

Some impossibility results for inference with cluster dependence with large clusters

Authors	Kojevnikov, Denis; Song, Kyungchul
Published in	Journal of Econometrics
DOI	10.1016/j.jeconom.2023.105524
Publication Date	2023-12
Document Version	publishersversion
Link	https://research.tilburguniversity.edu/en/publications/80b8e4ed-54bc-4a34-883f-f581104932ae
Citation	Kojevnikov, D & Song, K 2023, 'Some impossibility results for inference with cluster dependence with large clusters', Journal of Econometrics, vol. 237, no. 2, 105524. https://doi.org/10.1016/j.jeconom.2023.105524
Download Date	2026-05-17 12:49:58
Rights	<p>General rights</p> <p>Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.</p> <ul style="list-style-type: none"> - Users may download and print one copy of any publication from the public portal for the purpose of private study or research. - You may not further distribute the material or use it for any profit-making activity or commercial gain - You may freely distribute the URL identifying the publication in the public portal" <p>Take down policy</p> <p>If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.</p>



Some impossibility results for inference with cluster dependence with large clusters[☆]

Denis Kojevnikov, Kyungchul Song^{*}

Tilburg University, Netherlands
University of British Columbia, Canada

ARTICLE INFO

JEL classification:

C01
C12
C13

Keywords:

Consistent discrimination
Local dependence
Unknown dependence structure
Consistent estimation of long-run variance
Cluster dependence
Log likelihood process

ABSTRACT

This paper focuses on a setting with observations having a cluster dependence structure and presents two main impossibility results. First, we show that when there is only one large cluster, i.e., the researcher does not have any knowledge on the dependence structure of the observations, it is not possible to consistently discriminate the mean. When within-cluster observations satisfy the uniform central limit theorem, we also show that a sufficient condition for consistent \sqrt{n} -discrimination of the mean is that we have at least two large clusters. This result shows some limitations for inference when we lack information on the dependence structure of observations. Our second result provides a necessary and sufficient condition for the cluster structure that the long run variance is consistently estimable. Our result implies that when there is at least one large cluster, the long run variance is not consistently estimable.

1. Introduction

Statistical inference from data usually begins by imposing a form of a dependence structure on the data, by specifying which groups of observations exhibit strong within-group dependence. Various tools of asymptotic inference such as the law of large numbers and the central limit theorem are available for many typically imposed dependence structures. A standard case is the independence assumption or an assumption on time series dependence. However, it is well known that in the case of cross-sectional dependence, a researcher is often less confident about the correctness of the dependence structure used, despite its crucial role for inference.

A popular way to deal with this challenge is to use cluster dependence modeling, where the dependence structure among observations within each cluster is left unspecified, while independence is imposed between observations from different clusters. The inference procedures when there are many clusters are well known and can be analyzed using standard methods of asymptotic inference. However, less is known about the case where there are large clusters, and the dependence structure within such a cluster is unknown. Cameron et al. (2008) proposed a wild bootstrap procedure and showed by simulations that their tests perform well even when there are a small number of clusters. The robustness of this result was confirmed by MacKinnon and Webb (2017) even when the sizes of the clusters are highly heterogeneous. This large cluster issue has also drawn interest in the literature of difference-in-differences when there are only few treated clusters (see Conley and Taber (2011), Hagemann (2019), and MacKinnon

[☆] We thank Michael Leung and James MacKinnon for valuable comments. We thank the Associate Editor and three anonymous referees for their constructive comments which helped us improve our paper's exposition and results. All errors are ours. Song acknowledges financial support from Social Sciences and Humanities Research Council of Canada.

^{*} Correspondence to: Vancouver School of Economics, University of British Columbia, 6000 Iona Drive, Vancouver, V6T 1L4, Canada.
E-mail address: kysong@mail.ubc.ca (K. Song).

and Webb (2020), and references therein). Djogbenou et al. (2019) studied inference on regression models with clustered errors. They provided conditions for the cluster sizes so that asymptotic and bootstrap inferences are asymptotically valid. They showed that their conditions exclude the presence of a large cluster.

