j} < 0 \right\},
\end{aligned}$$

$$\mathcal{B}_n^{(2)}(g) \equiv \left\{ \sum_{j \in \mathcal{J}_{\text{same}}(g,1_q)} \widehat{S}_{n,j} < 0, \sum_{j \in \mathcal{J}_{\text{diff}}(g,1_q)} \widehat{S}_{n,j} > 0 \right\},$$

for any $g \in \mathbb{G}_{U,-1}$.

Therefore, for any $\mathcal{H}_1 \in \mathbb{H}_{k-1}$,

$$\begin{aligned}
\mathcal{E}_n(\mathcal{H}_1) &= \left(\bigcap_{h_2 \in \mathbb{G}_{U,-1} \setminus \mathcal{H}_1} \{T_n > T_n(h_2)\} \right) \bigcap \left(\bigcap_{h_1 \in \mathcal{H}_1} \{T_n(h_1) > T_n\} \right) \\
&= \left(\bigcap_{h_2 \in \mathbb{G}_{U,-1} \setminus \mathcal{H}_1} \{\mathcal{A}_n^{(1)}(h_2) \cup \mathcal{A}_n^{(2)}(h_2)\} \right) \bigcap \left(\bigcap_{h_1 \in \mathcal{H}_1} \{\mathcal{B}_n^{(1)}(h_1) \cup \mathcal{B}_n^{(2)}(h_1)\} \right) \\
&= \left(\bigcap_{h_2 \in \mathbb{G}_{U,-1} \setminus \mathcal{H}_1} \{\mathcal{F}_n^{(1)}(h_2, \mathcal{H}_1) \cup \mathcal{F}_n^{(2)}(h_2, \mathcal{H}_1)\} \right) \bigcap \left(\bigcap_{h_1 \in \mathcal{H}_1} \{\mathcal{F}_n^{(1)}(h_1, \mathcal{H}_1) \cup \mathcal{F}_n^{(2)}(h_1, \mathcal{H}_1)\} \right) \\
&= \bigcup_{m_1=1}^{2} \bigcup_{m_2=1}^{2} \cdots \bigcup_{m_L=1}^{2} \left(\bigcap_{l=1}^{L} \mathcal{F}_n^{(m_l)}(\overline{g}_l, \mathcal{H}_1) \right) \\
&= \bigcup_{m \in \mathbb{M}} \left(\bigcap_{l=1}^{L} \mathcal{F}_n^{(m_l)}(\overline{g}_l, \mathcal{H}_1) \right), \qquad \text{(B.11)}
\end{aligned}$$

where the first equality follows from the definition of $\mathcal{E}_n(\mathcal{H}_1)$, the second equality follows from (B.9) and (B.10), the third equality follows from relabelling

$$\mathcal{F}_n^{(m_l)}(\overline{g}_l, \mathcal{H}) \equiv \begin{cases} \mathcal{A}_n^{(m_l)}(\overline{g}_l) & , \ \overline{g}_l \in \mathbb{G}_{U,-1} \setminus \mathcal{H}_1 \\ \mathcal{B}_n^{(m_l)}(\overline{g}_l) & , \ \overline{g}_l \in \mathcal{H}_1 \end{cases}$$

for each $l = 1, \ldots, L$, the fourth equality uses the property on the intersection and union of a finite number of sets (note that L is finite), and the last equality follows from labeling the unique elements in $\mathbb{G}_{U,-1}$ as $\{\overline{g}_1, \ldots, \overline{g}_L\}$ as in the last paragraph of Section 3.1 and writing $\mathbb{M} \equiv \{1,2\}^L$ as in the statement of the proposition to simplify the L unions.

Next, for any $m, \widetilde{m} \in \mathbb{M}$ with $m \neq \widetilde{m}$, there must exist at least one $v = 1, \ldots, L$ such that $m_v \neq \widetilde{m}_v$. Assume without loss of generality that $m_v = 1$ and $\widetilde{m}_v = 2$. Then, it must be the case that $\mathcal{F}_n^{(m_v)}(\overline{g}_v, \mathcal{H}_1) \cap \mathcal{F}_n^{(\widetilde{m}_v)}(\overline{g}_v, \mathcal{H}_1) = \emptyset$ because $\mathcal{A}_n^{(1)}(\overline{g}_v) \cap \mathcal{A}_n^{(2)}(\overline{g}_v) = \emptyset$ if $\overline{g}_v \in \mathbb{G}_{U,-1} \setminus \mathcal{H}_1$ and $\mathcal{B}_n^{(1)}(\overline{g}_v) \cap \mathcal{B}_n^{(2)}(\overline{g}_v) = \emptyset$ if $\overline{g}_v \in \mathcal{H}_1$. This follows because for a given $g \in \mathbb{G}_{U,-1}$, $\mathcal{A}_n^{(1)}(g)$ requires

$$\sum_{j \in \mathcal{J}} \widehat{S}_{n,j} > 0,$$

whereas $\mathcal{A}_n^{(2)}(g)$ requires

$$\sum_{j\in\mathcal{J}} \widehat{S}_{n,j} < 0.$$

Similarly, for a given $g \in \mathbb{G}_{U,-1}$, $\mathcal{B}_n^{(1)}(g)$ requires

$$\sum_{j\in\mathcal{J}_{\text{same}}(g,1_q)} \widehat{S}_{n,j} - \sum_{j\in\mathcal{J}_{\text{diff}}(g,1_q)} \widehat{S}_{n,j} > 0,$$

whereas $\mathcal{B}_n^{(2)}(g)$ requires

$$\sum_{j\in\mathcal{J}_{\text{same}}(g,1_q)} \widehat{S}_{n,j} - \sum_{j\in\mathcal{J}_{\text{diff}}(g,1_q)} \widehat{S}_{n,j} < 0.$$

As a result,

$$\left(\bigcap_{l=1}^{L} \mathcal{F}_n^{(m_l)}(\overline{g}_l, \mathcal{H}_1)\right) \bigcap \left(\bigcap_{l=1}^{L} \mathcal{F}_n^{(\widetilde{m}_l)}(\overline{g}_l, \mathcal{H}_1)\right) = \emptyset, \tag{B.12}$$

whenever $m \neq \widetilde{m}$.

Using (B.11), (B.12), the inclusion-exclusion principle, and the fact that $|\mathbb{M}|$ is finite, it follows that

$$\lim_{n\to\infty} \mathbb{P}_\delta[\mathcal{E}_n(\mathcal{H}_1)] = \sum_{m\in\mathbb{M}} \lim_{n\to\infty} \mathbb{P}_\delta\left[\left(\bigcap_{l=1}^{L} \mathcal{F}_n^{(m_l)}(\overline{g}_l, \mathcal{H}_1)\right)\right]. \tag{B.13}$$

For each $l = 1, \ldots, L$, the event $\mathcal{F}_n^{(m_l)}(\overline{g}_l, \mathcal{H}_1)$ can be written as $\{F_l \widehat{S}_n > 0_{2\times 1}\}$, where $0_{2\times 1} \equiv (0,0)'$, $\widehat{S}_n \equiv (\widehat{S}_{n,1}, \ldots, \widehat{S}_{n,q})'$ and F_l is a $2 \times q$ matrix defined as follows.

- If $\overline{g}_l \in \mathbb{G}_{U,-1} \backslash \mathcal{H}_1$ and $m_l = 1$, then the (i,j)-entry of F_l is defined as

$$\begin{cases} 1 & \text{, if } (i = 1 \text{ and } j \in \mathcal{J}_{\text{same}}(\overline{g}_l, 1_q)) \text{ or } (i = 2 \text{ and } j \in \mathcal{J}_{\text{diff}}(\overline{g}_l, 1_q)) \\ 0 & \text{, otherwise} \end{cases}. \tag{B.14}$$

- If $\overline{g}_l \in \mathbb{G}_{U,-1} \backslash \mathcal{H}_1$ and $m_l = 2$, then the (i,j)-entry of F_l is defined as

$$\begin{cases} -1 & \text{, if } (i = 1 \text{ and } j \in \mathcal{J}_{\text{same}}(\overline{g}_l, 1_q)) \text{ or } (i = 2 \text{ and } j \in \mathcal{J}_{\text{diff}}(\overline{g}_l, 1_q)) \\ 0 & \text{, otherwise} \end{cases}. \tag{B.15}$$

- If $\overline{g}_l \in \mathcal{H}_1$ and $m_l = 1$, then the (i,j)-entry of F_l is defined as

$$\begin{cases} 1 & \text{, if } i = 1 \text{ and } j \in \mathcal{J}_{\text{same}}(\overline{g}_l, 1_q) \\ -1 & \text{, if } i = 2 \text{ and } j \in \mathcal{J}_{\text{diff}}(\overline{g}_l, 1_q) \\ 0 & \text{, otherwise} \end{cases}. \tag{B.16}$$

- If $\overline{g}_l \in \mathcal{H}_1$ and $m_l = 2$, then the (i,j)-entry of F_l is defined as

$$\begin{cases} -1 & \text{, if } i = 1 \text{ and } j \in \mathcal{J}_{\text{same}}(\overline{g}_l, 1_q) \\ 1 & \text{, if } i = 2 \text{ and } j \in \mathcal{J}_{\text{diff}}(\overline{g}_l, 1_q) \\ 0 & \text{, otherwise} \end{cases}. \tag{B.17}$$

Let F be the matrix that stacks $F_1, \ldots, F_L$ by row. Using Lemma A.2 and the continuous mapping theorem, it follows that

$$F\widehat{S}_n \xrightarrow{d} FZ + F\xi\delta, \tag{B.18}$$

where $Z \equiv (Z_1, \ldots, Z_q)'$ and $\xi \equiv (\xi_1, \ldots, \xi_q)'$ such that $Z_j \equiv c'S_j \sim N(0, \sigma_j^2)$ and $\sigma_j^2 \equiv c'\Sigma_j c$ for all $j \in \mathcal{J}$ as defined in Lemma A.2. In addition, $Z_j \perp\!\!\!\perp Z_k$ for any $j \neq k$. Let

$$\mathcal{F}^{(m_l)}(\overline{g}_l, \mathcal{H}_1, \delta) \equiv \begin{cases} \mathcal{A}^{(m_l)}(\overline{g}_l, \delta) & \text{, } \overline{g}_l \in \mathbb{G}_{U,-1} \backslash \mathcal{H}_1 \\ \mathcal{B}^{(m_l)}(\overline{g}_l, \delta) & \text{, } \overline{g}_l \in \mathcal{H}_1 \end{cases},$$

for any $l = 1, \ldots, L$, and

$$\mathcal{A}^{(1)}(g,\delta) \equiv \left\{ \sum_{j \in \mathcal{J}_{\text{same}}(g,1_q)} Z_j > -\sum_{j \in \mathcal{J}_{\text{same}}(g,1_q)} \xi_j\delta, \sum_{j \in \mathcal{J}_{\text{diff}}(g,1_q)} Z_j > -\sum_{j \in \mathcal{J}_{\text{diff}}(g,1_q)} \xi_j\delta \right\},$$

$$\mathcal{A}^{(2)}(g,\delta) \equiv \left\{ \sum_{j \in \mathcal{J}_{\text{same}}(g,1_q)} Z_j < -\sum_{j \in \mathcal{J}_{\text{same}}(g,1_q)} \xi_j\delta, \sum_{j \in \mathcal{J}_{\text{diff}}(g,1_q)} Z_j < -\sum_{j \in \mathcal{J}_{\text{diff}}(g,1_q)} \xi_j\delta \right\},$$

$$\mathcal{B}^{(1)}(g,\delta) \equiv \left\{ \sum_{j \in \mathcal{J}_{\text{same}}(g,1_q)} Z_j > -\sum_{j \in \mathcal{J}_{\text{same}}(g,1_q)} \xi_j\delta, \sum_{j \in \mathcal{J}_{\text{diff}}(g,1_q)} Z_j < -\sum_{j \in \mathcal{J}_{\text{diff}}(g,1_q)} \xi_j\delta \right\},$$

$$\mathcal{B}^{(2)}(g,\delta) \equiv \left\{ \sum_{j \in \mathcal{J}_{\text{same}}(g,1_q)} Z_j < -\sum_{j \in \mathcal{J}_{\text{same}}(g,1_q)} \xi_j\delta, \sum_{j \in \mathcal{J}_{\text{diff}}(g,1_q)} Z_j > -\sum_{j \in \mathcal{J}_{\text{diff}}(g,1_q)} \xi_j\delta \right\},$$

for any $g \in \mathbb{G}_{U,-1}$.

Let $\mathfrak{B} \equiv \{(x_1, \ldots, x_{2L}) : x_l > 0 \text{ for each } l = 1, \ldots, 2L\}$, so that $\{F\widehat{S}_n > 0\}$ can be written as $\{F\widehat{S}_n \in \mathfrak{B}\}$. Denote $\partial\mathfrak{B}$ as the boundary of $\mathfrak{B}$. Note that $\partial\mathfrak{B} \subseteq \bigcup_{l=1}^{2L}\{x_l = 0\}$. Now, $\mathbb{P}[\sum_{j\in\mathcal{J}_{\text{same}}(\overline{g}_l,1_q)} Z_j = -\sum_{j\in\mathcal{J}_{\text{same}}(\overline{g}_l,1_q)} \xi_j\delta] = 0$ for each $l = 1, \ldots, L$ by Assumption 3.3. This follows by writing $w_0 = \sum_{j\in\mathcal{J}_{\text{same}}(g,1_q)} \xi_j\delta$, $w_j = 1$ for $j \in \mathcal{J}_{\text{same}}(\overline{g}_l, 1_q)$ and $w_j = 0$ for $j \notin \mathcal{J}_{\text{same}}(\overline{g}_l, 1_q)$ in Assumption 3.3. For the same reason, $\mathbb{P}[\sum_{j\in\mathcal{J}_{\text{diff}}(\overline{g}_l,1_q)} Z_j = -\sum_{j\in\mathcal{J}_{\text{diff}}(\overline{g}_l,1_q)} \xi_j\delta] = 0$ for each $l = 1, \ldots, L$. Therefore, $\mathbb{P}[FZ + F\xi\delta \in \partial\mathcal{B}] = 0$ by applying the union bound. Hence, using the Portmanteau lemma (see, e.g., Chapter 2 of van der Vaart (2000)),

$$\lim_{n\to\infty} \mathbb{P}_\delta\left[\left(\bigcap_{l=1}^{L} \mathcal{F}_n^{(m_l)}(\overline{g}_l, \mathcal{H}_1)\right)\right] = \lim_{n\to\infty} \mathbb{P}_\delta[F\widehat{S}_n \in \mathfrak{B}] = \mathbb{P}\left[\left(\bigcap_{l=1}^{L} \mathcal{F}^{(m_l)}(\overline{g}_l, \mathcal{H}_1, \delta)\right)\right].$$

The proof is complete by defining $V_{\text{same}}(g, \delta)$ and $V_{\text{diff}}(g, \delta)$ as in the statement of the lemma. □

Proof of Theorem 3.4. To begin with, define $R_n \equiv \mathbb{1}[T_n \neq T_n(g) \text{ for any } g \in \mathbb{G}_{U,-1}]$. Following the definition in (7), write the local asymptotic power as the following two parts

$$\pi(\delta, \alpha) = \lim_{n\to\infty} \mathbb{P}_\delta[T_n > \widehat{\text{cv}}_n(1-\alpha)] = P_1 + P_2,$$

where

$$P_1 \equiv \lim_{n\to\infty} \mathbb{P}_\delta[T_n > \widehat{\text{cv}}_n(1-\alpha), R_n = 1], \tag{B.19}$$

$$P_2 \equiv \lim_{n\to\infty} \mathbb{P}_\delta[T_n > \widehat{\text{cv}}_n(1-\alpha), R_n = 0]. \tag{B.20}$$

Here, P_1 does not allow for "ties" in $\{T_n > \widehat{\text{cv}}_n(1-\alpha)\}$ whereas P_2 allows for "ties." The goal is to derive an expression for P_1 and show that $P_2 = 0$.

Part 1: Computation of P_1

Recall that $\{T_n > \widehat{\text{cv}}_n(1-\alpha)\}$ is the same as the event that T_n is one of the largest $K \equiv \lfloor \alpha|\mathbb{G}_U| \rfloor$ terms in $\{T_n(g) : g \in \mathbb{G}_U\}$. Let $\mathbb{H}_k$ be the collection of all distinct size k subsets of $\mathbb{G}_{U,-1}$ as in the statement of this proposition. Hence, the probability in (B.19) can be written as

$$\mathbb{P}_\delta[T_n > \widehat{\text{cv}}_n(1-\alpha), R_n = 1] = \sum_{k=1}^{K} P_{n,1,k}, \tag{B.21}$$

where $P_{n,1,k}$ is the probability that T_n is the kth largest term in $\{T_n(g) : g \in \mathbb{G}_U\}$, and is

defined as

$$P_{n,1,k} \equiv \mathbb{P}_\delta \left[\bigcup_{\mathcal{H} \in \mathbb{H}_{k-1}} \mathcal{E}_n(\mathcal{H}) \right], \tag{B.22}$$

where

$$\mathcal{E}_n(\mathcal{H}) \equiv \{T_n > T_n(h_1) \text{ for all } h_1 \in \mathbb{G}_{U,-1} \backslash \mathcal{H} \text{ and } T_n(h_2) > T_n \text{ for all } h_2 \in \mathcal{H}\},$$

with the convention that $\mathcal{E}_n(\mathcal{H}) = \{T_n > T_n(h_1) \text{ for all } h_1 \in \mathbb{G}_{U,-1}\}$ when $k = 1$. The strict inequalities in $\mathcal{E}_n(\mathcal{H})$ follow from (B.19) that $T_n \neq T_n(g)$ for any $g \in \mathbb{G}_{U,-1}$ and the definition of $\widehat{\text{cv}}_n(1-\alpha)$.

For any $\mathcal{H}, \widetilde{\mathcal{H}} \in \mathbb{H}_{k-1}$ with $\mathcal{H} \neq \widetilde{\mathcal{H}}$, it must be that

$$\mathcal{E}_n(\mathcal{H}) \cap \mathcal{E}_n(\widetilde{\mathcal{H}}) = \emptyset. \tag{B.23}$$

This is because there exists at least one $\widetilde{h} \in \widetilde{\mathcal{H}}$ such that $T_n(\widetilde{h}) > T_n$ in the event $\mathcal{E}_n(\widetilde{\mathcal{H}})$ but $T_n > T_n(\widetilde{h})$ in the event $\mathcal{E}_n(\mathcal{H})$. Otherwise, $\mathcal{H} = \widetilde{\mathcal{H}}$. Since K and $|\mathbb{H}_{k-1}|$ are finite, it follows that

$$P_1 = \sum_{k=1}^{K} \sum_{\mathcal{H} \in \mathbb{H}_{k-1}} \lim_{n \to \infty} \mathbb{P}_\delta[\mathcal{E}_n(\mathcal{H})], \tag{B.24}$$

by the inclusion-exclusion principle and (B.23).