There are several methods proposed to deal with the problem of inference with large clusters. Donald and Lang (2007) and Bester et al. (2011) considered linear models and proposed inference where the asymptotic distribution of the long run variance estimator is fully known. This approach is related to the HAR (Heteroskedasticity-Autocorrelation Robust) inference of Kiefer and Vogelsang (2002) and Sun (2014) in time series, which uses a normalization by an inconsistent long run variance estimator that has a stochastic limit.

Ibragimov and Müller (2010) proposed a t -test approach based on within-cluster estimators together with a t -distribution, where the degree of freedom in the t distribution is determined by the number of clusters. They used the result of Bakirov and Szekely (2005) and showed that their approach is asymptotically valid, even if the variances of the cluster specific estimators are different across the clusters. Ibragimov and Müller (2016) extended these results to the problem of two-sample comparison and developed a testing procedure for the level of clustering.

Some studies adopted the approach of randomized testing to deal with cluster dependence with large clusters. Canay et al. (2017) developed asymptotic inference procedures when the inference involves statistics whose limiting distribution satisfy symmetry properties. Hagemann (2019) proposed randomized tests for treatment effects when there are only a small number of clusters. Like Ibragimov and Müller (2010), both proposals assumed large sample properties for within-cluster statistics. A recent work by Canay et al. (2021) used the analog between wild bootstrap and randomized tests, and provided conditions under which the wild bootstrap for cluster-dependent regression models is asymptotically valid when there are only a small number of clusters.

Our paper focuses on observations with a cluster dependence structure and explores implications on statistical inference when there are large clusters. First, we show that when the sample consists of large clusters, the mean cannot be consistently discriminated if there is only one cluster, i.e., the researcher does not have any knowledge on the dependence structure of the data. Furthermore, when the observations form large clusters and within-cluster observations satisfy the uniform central limit theorem, a sufficient condition for the mean to be consistently discriminated at the rate of \sqrt{n} is that the sample consists of at least two large clusters.

This impossibility result has a significant implication in a setting where the researcher does not know the dependence structure of observations. In such a case, consistent discrimination of the mean is not possible with uniform-in- P asymptotic size control. Note that Song (2016) proposed a randomized subsampling approach, and Leung (2021) provided a set of general conditions for the approach to produce asymptotically valid inference. Both focus on a setting where no knowledge on the dependence structure is required. Among other things, their results show that the mean is consistently discriminated. Our impossibility result on consistent discrimination considers a setting where there is no uniform upper bound of the long run variance in the null model, and this setting is excluded by part of their conditions. Hence, their results do not contradict our impossibility result.

Our second result is concerned with consistent estimation of long run variances. More specifically, suppose that $X_n = [X_{n,1}, \dots, X_{n,n}]^T$ is a given random vector of dimension n , where each observation $X_{n,i}$ has the same mean μ . Let us define the long-run variance of X_n as follows¹:

$$\sigma_{LR}^2 = \text{Var} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n X_{n,i} \right). \quad (1.1)$$

Recently, Hansen and Lee (2019) derived an asymptotic distribution theory for clustered data, including a law of large numbers and a central limit theorem. One of their results presents a condition for the cluster structure that is necessary and sufficient for the weak law of large numbers to hold for the sample average of the clustered observations. Our paper shows that the same condition is in fact necessary and sufficient for the consistent estimability of the long run variance of the clustered observations as well. Our condition for the cluster structure also implies that when there is at least one large cluster, i.e., the researcher does not know the dependence structure on a nonnegligible portion of the data, the long-run variance is not consistently estimable. It is not hard to show that the existing cluster-robust variance estimators are inconsistent when the cluster structure is severely misspecified. However, to the best of our knowledge, it has not been known whether there exists any consistent estimator of the long run variance when there is a lack of knowledge on the dependence structure on a nonnegligible portion of the data. Our result gives a negative answer to this question.