Using Lemma 3.2 and (B.24), it follows that

$$P_1 = \sum_{k=1}^{K} \sum_{\mathcal{H} \in \mathbb{H}_{k-1}} \sum_{m \in \mathbb{M}} \mathbb{P}\left[\left(\bigcap_{l=1}^{L} \mathcal{F}^{(m_l)}(\overline{g}_l, \mathcal{H}) \right) \right],$$

using the same notations from Lemma 3.2.

Part 2: Computation of P_2

Consider the case where there can be ties of T_n with some $T_n(g)$ where $g \in \mathbb{G}_{U,-1}$. Recall that k refers to T_n being the kth largest term. It is not possible to have such ties when $k = 1$ because this case requires $T_n > T_n(g)$ for all $g \in \mathbb{G}_{U,-1}$. Thus, $P_2 = 0$ if $k = 1$.

Now, consider $k \geq 2$. The goal is also to show that the limiting probability in this case is 0. For each $k = 2, \ldots, K$, let $\mathbb{O}(\mathcal{H})$ be the set that contains all subsets of $\mathcal{H}$ with size $1, \ldots, |\mathcal{H}|$ (i.e., it is the power set of $\mathcal{H}$ excluding the empty set). By similar reasoning as

in the last part and using (B.20), P_2 can be written as

$$P_2 = \lim_{n\to\infty} \sum_{k=2}^{K} P_{n,2,k}, \tag{B.25}$$

where

$$P_{n,2,k} \equiv \sum_{\mathcal{H}\in\mathbb{H}_{k-1}} \sum_{\mathcal{O}\in\mathbb{O}(\mathcal{H})} \lim_{n\to\infty} \mathbb{P}_\delta[\widetilde{\mathcal{E}}_n(\mathcal{H},\mathcal{O})], \tag{B.26}$$

and

$$\begin{aligned}\widetilde{\mathcal{E}}_n(\mathcal{H},\mathcal{O}) \equiv \{&T_n > T_n(h_1) \text{ for all } h_1 \in \mathbb{G}_{U,-1}\backslash\mathcal{H},\\ &T_n(h_2) > T_n \text{ for all } h_2 \in \mathcal{H}\backslash\mathcal{O},\\ &T_n(h_3) = T_n \text{ for all } h_3 \in \mathcal{O}\},\end{aligned}$$

noting that $\mathcal{O}$ is a subset of $\mathcal{H}$ by construction.

The interpretation of $P_{n,2,k}$ in (B.26) is similar to before. It collects the probability of all events such that T_n is the kth largest term while allowing for some ties as represented by the set $\mathcal{O}$. Equation (B.26) follows because $|\mathbb{H}_{k-1}|$ and $|\mathbb{O}(\mathcal{H})|$ are finite, and that $\mathcal{E}_n(\mathcal{H},\mathcal{O}) \cap \mathcal{E}_n(\widetilde{\mathcal{H}},\widetilde{\mathcal{O}}) = \emptyset$ unless $\mathcal{H} = \widetilde{\mathcal{H}}$ and $\mathcal{O} = \widetilde{\mathcal{O}}$.

Now, applying Lemma A.1 with $h = 1_q$ and using the same derivation as for (B.11) gives

$$\widetilde{\mathcal{E}}_n(\mathcal{H},\mathcal{O}) = \bigcup_{m\in\mathbb{M}} \left(\bigcap_{l=1}^{L} \widetilde{\mathcal{F}}_n^{(m_l)}(\overline{g}_l,\mathcal{H},\mathcal{O}) \right), \tag{B.27}$$

where

$$\widetilde{\mathcal{F}}_n^{(m_l)}(\overline{g}_l,\mathcal{H},\mathcal{O}) \equiv \begin{cases} \mathcal{A}_n^{(m_l)}(\overline{g}_l) & , \overline{g}_l \in \mathbb{G}_{U,-1}\backslash\mathcal{H} \\ \mathcal{B}_n^{(m_l)}(\overline{g}_l) & , \overline{g}_l \in \mathcal{H}\backslash\mathcal{O} \\ \mathcal{C}_n^{(m_l)}(\overline{g}_l) & , \overline{g}_l \in \mathcal{O} \end{cases},$$

$$\mathcal{C}_n^{(1)}(g) \equiv \left\{ \sum_{j\in\mathcal{J}_{\text{same}}(g,1_q)} \widehat{S}_{n,j} = 0 \right\},$$

$$\mathcal{C}_n^{(2)}(g) \equiv \left\{ \sum_{j\in\mathcal{J}_{\text{same}}(g,1_q)} \widehat{S}_{n,j} \neq 0, \sum_{j\in\mathcal{J}_{\text{diff}}(g,1_q)} \widehat{S}_{n,j} = 0 \right\},$$

$\mathcal{A}_n^{(m_l)}(\overline{g}_l)$ and $\mathcal{B}_n^{(m_l)}(\overline{g}_l)$ are as defined in the proof of Lemma 3.2.

Similar to the proof of (B.12), consider any $m, \widetilde{m} \in \mathbb{M}$ with $m \neq \widetilde{m}$, there must exist at least one $v = 1, \ldots, L$ such that $m_v \neq \widetilde{m}_v$. Assume without loss of generality that $m_v = 1$ and $\widetilde{m}_v = 2$. Then, it must be the case that $\widetilde{\mathcal{F}}_n^{(m_v)}(\overline{g}_v, \mathcal{H}, \mathcal{O}) \cap \widetilde{\mathcal{F}}_n^{(\widetilde{m}_v)}(\overline{g}_v, \mathcal{H}, \mathcal{O}) = \emptyset$. This follows because $\mathcal{A}_n^{(1)}(\overline{g}_v) \cap \mathcal{A}_n^{(2)}(\overline{g}_v) = \emptyset$ if $\overline{g}_v \in \mathbb{G}_{U,-1} \backslash \mathcal{H}$, $\mathcal{B}_n^{(1)}(\overline{g}_v) \cap \mathcal{B}_n^{(2)}(\overline{g}_v) = \emptyset$ if $\overline{g}_v \in \mathcal{H} \backslash \mathcal{O}$, and $\mathcal{C}_n^{(1)}(\overline{g}_v) \cap \mathcal{C}_n^{(2)}(\overline{g}_v) = \emptyset$ if $\overline{g}_v \in \mathcal{O}$. As a result,

$$\left(\bigcap_{l=1}^{L} \widetilde{\mathcal{F}}_n^{(m_l)}(\overline{g}_l, \mathcal{H}, \mathcal{O}) \right) \bigcap \left(\bigcap_{l=1}^{L} \widetilde{\mathcal{F}}_n^{(\widetilde{m}_l)}(\overline{g}_l, \mathcal{H}, \mathcal{O}) \right) = \emptyset, \tag{B.28}$$

whenever $m \neq \widetilde{m}$. Using (B.26), (B.28), (B.29), the inclusion-exclusion principle, and the fact that $|\mathbb{M}|$ is finite, it follows that

$$\lim_{n \to \infty} \mathbb{P}_\delta[\widetilde{\mathcal{E}}_n(\mathcal{H}, \mathcal{O})] = \sum_{m \in \mathbb{M}} \lim_{n \to \infty} \mathbb{P}_\delta \left[\left(\bigcap_{l=1}^{L} \widetilde{\mathcal{F}}_n^{(m_l)}(\overline{g}_l, \mathcal{H}, \mathcal{O}) \right) \right] \tag{B.29}$$

The next step is similar to the discussion in the proof of Lemma 3.2. Fix a $\mathcal{H} \in \mathbb{H}_{k-1}$ and $\mathcal{O} \in \mathbb{O}(\mathcal{H})$. Note that for each $l = 1, \ldots, L$, the event $\widetilde{\mathcal{F}}_n^{(m_l)}(\overline{g}_l, \mathcal{H}, \mathcal{O})$ can be written as $\{\widetilde{F}_l \widehat{S}_n > 0_{2\times 1}\}$ if $\overline{g}_l \notin \mathcal{O}$, where $0_{2\times 1} \equiv (0,0)'$, $\widehat{S}_n \equiv (\widehat{S}_{n,1}, \ldots, \widehat{S}_{n,q})'$ and $\widetilde{F}_l$ is a $2 \times q$ matrix defined as follows, again by recalling that $\mathcal{O}$ is a subset of $\mathcal{H}$ by construction.

- If $\overline{g}_l \in \mathbb{G}_{U,-1} \backslash \mathcal{H}$ and $m_l = 1$, then the (i,j)-entry of $\widetilde{F}_l$ is defined as in (B.14).
- If $\overline{g}_l \in \mathbb{G}_{U,-1} \backslash \mathcal{H}$ and $m_l = 2$, then the (i,j)-entry of $\widetilde{F}_l$ is defined as in (B.15).
- If $\overline{g}_l \in \mathcal{H} \backslash \mathcal{O}$ and $m_l = 1$, then the (i,j)-entry of $\widetilde{F}_l$ is defined as in (B.16).
- If $\overline{g}_l \in \mathcal{H} \backslash \mathcal{O}$ and $m_l = 2$, then the (i,j)-entry of $\widetilde{F}_l$ is defined as in (B.17).

If $\overline{g}_l \in \mathcal{O}$, then the event $\widetilde{\mathcal{F}}_n^{(m_l)}(\overline{g}_l, \mathcal{H}, \mathcal{O})$ can be written as follows.

- If $\overline{g}_l \in \mathcal{O}$ and $m_l = 1$, the event $\widetilde{\mathcal{F}}_n^{(m_l)}(\overline{g}_l, \mathcal{H}, \mathcal{O})$ can be written as $\{\widetilde{F}_l' \widehat{S}_n = 0\}$, where $\widetilde{F}_l$ is a vector of q components. The jth-entry of $\widetilde{F}_l$ is defined as

$$\begin{cases} 1 & \text{, if } j \in \mathcal{J}_{\text{same}}(\overline{g}_l, 1_q) \\ 0 & \text{, otherwise} \end{cases}.$$

- If $\overline{g}_l \in \mathcal{O}$ and $m_l = 2$, the event $\widetilde{\mathcal{F}}_n^{(m_l)}(\overline{g}_l, \mathcal{H}, \mathcal{O})$ can be written as $\{\widetilde{F}_{l,1}' \widehat{S}_n \neq$

$0, \widetilde{F}'_{l,2}\widehat{S}_n = 0\}$, where $\widetilde{F}_{l,i}$ is the ith row of $\widetilde{F}_l$. The (i,j)-entry of $\widetilde{F}_l$ is defined as

$$\begin{cases} 1 & \text{, if } (i = 1 \text{ and } j \in \mathcal{J}_{\text{same}}(\overline{g}_l, 1_q)) \text{ or } (i = 2 \text{ and } j \in \mathcal{J}_{\text{diff}}(\overline{g}_l, 1_q)) \\ 0 & \text{, otherwise} \end{cases}.$$

Let $\widetilde{F}$ be the matrix that stacks $\widetilde{F}_1, \ldots, \widetilde{F}_L$ by row. Using Lemma A.2 and the continuous mapping theorem,

$$\widetilde{F}\widehat{S}_n \xrightarrow{d} \widetilde{F}Z + \widetilde{F}\xi\delta,$$

where $Z \equiv (Z_1, \ldots, Z_q)'$ and $\xi \equiv (\xi_1, \ldots, \xi_q)'$ such that $Z_j \equiv c'S_j \sim N(0, \sigma_j^2)$ and $\sigma_j^2 \equiv c'\Sigma_j c$ for all $j \in \mathcal{J}$ as before. In addition, $Z_j \perp\!\!\!\perp Z_k$ for any $j \neq k$.

Recall that for each $l = 1, \ldots, L$, $\mathbb{P}[\sum_{j \in \mathcal{J}_{\text{same}}(\overline{g}_l, 1_q)} Z_j = -\sum_{j \in \mathcal{J}_{\text{same}}(\overline{g}_l, 1_q)} \xi_j \delta] = 0$ and $\mathbb{P}[\sum_{j \in \mathcal{J}_{\text{diff}}(\overline{g}_l, 1_q)} Z_j = -\sum_{j \in \mathcal{J}_{\text{diff}}(\overline{g}_l, 1_q)} \xi_j \delta] = 0$ by Assumption 3.3 and applying the arguments used in the proof of Lemma 3.2. Therefore, using the same reasoning as before and applying the Portmanteau lemma (see, e.g., Chapter 2 of van der Vaart (2000)), it follows that

$$\lim_{n \to \infty} \mathbb{P}_\delta \left[\left(\bigcap_{l=1}^{L} \widetilde{\mathcal{F}}_n^{(m_l)}(\overline{g}_l, \mathcal{H}, \mathcal{O}) \right) \right] = \mathbb{P} \left[\left(\bigcap_{l=1}^{L} \widetilde{\mathcal{F}}^{(m_l)}(\overline{g}_l, \mathcal{H}, \mathcal{O}, \delta) \right) \right], \tag{B.30}$$

where

$$\widetilde{\mathcal{F}}^{(m_l)}(\overline{g}_l, \mathcal{H}, \mathcal{O}, \delta) \equiv \begin{cases} \mathcal{A}^{(m_l)}(\overline{g}_l, \delta) & \text{, } \overline{g}_l \in \mathbb{G}_{U,-1} \backslash \mathcal{H} \\ \mathcal{B}^{(m_l)}(\overline{g}_l, \delta) & \text{, } \overline{g}_l \in \mathcal{H} \backslash \mathcal{O} \\ \mathcal{C}^{(m_l)}(\overline{g}_l, \delta) & \text{, } \overline{g}_l \in \mathcal{O} \end{cases},$$

for any $l = 1, \ldots, L$. In the above, $\mathcal{A}^{(m_l)}(\overline{g}_l, \delta)$ and $\mathcal{B}^{(m_l)}(\overline{g}_l, \delta)$ are as defined in the proof of Lemma 3.2, and

$$\mathcal{C}^{(1)}(g, \delta) \equiv \left\{ \sum_{j \in \mathcal{J}_{\text{same}}(g, 1_q)} Z_j = -\sum_{j \in \mathcal{J}_{\text{same}}(g, 1_q)} \xi_j \delta, \right\},$$

$$\mathcal{C}^{(2)}(g, \delta) \equiv \left\{ \sum_{j \in \mathcal{J}_{\text{same}}(g, 1_q)} Z_j \neq -\sum_{j \in \mathcal{J}_{\text{same}}(g, 1_q)} \xi_j \delta, \sum_{j \in \mathcal{J}_{\text{diff}}(g, 1_q)} Z_j = -\sum_{j \in \mathcal{J}_{\text{diff}}(g, 1_q)} \xi_j \delta \right\},$$

for any $g \in \mathbb{G}_{U,-1}$.

In the event $\mathcal{C}^{(1)}(\overline{g}_l, \delta)$ for each $l = 1, \ldots, L$, $\mathcal{J}_{\text{same}}(\overline{g}_l, 1_q)$ is nonempty. As before,

in terms of the notations in Assumption 3.3, this means $w_0 = \sum_{j \in \mathcal{J}_{\text{same}}(g, 1_q)} \xi_j \delta$, $w_j = 1$ for $j \in \mathcal{J}_{\text{same}}(\overline{g}_l, 1_q)$ and $w_j = 0$ for $j \notin \mathcal{J}_{\text{same}}(\overline{g}_l, 1_q)$. The event $\mathcal{C}^{(1)}(\overline{g}_l, \delta)$ occurs with probability 0 by Assumption 3.3. Similarly, the event $\mathcal{C}^{(2)}(\overline{g}_l, \delta)$ occurs with probability 0 by Assumption 3.3. It follows that the probability in (B.30) equals 0.

The above argument works with any $\mathcal{H} \in \mathbb{H}_{k-1}$ and $\mathcal{O} \in \mathbb{O}(\mathcal{H})$. This implies that $P_2 = 0$ by (B.26).

Conclusion

Combining the results in the two parts, it follows that $\pi(\delta, \alpha) = P_1$. □

Proof of Corollary 3.5. For any positive integer q, $\alpha \in [\frac{1}{2^{q-1}}, \frac{1}{2^{q-2}})$ corresponds to the case where $K = \lfloor \alpha |\mathbb{G}_U| \rfloor = 1$. Using Theorem 3.4, the rejection probability is

$$\lim_{n \to \infty} \mathbb{P}_\delta[T_n > \widehat{\text{cv}}_n(1-\alpha)] = \sum_{m \in \mathbb{M}} \lim_{n \to \infty} \mathbb{P}_\delta \left[\bigcap_{l=1}^{L} \mathcal{F}_n^{(m_l)}(\overline{g}_l) \right], \tag{B.31}$$

where

$$\mathcal{F}_n^{(1)}(g) \equiv \left\{ \sum_{j \in \mathcal{J}_{\text{same}}(g, 1_q)} \widehat{S}_{n,j} > 0, \sum_{j \in \mathcal{J}_{\text{diff}}(g, 1_q)} \widehat{S}_{n,j} > 0 \right\},$$
$$\mathcal{F}_n^{(2)}(g) \equiv \left\{ \sum_{j \in \mathcal{J}_{\text{same}}(g, 1_q)} \widehat{S}_{n,j} < 0, \sum_{j \in \mathcal{J}_{\text{diff}}(g, 1_q)} \widehat{S}_{n,j} < 0 \right\},$$

for any $g \in \mathbb{G}_{U,-1}$. Note that $\mathcal{F}_n^{(1)}(g) \cap \mathcal{F}_n^{(2)}(h) = \emptyset$ for any $g, h \in \mathbb{G}_{U,-1}$ with $g \neq h$. This is because for any $g \in \mathbb{G}_{U,-1}$, the event $\mathcal{F}_n^{(1)}(g)$ implies

$$\sum_{j \in \mathcal{J}} \widehat{S}_{n,j} = \sum_{j \in \mathcal{J}_{\text{same}}(g, 1_q)} \widehat{S}_{n,j} + \sum_{j \in \mathcal{J}_{\text{diff}}(g, 1_q)} \widehat{S}_{n,j} > 0,$$

and the event $\mathcal{F}_n^{(2)}(g)$ implies

$$\sum_{j \in \mathcal{J}} \widehat{S}_{n,j} = \sum_{j \in \mathcal{J}_{\text{same}}(g, 1_q)} \widehat{S}_{n,j} + \sum_{j \in \mathcal{J}_{\text{diff}}(g, 1_q)} \widehat{S}_{n,j} < 0.$$