There has long been a strand of literature that studies impossibility of estimation and inference. (See, e.g., Bahadur and Savage (1956), Dufour (1997), Pötscher (2002), Romano (2004), and Bertanha and Moreira (2020).) The impossibility of consistent estimation of a long run variance in this paper is related to Pötscher (2002) who established a minimax risk lower bound for a general estimation problem. Among others, his result can be used to prove the impossibility of consistent estimation uniform in P as shown in Corollary 3.2 there. However, we cannot apply this corollary in our setting, because our probability model is not indexed by a set of parameters fixed independently of the sample size, such as \mathcal{H} in his paper. This stems from our setting where we have to deal with the *joint distribution of the entire sample* whose dependence structure varies in the model as n changes. Bertanha and Moreira (2020) studied impossibility results of two types: indistinguishability of the null hypothesis from the alternative hypothesis and unbounded confidence sets. Their study of impossibility of the first type is related to impossibility of consistent discrimination

¹ Note that when there is a common shock, say, C_n , such as cluster-specific fixed effects with few clusters, the analysis in this paper carries over to this case with σ_{LR}^2 replaced by the conditional variance given common shock C_n . Our impossibility results do not depend on whether there is a common shock of this form in the data or not. For simplicity, we consider a setting without such cluster-specific fixed effects.

of the mean in our paper. For this result, they assume that for each probability in the alternative hypothesis, there is a sequence of probabilities under the null hypothesis that weakly converge to this probability. Our setting does not satisfy this assumption in general. Hence, our result does not fall into their framework. Menzel (2021) recently developed and verified the validity of a bootstrap procedure in multi-way clustered observations with two or more dimensions. Part of his results shows that it is not possible to consistently estimate the distribution of the cluster dependent observations. Our results are not the special case of his results, because our impossibility result holds for models that exclude the counterexample that he used to prove the impossibility result. In particular, our cluster dependence accommodates within-cluster heterogeneity in terms of marginal distributions and dependence structures.

The rest of the paper is organized as follows. The next section studies the consistent discrimination of the mean. Section 3 is devoted to presenting the result of the impossibility of consistent estimation of the long run variance. In Section 4, we illustrate the implication of our results for the case of difference-in-difference models. In Section 5, we conclude. The mathematical proofs are found in the Appendix.

2. Cluster dependence

Let $X_n = [X_{n,1}, \dots, X_{n,n}]^T$ be a random vector with a joint distribution P_n which belongs to the class of distributions \mathcal{P}_n . Throughout the paper, we assume that for each $P_n \in \mathcal{P}_n$,

$$\mathbf{E}[X_{n,1}] = \mathbf{E}[X_{n,2}] = \dots = \mathbf{E}[X_{n,n}],$$

and $\sigma_{LR}^2 < \infty$, where σ_{LR}^2 is defined in (1.1). In many situations, the dependence structure is partially observed. Here we consider cluster dependence, where the dependence structure is entirely unknown within each cluster, and observations are independent between clusters. Let $N_{n,m}, m = 1, \dots, M_n$, be a partition of $N_n = \{1, \dots, n\}$ such that $|N_{n,m}| = n_m$ for each $m = 1, \dots, M_n$, so that $\sum_{m=1}^{M_n} n_m = n$. Define $\mathcal{M}_n = \{N_{n,m} : m = 1, \dots, M_n\}$ and call it a **cluster structure**. Throughout the paper we assume that $(X_{n,i})_{i \in N_{n,m}}$ are independent across m 's under all $P_n \in \mathcal{P}_n$, i.e., the joint distribution of X_n has a cluster dependence structure. For future references, we define

$$\bar{X}_{n,m} = \frac{1}{n_m} \sum_{i \in N_{n,m}} X_{n,i} \quad \text{and} \quad \sigma_{n,m}^2 = \text{Var} \left(\sqrt{n_m} \bar{X}_{n,m} \right),$$

so that $\bar{X}_{n,m}$ represents the within-cluster mean of $X_{n,i}$'s and $\sigma_{n,m}^2$ represents the within-cluster long-run variance of $X_{n,i}$'s.

Our impossibility results rely on the assumption that the probability model, \mathcal{P}_n , includes Gaussian experiments with what we call local-to-independence common shocks. For each cluster $m = 1, \dots, M_n$, and for $\delta > 0$ and $\sigma^2 > 0$, we define

$$\Sigma_{n,m}(\sigma^2, \delta) = \sigma^2 \left(\left(1 - \frac{\delta}{n_m} \right) I_{n_m} + \frac{\delta}{n_m} \mathbf{1}_{n_m} \mathbf{1}_{n_m}^T \right),$$

where I_{n_m} denotes the n_m -dimensional identity matrix and $\mathbf{1}_{n_m}$ is the n_m -dimensional column vector of ones. Let $\Sigma_n(\sigma^2, \delta)$ be the $n \times n$ block diagonal matrix whose m -th block is given by $\Sigma_{n,m}(\sigma^2, \delta)$. Suppose that $\Sigma_n(\sigma^2, \delta)$ is positive definite. Then, for each $\mu_n \in \mathbf{R}^n$, we let $\Phi(\mu_n, \Sigma_n(\sigma^2, \delta))$ be the multivariate normal distribution with mean μ_n and covariance matrix $\Sigma_n(\sigma^2, \delta)$. Define

$$\mathcal{P}_{n,\mathcal{N}} = \left\{ \Phi(\bar{\mu} \mathbf{1}_n, \Sigma_n(\sigma^2, \delta)) : \sigma^2 > 0, \delta \in (0, 1] \text{ and } \bar{\mu} \in \mathbf{R} \right\}.$$

The set $\mathcal{P}_{n,\mathcal{N}}$ represents a Gaussian model, where each member is a multivariate normal distribution with a common mean and an equal covariance. We call each $\Phi(\mu_n, \Sigma_n(\sigma^2, \delta))$ the **local-to-independence common shock (LTIC) Gaussian distribution** with parameters σ^2 and δ . This Gaussian distribution represents the cross-sectional dependence structure of $X_{n,i}$'s generated as follows:

$$X_{n,i} = \mu_{n,i} + \varepsilon_i \sqrt{1 - \frac{\delta}{n_m}} + \eta_m \sqrt{\frac{\delta}{n_m}}, \quad \text{whenever } i \in N_m,$$

where $\mu_{n,i}$ is the i -th entry of μ_n , ε_i 's are i.i.d. normal random variables with mean zero and variance σ^2 , and $\eta_m, m = 1, \dots, M_n$, are i.i.d. normal random variables with mean zero and variance σ^2 , independent of ε_i 's. Each random variable η_m represents a within-cluster ‘‘common shock’’, and creates the within-cluster global dependence among $X_{n,i}$'s. The influence of this common shock on the random variable $X_{n,i}$ diminishes at the rate of $\sqrt{n_m}$.

3. Consistent discrimination of the mean

3.1. Consistent discrimination of the mean

Let us explore the consistent discrimination of the mean under the general cluster dependence structure. We introduce the notion of consistent discrimination formally. Let \mathcal{P}_n be a set of the distributions of $X_n \in \mathbf{R}^n$ such that $\mathbf{E}[X_{n,i}]$ is identical across i for each $n \geq 1$. Let $\mathcal{P}_{n,0} = \{P_n \in \mathcal{P}_n : \mathbf{E}[X_{n,i}] = 0\}$, i.e., the set of probabilities under the null hypothesis of $\mathbf{E}[X_{n,i}] = 0$.

Definition 3.1. The mean of $X_{n,i}$, $n \geq 1$, is **consistently discriminated at level $\alpha \in (0, 1)$ in model \mathcal{P}_n** , if there is a sequence of (potentially randomized) tests $\{\varphi_n\}_{n \geq 1}$ such that

$$\limsup_{n \rightarrow \infty} \mathbf{E}[\varphi_n(X_n)] \leq \alpha,$$

along any sequence $P_{n,0} \in \mathcal{P}_{n,0}$, and

$$\liminf_{n \rightarrow \infty} \mathbf{E}[\varphi_n(X_n)] = 1,$$

along any sequence $P_n \in \mathcal{P}_n$ such that $\liminf_{n \geq 1} \mathbf{E}[X_{n,i}]/\sigma_{LR} > 0$.

The following theorem shows that when the sample consists of nonnegligible clusters, a necessary condition for the consistent discrimination of the mean is that there exist at least two clusters.