If $\{m_l\}_{l=1}^L$ are not all the same, it follows that

$$\mathbb{P}_\delta \left[\bigcap_{l=1}^{L} \mathcal{F}_n^{(m_l)}(\overline{g}_l) \right] = 0.$$

Therefore,

$$\lim_{n\to\infty} \mathbb{P}_\delta[T_n > \widehat{\mathrm{cv}}_n(1-\alpha)] = \lim_{n\to\infty} \mathbb{P}_\delta\left[\bigcap_{l=1}^{L} \mathcal{F}_n^{(1)}(\overline{g}_l)\right] + \lim_{n\to\infty} \mathbb{P}_\delta\left[\bigcap_{l=1}^{L} \mathcal{F}_n^{(2)}(\overline{g}_l)\right]. \tag{B.32}$$

Note that the intersections from $l = 1, \ldots, L$ considers all the sign changes in $\mathbb{G}_{U,-1}$. By expanding the intersection,

$$\begin{aligned}
\bigcap_{l=1}^{L} \mathcal{F}_n^{(1)}(\overline{g}_l) &= \bigcap_{j_1\in\mathcal{J}} \{\widehat{S}_{n,j_1} > 0\} \bigcap_{\{j_1,j_2\}\subset\mathcal{J}} \{\widehat{S}_{n,j_1} + \widehat{S}_{n,j_2} > 0\} \\
&\quad \cdots \bigcap_{\{j_1,j_2,\ldots,j_{q-1}\}\subset\mathcal{J}} \{\widehat{S}_{n,j_1} + \widehat{S}_{n,j_2} + \cdots + \widehat{S}_{n,j_{q-1}} > 0\} \\
&= \bigcap_{j_1\in\mathcal{J}} \{\widehat{S}_{n,j_1} > 0\},
\end{aligned}$$

because the partial sums being positive is implied by the individual terms being positive. For the same reason, it follows that

$$\begin{aligned}
\bigcap_{l=1}^{L} \mathcal{F}_n^{(2)}(\overline{g}_l) &= \bigcap_{j_1\in\mathcal{J}} \{\widehat{S}_{n,j_1} < 0\} \bigcap_{\{j_1,j_2\}\subset\mathcal{J}} \{\widehat{S}_{n,j_1} + \widehat{S}_{n,j_2} < 0\} \\
&\quad \cdots \bigcap_{\{j_1,j_2,\ldots,j_{q-1}\}\subset\mathcal{J}} \{\widehat{S}_{n,j_1} + \widehat{S}_{n,j_2} + \cdots + \widehat{S}_{n,j_{q-1}} < 0\} \\
&= \bigcap_{j_1\in\mathcal{J}} \{\widehat{S}_{n,j_1} < 0\}.
\end{aligned}$$

Therefore,

$$\mathbb{P}_\delta\left[\bigcap_{l=1}^{L} \mathcal{F}_n^{(1)}(\overline{g}_l)\right] = \mathbb{P}_\delta\left[\bigcap_{j=1}^{q} \{\widehat{S}_{n,j} > 0\}\right] \tag{B.33}$$

and

$$\mathbb{P}_\delta\left[\bigcap_{l=1}^{L} \mathcal{F}_n^{(2)}(\overline{g}_l)\right] = \mathbb{P}_\delta\left[\bigcap_{j=1}^{q} \{\widehat{S}_{n,j} < 0\}\right]. \tag{B.34}$$

Suppose that Assumption 2.2 holds and consider the local alternative that $c'\beta = \lambda + \frac{\delta}{\sqrt{n}}$. Using Lemma A.2, applying (B.31) to (B.34), together with the assumption that the

clusters are independent, it follows that

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}_\delta[T_n > \widehat{\mathrm{cv}}_n(1-\alpha)] &= \prod_{j=1}^{q} \mathbb{P}[Z_j + \xi_j\delta < 0] + \prod_{j=1}^{q} \mathbb{P}[Z_j + \xi_j\delta > 0] \\ &= \prod_{j\in\mathcal{J}} \Phi\left(-\frac{\xi_j\delta}{\sigma_j}\right) + \prod_{j\in\mathcal{J}} \left[1 - \Phi\left(-\frac{\xi_j\delta}{\sigma_j}\right)\right],\end{aligned}$$

where $\Phi(\cdot)$ is the standard normal cdf, and by noting that $\mathbb{P}[Z_j + \xi_j\delta = 0] = 0$ for any $j = 1, \ldots, q$. □

Proof of Property 3.6. Let $\phi(\cdot)$ be the standard normal pdf. The proofs of the properties are as follows.

1. We have

$$\frac{\partial \pi_{\mathrm{L}}(\delta)}{\partial \delta} = -\sum_{j\in\mathcal{J}} \frac{\xi_j}{\sigma_j} \phi\left(-\frac{\xi_j\delta}{\sigma_j}\right) \prod_{l\neq j} \Phi\left(-\frac{\xi_l\delta}{\sigma_l}\right),$$

   and

$$\frac{\partial \pi_{\mathrm{R}}(\delta)}{\partial \delta} = \sum_{j\in\mathcal{J}} \frac{\xi_j}{\sigma_j} \phi\left(-\frac{\xi_j\delta}{\sigma_j}\right) \prod_{l\neq j} \left[1 - \Phi\left(-\frac{\xi_l\delta}{\sigma_l}\right)\right].$$

   The results follow because ξ_j, σ_j, $\phi\left(-\frac{\xi_j\delta}{\sigma_j}\right)$, and $\Phi\left(-\frac{\xi_j\delta}{\sigma_j}\right)$ are nonnegative for all $j \in \mathcal{J}$.

2. The result follows because

$$\pi_{\mathrm{L}}(0) = \prod_{j\in\mathcal{J}} \Phi(0) = \frac{1}{2^q}$$

   and

$$\pi_{\mathrm{R}}(0) = \prod_{j\in\mathcal{J}} [1 - \Phi(0)] = \frac{1}{2^q}.$$

3. Note that $\xi_j > 0$ and $\sigma_j > 0$ for any $j \in \mathcal{J}$. If $\delta < 0$, then

$$\Phi\left(-\frac{\xi_j\delta}{\sigma_j}\right) > \Phi(0) > 1 - \Phi\left(-\frac{\xi_j\delta}{\sigma_j}\right)$$

   for any $j \in \mathcal{J}$. Hence, $\pi_{\mathrm{L}}(\delta) > \frac{1}{2^q} > \pi_{\mathrm{R}}(\delta)$ by multiplying the terms of $j \in \mathcal{J}$.

If $\delta > 0$, then

$$\Phi\left(-\frac{\xi_j \delta}{\sigma_j}\right) < \Phi(0) < 1 - \Phi\left(-\frac{\xi_j \delta}{\sigma_j}\right)$$

for any $j \in \mathcal{J}$. Hence, the result follows from multiplying the terms over $j \in \mathcal{J}$. Similarly, $\pi_{\mathrm{R}}(\delta) > \frac{1}{2^{\eta}} > \pi_{\mathrm{L}}(\delta)$ under this case.

□

Proof of Proposition 4.6.

1. To begin with, I show Assumption 4.4.1 holds under Assumption 4.5. Since $\widetilde{\Omega} = \{\omega^\star\}$, the statement I need to show is

$$\lim_{n\to\infty} \mathbb{P}[\widehat{\omega}_n = \omega^\star] = 1. \tag{B.35}$$

Here, α and δ are fixed throughout the proof, so I write $\widehat{\pi}_n(\omega) \equiv \widehat{\pi}_n(\omega, \delta, \alpha)$ for notational simplicity. Let $\widehat{\Delta}_n(\omega) \equiv \widehat{\pi}_n(\omega) - \pi(\omega)$ for each $\omega \in \Omega$. By Assumption 4.5.1 and the continuous mapping theorem, $\widehat{\pi}_n(\omega) - \widehat{\pi}_n(\omega^\star) \xrightarrow{p} \pi(\omega) - \pi(\omega^\star)$ for any $\omega \in \Omega \backslash \{\omega^\star\}$. For a given $\omega \in \Omega \backslash \{\omega^\star\}$, this means for any $\epsilon(\omega) > 0$,

$$\lim_{n\to\infty} \mathbb{P}[|\widehat{\Delta}_n(\omega) - \widehat{\Delta}_n(\omega^\star)| > \epsilon(\omega)] = 0. \tag{B.36}$$

Next, by Assumptions 4.5.2 and 4.5.3, since $\omega^\star$ is the unique solution to (11), this means there exists $\eta > 0$ such that

$$\pi(\omega^\star) - \pi(\omega) > \eta \tag{B.37}$$

for all $\omega \in \Omega \backslash \{\omega^\star\}$.

Note that

$$\begin{aligned}
\mathbb{P}[\widehat{\omega}_n = \omega^\star] &\geq \mathbb{P}\left[\bigcap_{\omega \in \Omega\backslash\{\omega^\star\}} \{\widehat{\pi}_n(\omega^\star) > \widehat{\pi}_n(\omega)\}\right] \\
&= \mathbb{P}\left[\bigcap_{\omega \in \Omega\backslash\{\omega^\star\}} \{\pi(\omega^\star) + \widehat{\Delta}_n(\omega^\star) > \pi(\omega) + \widehat{\Delta}_n(\omega)\}\right] \\
&= \mathbb{P}\left[\bigcap_{\omega \in \Omega\backslash\{\omega^\star\}} \{\pi(\omega^\star) - \pi(\omega) > \widehat{\Delta}_n(\omega) - \widehat{\Delta}_n(\omega^\star)\}\right]
\end{aligned}$$

$$
\begin{aligned}
&= 1 - \mathbb{P}\left[\bigcup_{\omega\in\Omega\setminus\{\omega^\star\}} \{\pi(\omega^\star) - \pi(\omega) \le \widehat{\Delta}_n(\omega) - \widehat{\Delta}_n(\omega^\star)\}\right] \\
&\ge 1 - \sum_{\omega\in\Omega\setminus\{\omega^\star\}} \mathbb{P}\left[\pi(\omega^\star) - \pi(\omega) \le \widehat{\Delta}_n(\omega) - \widehat{\Delta}_n(\omega^\star)\right] \\
&\ge 1 - \sum_{\omega\in\Omega\setminus\{\omega^\star\}} \mathbb{P}\left[\pi(\omega^\star) - \pi(\omega) \le |\widehat{\Delta}_n(\omega) - \widehat{\Delta}_n(\omega^\star)|\right] \\
&\ge 1 - \sum_{\omega\in\Omega\setminus\{\omega^\star\}} \mathbb{P}\left[\eta \le |\widehat{\Delta}_n(\omega) - \widehat{\Delta}_n(\omega^\star)|\right], \qquad \text{(B.38)}
\end{aligned}
$$

where the first line follows from noting that $\pi_n(\omega^\star) > \widehat{\pi}_n(\omega)$ for any $\omega \in \Omega\setminus\{\omega^\star\}$ implies $\widehat{\omega}_n = \omega^\star$, the second line follows from the definition of $\widehat{\Delta}_n(\omega)$, the third line follows from rearranging terms, the fourth line follows from probability rules, the fifth line follows from the union bound, the sixth line follows from noting that the events in the fifth line imply the events in the sixth line, and the last line follows from (B.37).

Recall that $|\Omega| < \infty$. By choosing $\epsilon(\omega) = \frac{\eta}{2}$ in (B.36) for each $\omega \in \Omega\setminus\{\omega^\star\}$, it follows that

$$
\lim_{n\to\infty} \mathbb{P}[\widehat{\omega}_n = \omega^\star] \ge 1 - \sum_{\omega\in\Omega\setminus\{\omega^\star\}} \lim_{n\to\infty} \mathbb{P}\left[\eta \le |\widehat{\Delta}_n(\omega) - \widehat{\Delta}_n(\omega^\star)|\right] = 1. \qquad \text{(B.39)}
$$

Therefore, (B.35) follows from the squeeze theorem and (B.39).

2. For Assumption 4.4.2, it is equivalent to

$$
\lim_{n\to\infty} |\mathbb{P}[\phi_n(\widehat{\omega}_n) = 1 | \widehat{\omega}_n = \omega^\star] - \mathbb{P}[\phi_n(\omega^\star) = 1]| = 0, \qquad \text{(B.40)}
$$

since $\widetilde{\Omega} = \{\omega^\star\}$ by Assumption 4.5.2.

Note that

$$
\begin{aligned}
&\mathbb{P}[\phi_n(\widehat{\omega}_n) = 1 | \widehat{\omega}_n = \omega^\star] - \mathbb{P}[\phi_n(\omega^\star) = 1] \\
&\quad = \frac{\mathbb{P}[\phi_n(\widehat{\omega}_n) = 1, \widehat{\omega}_n = \omega^\star]}{\mathbb{P}[\widehat{\omega}_n = \omega^\star]} - \mathbb{P}[\phi_n(\omega^\star) = 1] \\
&\quad = \frac{\mathbb{P}[\phi_n(\omega^\star) = 1, \widehat{\omega}_n = \omega^\star]}{\mathbb{P}[\widehat{\omega}_n = \omega^\star]} - \mathbb{P}[\phi_n(\omega^\star) = 1] \\
&\quad = \frac{\mathbb{P}[\phi_n(\omega^\star) = 1] - \mathbb{P}[\phi_n(\omega^\star) = 1, \widehat{\omega}_n \ne \omega^\star]}{\mathbb{P}[\widehat{\omega}_n = \omega^\star]} - \mathbb{P}[\phi_n(\omega^\star) = 1]
\end{aligned}
$$

$$= \mathbb{P}[\phi_n(\omega^\star) = 1] \left(\frac{1}{\mathbb{P}[\widehat{\omega}_n = \omega^\star]} - 1 \right) - \frac{\mathbb{P}[\phi_n(\omega^\star) = 1, \widehat{\omega}_n \neq \omega^\star]}{\mathbb{P}[\widehat{\omega}_n = \omega^\star]}. \tag{B.41}$$

Since $\mathbb{P}[\phi_n(\omega^\star) = 1, \widehat{\omega}_n \neq \omega^\star] \leq \mathbb{P}[\widehat{\omega}_n \neq \omega^\star]$, it follows from (B.35) that

$$\lim_{n\to\infty} \mathbb{P}[\phi_n(\omega^\star) = 1, \widehat{\omega}_n \neq \omega^\star] \leq 1 - \lim_{n\to\infty} \mathbb{P}[\widehat{\omega}_n = \omega^\star] = 0.$$

Hence,

$$\lim_{n\to\infty} \mathbb{P}[\phi_n(\omega^\star) = 1, \widehat{\omega}_n \neq \omega^\star] = 0 \tag{B.42}$$

by the squeeze theorem. Since CRS controls size of any $\omega \in \Omega$ under Assumption 2.2, it must also hold for $\omega^\star$, i.e.,

$$\lim_{n\to\infty} \mathbb{P}[\phi_n(\omega^\star) = 1] \leq \alpha. \tag{B.43}$$

Using (B.35), (B.41) to (B.43), it follows that

$$\lim_{n\to\infty} \left(\mathbb{P}[\phi_n(\widehat{\omega}_n) = 1 | \widehat{\omega}_n = \omega] - \mathbb{P}[\phi_n(\omega) = 1] \right) = 0.$$

This implies that Assumption 4.4.2 holds.

□

Proof of Theorem 4.7. Under Assumption 2.2 with $\mathcal{J}$ replaced by any $\omega \in \Omega$ and the null hypothesis, the following holds

$$\lim_{n\to\infty} \mathbb{P}[\phi_n(\omega) = 1] \leq \alpha \qquad \text{for each } \omega \in \Omega. \tag{B.44}$$

This is because for any given grouping of clusters $\omega \in \Omega$, CRS controls size.

Next, consider the test based on the data-driven procedure under the null hypothesis, where $\widehat{\omega}_n$ is the grouping of clusters given by the procedure:

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}[\phi_n(\widehat{\omega}_n) = 1] \\
&= \lim_{n\to\infty} \sum_{\omega\in\Omega} \mathbb{P}[\phi_n(\widehat{\omega}_n) = 1 | \widehat{\omega}_n = \omega] \mathbb{P}[\widehat{\omega}_n = \omega] \\
&= \lim_{n\to\infty} \sum_{\omega\in\Omega} \mathbb{P}[\phi_n(\omega) = 1 | \widehat{\omega}_n = \omega] \mathbb{P}[\widehat{\omega}_n = \omega] \\
&= \lim_{n\to\infty} \sum_{\omega\in\Omega} \mathbb{P}[\phi_n(\omega) = 1 | \widehat{\omega}_n = \omega] \left(\mathbb{P}[\widehat{\omega}_n = \omega, \widehat{\omega}_n \in \widetilde{\Omega}] + \mathbb{P}[\widehat{\omega}_n = \omega, \widehat{\omega}_n \notin \widetilde{\Omega}] \right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{\omega\in\Omega} \lim_{n\to\infty} |\mathbb{P}[\phi_n(\omega)=1|\widehat{\omega}_n=\omega] - \mathbb{P}[\phi_n(\omega)=1] + \mathbb{P}[\phi_n(\omega)=1]| \\
&\qquad \cdot \lim_{n\to\infty} \left(\mathbb{P}[\widehat{\omega}_n=\omega, \widehat{\omega}_n\in\widetilde{\Omega}] + \mathbb{P}[\widehat{\omega}_n=\omega, \widehat{\omega}_n\notin\widetilde{\Omega}]\right) \\
&= \sum_{\omega\in\Omega} \lim_{n\to\infty} |\mathbb{P}[\phi_n(\omega)=1|\widehat{\omega}_n=\omega] - \mathbb{P}[\phi_n(\omega)=1] + \mathbb{P}[\phi_n(\omega)=1]| \\
&\qquad \cdot \lim_{n\to\infty} \mathbb{P}[\widehat{\omega}_n=\omega, \widehat{\omega}_n\in\widetilde{\Omega}] \\
&\leq \sum_{\omega\in\Omega} \lim_{n\to\infty} |\mathbb{P}[\phi_n(\omega)=1]| \cdot \lim_{n\to\infty} \mathbb{P}[\widehat{\omega}_n=\omega, \widehat{\omega}_n\in\widetilde{\Omega}] \\
&\leq \alpha \sum_{\omega\in\Omega} \lim_{n\to\infty} \mathbb{P}[\widehat{\omega}_n=\omega, \widehat{\omega}_n\in\widetilde{\Omega}] \\
&= \alpha \lim_{n\to\infty} \sum_{\omega\in\Omega} \mathbb{P}[\widehat{\omega}_n=\omega, \widehat{\omega}_n\in\widetilde{\Omega}] \\
&\leq \alpha,
\end{aligned}
$$