Theorem 3.1. Suppose that $\mathcal{P}_{n,\mathcal{N}} \subset \mathcal{P}_n$ for each $n \geq 1$. Suppose further that $\alpha \in (0, 1/2)$, and $M_n = 1$ for each $n \geq 1$. Then, the mean cannot be consistently discriminated at level α .

The theorem implies that when we do not know the local dependence structure of the random variables (i.e., $M_n = 1$), it is not possible to consistently discriminate the mean.

3.2. Consistent \sqrt{n} -discrimination of the mean

We introduce the following notion of consistent \sqrt{n} -discrimination.

Definition 3.2. The mean of $X_{n,i}$, $n \geq 1$, is **consistently \sqrt{n} -discriminated at level $\alpha \in (0, 1)$ in model \mathcal{P}_n** , if there is a sequence of (potentially randomized) tests $\{\varphi_n\}_{n \geq 1}$ such that

$$\limsup_{n \rightarrow \infty} \mathbf{E}[\varphi_n(X_n)] \leq \alpha,$$

along any sequence $P_n \in \mathcal{P}_{n,0}$, and

$$\liminf_{n \rightarrow \infty} \mathbf{E}[\varphi_n(X_n)] = 1,$$

along any sequence $P_n \in \mathcal{P}_n$ such that $\lim_{n \rightarrow \infty} \sqrt{n} \mathbf{E}[X_{n,i}]/\sigma_{LR} = \infty$.

We consider consistent discrimination against alternatives after normalizing by σ_{LR} (which depends on n), so that when σ_{LR}^2 is larger, we focus on the alternative hypothesis that is farther away from the null hypothesis. Hence, if it is not possible to consistently \sqrt{n} -discriminate the mean, it is not necessarily due to the long run variance increasing to infinity fast.

The consistent \sqrt{n} -discrimination is often obtained when the parameter is in a finite dimensional space, and one knows the local dependence structure of the observations. To illustrate this point, suppose that $X_{n,1}, \dots, X_{n,n}$'s are i.i.d. Then, often we have

$$\frac{\sqrt{n}(\bar{X}_n - \mathbf{E}[X_{n,i}])}{\hat{\sigma}_n} \rightarrow_d \mathcal{N}(0, 1),$$

where $\hat{\sigma}_n^2 \rightarrow_p \sigma^2 = \text{Var}(X_{n,1}) > 0$, and $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_{n,i}$. Let us consider the usual t -test as follows:

$$\varphi_n(X_n) = 1 \left\{ \frac{\sqrt{n}\bar{X}_n}{\hat{\sigma}_n} > z_{1-\alpha} \right\}.$$

Under the Pitman local alternatives such that $\mathbf{E}[X_{n,i}] = \bar{\mu}/\sqrt{n}$, $\bar{\mu} > 0$, we have

$$\liminf_{n \rightarrow \infty} \mathbf{E}[\varphi_n(X_n)] = 1 - \Phi(z_{1-\alpha} - \bar{\mu}),$$

where Φ denotes the CDF of $\mathcal{N}(0, 1)$. The last term converges to 1 as $\bar{\mu} \rightarrow \infty$. Hence, the mean is consistently \sqrt{n} -discriminated. The discrimination results extend to the case with locally dependent observations where we know the local dependence structure and the long run variance is consistently estimable.

However, when we do not know the dependence structure, the consistent \sqrt{n} -discrimination of the mean is not guaranteed. We make this explicit in the following corollary which follows immediately from [Theorem 3.1](#).

Corollary 3.1. Suppose that $\mathcal{P}_{n,\mathcal{N}} \subset \mathcal{P}_n$ for each $n \geq 1$. Suppose further that $\alpha \in (0, 1/2)$, and $M_n = 1$. Then, the consistent \sqrt{n} -discrimination of the mean at level α is not possible.

On the other hand, if we have at least two large clusters and do not know the dependence structure within each cluster, we can consistently \sqrt{n} -discriminate the mean as long as the within-cluster sample means are asymptotically normal, as shown in the following theorem.