where the first equality uses the law of total probability, the second equality applies the condition that $\widehat{\omega}_n = \omega$, the third equality uses the law of total probability, the fourth equality uses the fact that $|\Omega|$ is finite and adds and subtracts, the fifth equality uses Assumption 4.4.1 on the existence of the set $\widetilde{\Omega}$ such that $\lim_{n\to\infty} \mathbb{P}[\widehat{\omega}_n \notin \widetilde{\Omega}] = 0$, the first inequality uses Assumption 4.4.2, the triangle inequality, the next inequality uses (B.44), the next equality exchanges the summation and limit using the fact that $|\Omega|$ is finite again, and the last equality uses $\{\widehat{\omega}_n = \omega, \widehat{\omega}_n \in \widetilde{\Omega}\} \subseteq \{\widehat{\omega}_n = \omega\}$ and that probabilities over $\omega \in \Omega$ sum to 1. □

Proof of Proposition 4.10. Assume without loss of generality that $\delta > 0$. Hence, Algorithm 4.9 uses optimization problem (17) to approximate the solution of problem (13). Following the statement of the proposition, let $\pi^\star$ be the optimal value to optimization problem (13). In addition, let $\{z^\star_{j,r}\}$ be the solution to (13) such that it gives the objective value $\pi^\star$. Denote

$$
\pi^\star_{\mathrm{L}} \equiv \sum_{j=1}^{\bar{q}} \sum_{r=1}^{\bar{q}} z^\star_{j,r} \log \Psi_{j,r} \tag{B.45}
$$

and

$$
\pi^\star_{\mathrm{R}} \equiv \sum_{j=1}^{\bar{q}} \sum_{r=1}^{\bar{q}} z^\star_{j,r} \log(1-\Psi_{j,r}), \tag{B.46}
$$

so that $\pi^\star = \pi^\star_{\mathrm{L}} + \pi^\star_{\mathrm{R}}$. Moreover, let

$$
\mathcal{L} \equiv \left\{ \sum_{j=1}^{\bar{q}} \sum_{r=1}^{\bar{q}} z_{j,r} \log \Psi_{j,r} : \{z_{j,r}\} \text{ is a feasible solution to (13)} \right\}
$$

be the set that collects all unique values of $\sum_{j=1}^{\overline{q}} \sum_{r=1}^{\overline{q}} z_{j,r} \log \Psi_{j,r}$ based on the feasible solutions to (13). Since $\overline{q}$ is fixed, it follows that $|\mathcal{L}|$ is finite. Let $\pi_{\mathrm{L}}^{(1)} < \pi_{\mathrm{L}}^{(2)} < \cdots < \pi_{\mathrm{L}}^{(|\mathcal{L}|)}$ be the ordered and distinct values of in $\mathcal{L}$. Here, $\pi_{\mathrm{L}}^{(1)} > 0$ because δ, as well as $\xi_{j,r}$ and $\sigma_{j,r}$ for $j, r = 1, \ldots, \overline{q}$ are finite. In addition, $\pi_{\mathrm{L}}^{\star} \in \mathcal{L}$ because $\{z_{j,r}^{\star}\}$ is a feasible solution to (13).

Let $\eta \equiv \min_{\ell \in \{1,\ldots,|\mathcal{L}|-1\}} (\log \pi_{\mathrm{L}}^{(\ell+1)} - \log \pi_{\mathrm{L}}^{(\ell)})$. Note that $\eta > 0$ by construction. In addition, let $A_0 \equiv \lceil \frac{2(\log \pi_{\mathrm{L}}^{(|\mathcal{L}|)} - \log \pi_{\mathrm{L}}^{(1)})}{\eta} \rceil$. Next, choose $\epsilon_{\mathrm{lb}} = \pi_{\mathrm{L}}^{(1)}$ and $\epsilon_{\mathrm{ub}} = \pi_{\mathrm{L}}^{(|\mathcal{L}|)}$. Partition $[\epsilon_{\mathrm{lb}}, \epsilon_{\mathrm{ub}}]$ into A_0 intervals, so that the length of each interval $\log \epsilon_a - \log \epsilon_{a-1}$ is at most $\frac{\eta}{2}$. For each interval, it can contain at most one $\pi_{\mathrm{L}}^{(\ell)}$ for $\ell = 1, \ldots, |\mathcal{L}|$ because the width of each interval is strictly less than η. If $\pi_{\mathrm{L}}^{\star}$ lies in the boundary of an interval, i.e., there exists a' such that $\epsilon_{a'} = \pi_{\mathrm{L}}^{\star}$, partition the interval $[\epsilon_{\mathrm{lb}}, \epsilon_{\mathrm{ub}}]$ into $A_0 + 1$ equally-spaced intervals instead, and check if $\pi_{\mathrm{L}}^{\star}$ lies in the interior of one of the intervals. If $\pi_{\mathrm{L}}^{\star}$ is still on the boundary, repeat the above step again by partitioning $[\epsilon_{\mathrm{lb}}, \epsilon_{\mathrm{ub}}]$ into finer intervals, until $\pi_{\mathrm{L}}^{\star}$ lies in the interior of one of the intervals. In addition, define $[\underline{\epsilon}, \overline{\epsilon}]$ as the subinterval that contains $\pi_{\mathrm{L}}^{\star}$.

Consider any choices of $\log \epsilon_{a'-1}$ and $\log \epsilon_{a'}$ with $\log \epsilon_{a'-1} \neq \underline{\epsilon}$ and $\log \epsilon_{a'} \neq \overline{\epsilon}$ such that (16) is feasible. Let $\{\widetilde{z}_{j,r}\}$ be the corresponding optimal solution to (16) with this choice of interval $[\log \epsilon_{a'-1}, \log \epsilon_{a'}]$. It must be the case that

$$\sum_{j=1}^{\overline{q}} \sum_{r=1}^{\overline{q}} \widetilde{z}_{j,r} \log \Psi_{j,r} \neq \pi_{\mathrm{L}}^{\star}$$

by the choice of a'. Assume to the contrary that

$$\sum_{j=1}^{\overline{q}} \sum_{r=1}^{\overline{q}} \widetilde{z}_{j,r} \log \Psi_{j,r} + \sum_{j=1}^{\overline{q}} \sum_{r=1}^{\overline{q}} \widetilde{z}_{j,r} \log(1 - \Psi_{j,r}) > \pi^{\star}. \tag{B.47}$$

Note that $\{\widetilde{z}_{j,r}\}$ is feasible to (13) with this choice of $[\log \epsilon_{a'-1}, \log \epsilon_{a'}]$ because (13) contains a subset of constraints of (16). But (B.47) contradicts that $\{z_{j,r}^{\star}\}$ is an optimal solution to (13) because $\{\widetilde{z}_{j,r}\}$ gives a higher optimal value.

Let (P) represents the optimization problem (16) with $\log \epsilon_{a-1} = \underline{\epsilon}$ and $\log \epsilon_a = \overline{\epsilon}$. Note that (P) is feasible since $\{z_{j,r}^{\star}\}$ is one such feasible solution. Let $\{\overline{z}_{j,r}\}$ be the optimal solution to (P). By construction, it must be the case that $\sum_{j=1}^{\overline{q}} \sum_{r=1}^{\overline{q}} \overline{z}_{j,r} \log \Psi_{j,r} = \pi_{\mathrm{L}}^{\star}$. In addition, $\{\overline{z}_{j,r}\}$ must also be feasible to (13) because (13) contains a subset of the constraints of (16).

Now, let $\overline{\pi}_{\mathrm{R}} \equiv \sum_{j=1}^{\overline{q}} \sum_{r=1}^{\overline{q}} \overline{z}_{j,r} \log(1 - \Psi_{j,r})$. It remains to show that $\overline{\pi}_{\mathrm{R}} = \pi_{\mathrm{R}}^{\star}$. First, suppose $\overline{\pi}_{\mathrm{R}} > \pi_{\mathrm{R}}^{\star}$. This implies that $\overline{\pi}_{\mathrm{L}} + \overline{\pi}_{\mathrm{R}} > \pi_{\mathrm{L}}^{\star} + \pi_{\mathrm{R}}^{\star}$. But $\{\overline{z}_{j,r}\}$ is feasible to (13), this contradicts that $\{z_{j,r}^{\star}\}$ is an optimal solution to (13). Next, suppose $\overline{\pi}_{\mathrm{R}} < \pi_{\mathrm{R}}^{\star}$. This implies that $\overline{\pi}_{\mathrm{L}} + \overline{\pi}_{\mathrm{R}} < \pi_{\mathrm{L}}^{\star} + \pi_{\mathrm{R}}^{\star}$. But $\{z_{j,r}^{\star}\}$ is also feasible to (P) because it satisfies the constraint

$$\sum_{j=1}^{\overline{q}} \sum_{r=1}^{\overline{q}} z_{j,r}^{\star} \log \Psi_{j,r} = z_{\mathrm{L}}^{\star} \in [\log \underline{\epsilon}, \log \overline{\epsilon}].$$

Thus, this contradicts that $\{\overline{z}_{j,r}\}$ is an optimal solution to (P). Therefore, it must be the case that $\overline{\pi}_{\mathrm{R}} = \pi_{\mathrm{R}}^{\star}$ and the proof is complete.

If $\delta < 0$, then the above proof can be modified by changing (17) to (16) and reversing the roles of the terms related to (B.45) and (B.46) appropriately. □

C More details on the empirical application

This section describes how the calibrated simulation exercise is conducted using the data of Dincecco and Katz (2016). I create a balanced panel based on their data. Let T be the longest time period of the data. Here, I focus on column (1) of Table 3 of Dincecco and Katz (2016), so that the regression follows (20). The calibrated simulation procedure is as follows.

1. Estimate the model to obtain $(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \{\widehat{\mu}_j\}_{j \in \mathcal{J}})$ and obtain the residuals as

$$\widehat{U}_{t,j} = Y_{t,j} - (\widehat{\beta}_0 + \widehat{\beta}_1 C_{t,j} + \widehat{\beta}_2 L_{t,j} + \widehat{\mu}_j),$$

   for all $j \in \mathcal{J}$ and $t \in \{1, \ldots, T\} \equiv \mathcal{T}$.

2. Use the residuals to fit an AR(1) model

$$U_{t,j} = \rho_j U_{t-1,j} + \epsilon_{t,j},$$

   where $\epsilon_{t,j} \sim N(0, v_j^2)$ for each country $j \in \mathcal{J}$. Hence, there is an estimate of $\widehat{\rho}_j$ and $\widehat{v}_j^2$ for each country $j \in \mathcal{J}$.

3. Conduct a Monte Carlo simulation with B replications. Each replication $b = 1, \ldots, B$ is as follows.

   (a) For each country $j \in \mathcal{J}$:

      i. Draw the residuals $\widetilde{U}_{t,j}^{(b)}$ based on the model estimated in Step 2 from $t = 1$

to $t = T$.

ii. Next, set the values of the treatment variables to create variation across countries and time.

- If country j has variation in $C_{t,j}$ in the original data, set $t_{C,j} = \lfloor \frac{3}{4}T \rfloor - 5j$. Otherwise, set $t_{C,j} = 0$.
- If country j has variation in $L_{t,j}$ in the original data, set $t_{L,j} = \lfloor \frac{3}{4}T \rfloor - 8(11 - j)$. Otherwise, set $t_{L,j} = 0$.

iii. Set $C_{t,j} = \mathbb{1}[t > t_{C,j}]$ and $L_{t,j} = \mathbb{1}[t > t_{L,j}]$ for all $t \in \mathcal{T}$.

(b) Suppose the goal is to perform inference on the coefficient of $C_{t,j}$. Then,

i. Let $\widehat{\beta}_1 + \Delta$ be the value of the alternative.

ii. Keep $\widehat{\beta}_2$ unchanged.

iii. Generate the outcome as

$$\widetilde{Y}_{t,j}^{(b)} = \widehat{\beta}_0 + (\widehat{\beta}_1 + \Delta)C_{t,j} + \widehat{\beta}_2 L_{t,j} + \widehat{\mu}_j + \widetilde{U}_{t,j}^{(b)},$$

for all $j \in \mathcal{J}$ and $t \in \mathcal{T}$.

(c) Suppose the goal is to perform inference on the coefficient of $L_{t,j}$. Then,

i. Let $\widehat{\beta}_2 + \Delta$ be the value of the alternative.

ii. Keep $\widehat{\beta}_1$ unchanged.

iii. Generate the outcome as

$$\widetilde{Y}_{t,j}^{(b)} = \widehat{\beta}_0 + \widehat{\beta}_1 C_{t,j} + (\widehat{\beta}_2 + \Delta)L_{t,j} + \widehat{\mu}_j + \widetilde{U}_{t,j}^{(b)},$$

for all $j \in \mathcal{J}$ and $t \in \mathcal{T}$.

The procedure for the other two columns is similar when time fixed effects and/or time trends are added to the regression equation.

References

ABADIE, A., S. ATHEY, G. W. IMBENS, AND J. M. WOOLDRIDGE (2022): "When Should You Adjust Standard Errors for Clustering?" *The Quarterly Journal of Economics*, 138, 1–35.

ALSAN, M. AND C. GOLDIN (2019): "Watersheds in Child Mortality: The Role of Effective Water and Sewerage Infrastructure, 1880–1920," *Journal of Political Economy*, 127, 586–638.

BERTRAND, M., E. DUFLO, AND S. MULLAINATHAN (2004): "How Much Should We Trust Differences-In-Differences Estimates?" *The Quarterly Journal of Economics*, 119, 249–275.

BESTER, A., T. CONLEY, AND C. HANSEN (2011): "Inference with dependent data using cluster covariance estimators," *Journal of Econometrics*, 165, 137–151.

BLOOM, N., B. EIFERT, A. MAHAJAN, D. MCKENZIE, AND J. ROBERTS (2013): " Does Management Matter? Evidence from India," *The Quarterly Journal of Economics*, 128, 1–51.

BURDE, D. AND L. L. LINDEN (2013): "Bringing Education to Afghan Girls: A Randomized Controlled Trial of Village-Based Schools," *American Economic Journal: Applied Economics*, 5, 27–40.

CAI, Y. (2024): "Some Finite Sample Properties of the Sign Test," *Working Paper*.

CAI, Y., I. A. CANAY, D. KIM, AND A. M. SHAIKH (2023): "On the Implementation of Approximate Randomization Tests in Linear Models with a Small Number of Clusters," *Journal of Econometric Methods*, 12, 85–103.

CAMERON, A., J. GELBACH, AND D. MILLER (2008): "Bootstrap-Based Improvements for Inference with Clustered Errors," *The Review of Economics and Statistics*, 90, 414–427.

CAMERON, A. C. AND D. L. MILLER (2015): "A practitioner's guide to cluster-robust inference," *Journal of human resources*, 50, 317–372.

CANAY, I. A., J. P. ROMANO, AND A. M. SHAIKH (2017a): "Randomization Tests under an Approximate Symmetry Assumption," *Econometrica*, 85, 1013–1030.

——— (2017b): "Supplement to 'Randomization Tests under an Approximate Symmetry Assumption'," *Econometrica Supplemental Material*, 85.

CANAY, I. A., A. SANTOS, AND A. M. SHAIKH (2021): "The Wild Bootstrap with a "small" number of "large" Clusters," *The Review of Economics and Statistics*, 2021.

CAO, J., C. HANSEN, D. KOZBUR, AND L. VILLACORTA (2022): "Inference for Dependent Data with Learned Clusters," *Working Paper*.

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

CONLEY, T., S. GONÇALVES, AND C. HANSEN (2018): "Inference with Dependent Data in Accounting and Finance Applications," *Journal of Accounting Research*, 56, 1139–1203.

CONLEY, T. G. AND C. R. TABER (2011): "Inference with "Difference in Differences" with a Small Number of Policy Changes," *The Review of Economics and Statistics*, 93, 113–125.

CROES, G. A. (1958): "A Method for Solving Traveling-Salesman Problems," *Operations Research*, 6, 791–812.

DINCECCO, M. AND G. KATZ (2016): "State Capacity and Long-run Economic Performance," *The Economic Journal*, 126, 189–218.

DJOGBENOU, A. A., J. G. MACKINNON, AND M. ØRREGAARD NIELSEN (2019): "Asymptotic theory and wild bootstrap inference with clustered errors," *Journal of Econometrics*, 212, 393–412.

FLOOD, M. M. (1956): "The Traveling-Salesman Problem," *Operations Research*, 4, 61–75.

GOODMAN, S. (2016): "Learning from the Test: Raising Selective College Enrollment by Providing Information," *The Review of Economics and Statistics*, 98, 671–684.

HAGEMANN, A. (2022): "Permutation inference with a finite number of heterogeneous clusters," *Working Paper*.

HOEFFDING, W. (1952): "The Large-Sample Power of Tests Based on Permutations of Observations," *The Annals of Mathematical Statistics*, 23, 169 – 192.

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

——— (2016): "Inference with Few Heterogeneous Clusters," *The Review of Economics and Statistics*, 98, 83–96.

LIN, S. AND B. W. KERNIGHAN (1973): "An Effective Heuristic Algorithm for the Traveling-Salesman Problem," *Operations Research*, 21, 498–516.

MACKINNON, J. G. AND M. D. WEBB (2017): "Wild Bootstrap Inference for Wildly Different Cluster Sizes," *Journal of Applied Econometrics*, 32, 233–254.

MACKINNON, J. G., M. ØRREGAARD NIELSEN, AND M. D. WEBB (2023): "Cluster-robust inference: A guide to empirical practice," *Journal of Econometrics*, 232, 272–299.

NEWEY, W. K. AND K. D. WEST (1987): "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.

VAN DER VAART, A. W. (2000): *Asymptotic statistics*, vol. 3, Cambridge university press.

WEBB, M. D. (2023): "Reworking wild bootstrap-based inference for clustered errors," *Canadian Journal of Economics/Revue canadienne d'économique*, 56, 839–858.