Theorem 3.2. Suppose that there exists a sub-partition $\mathcal{M}'_n \subset \mathcal{M}_n$ such that $|\mathcal{M}'_n| \geq 2$ for each $n \geq 1$, and

$$\liminf_{n \rightarrow \infty} \min_{N_{n,m} \in \mathcal{M}'_n} \frac{|N_{n,m}|}{n} > 0.$$

Suppose further that the set \mathcal{M}'_n satisfies that for each $P_n \in \mathcal{P}_n$ and for each $t \in \mathbf{R}$,

$$\max_{1 \leq m \leq M_n: N_{n,m} \in \mathcal{M}'_n} \left| P_n \left\{ \frac{\sqrt{n_m}(\bar{X}_{n,m} - \mathbf{E}[X_{n,i}])}{\sigma_{n,m}} \leq t \right\} - \Phi(t) \right| \rightarrow 0, \tag{3.1}$$

as $n \rightarrow \infty$.

Then, the mean is consistently \sqrt{n} -discriminated at level $\alpha \in (0, 1)$.

For the theorem, we construct a t -test statistic as in [Ibragimov and Müller \(2010\)](#) and show that using the test, we can consistently \sqrt{n} -discriminate the mean, without knowing the dependence structure within the clusters.

The asymptotic normality condition (3.1) is often satisfied if the within-cluster dependence is weak. As we show later, this does not mean that we can consistently estimate $\sigma_{n,m}$ for each cluster. (We will study this problem in the next section in detail.) Also, it is important to note that the within-cluster asymptotic normality is not enough to secure the consistent \sqrt{n} -discrimination of the mean, if there is only one cluster. In fact, the asymptotic normality condition alone does not exclude the possibility of $\mathcal{P}_{n,\mathcal{N}} \subset \mathcal{P}_n$, and in this case, [Corollary 3.1](#) shows that the mean is not consistently \sqrt{n} -discriminated.

As mentioned in the introduction, [Song \(2016\)](#) and [Leung \(2021\)](#) considered the approach of randomized subsampling inference when one does not know the dependence structure at all. Hence, their situation corresponds to the setting with $M_n = 1$. Their procedure requires the following assumption:

$$\sqrt{n} (\bar{X}_n - \mathbf{E}[X_{n,i}]) = O_p(1), \tag{3.2}$$

as $n \rightarrow \infty$. This assumption requires the existence of an upper bound of the long-run variance such that the upper bound does not change with n . In this case, it is not hard to see that we can consistently \sqrt{n} -discriminate the mean as long as the condition (3.1) holds. Indeed, we can consider the test where we reject the null hypothesis of $\mathbf{E}[X_{n,i}] = 0$ against $\mathbf{E}[X_{n,i}] > 0$ if and only if

$$\frac{\sqrt{n}\bar{X}_n}{\sqrt{c_n}} > z_{1-\alpha},$$

where c_n is a slowly increasing sequence such that $\sqrt{n}/c_n \rightarrow \infty$. In our setting of hypothesis testing, however, the set of probabilities \mathcal{P}_n does not have a finite upper bound for the long-run variance of the sample mean, reflecting the fact that such an upper bound is not known in practice. Thus, the assumption (3.2) does not hold uniformly over $P_n \in \mathcal{P}_n$ in our setting, and the results of [Song \(2016\)](#) and [Leung \(2021\)](#) do not contradict the impossibility result of [Theorem 3.1](#).

We would like to emphasize that the impossibility result in [Theorem 3.1](#) stems from the unknown cross-sectional dependence structure of the variables, where the model does not exclude the distributions in the LTIC Gaussian model. For example, [Romano \(2004\)](#) studied the Bahadur–Savage impossibility result ([Bahadur and Savage, 1956](#)), and proved that in the standard hypothesis testing with i.i.d. random variables, $X_i, i = 1, \dots, n$, the mean is consistently \sqrt{n} -discriminated, if the square of the location-scale normalized version of the random variable X_i satisfies the uniform integrability condition. However, our result shows that in a setting where the random variables are cross-sectionally dependent with an unknown dependence structure, the uniform integrability condition for the marginal distribution of X_i imposed on $\mathcal{P}_n, n \geq 1$, does not preclude the impossibility result in [Theorem 3.1](#). Thus, the impossibility of consistent discrimination or \sqrt{n} -discrimination in our paper cannot be simply resolved by imposing the uniform integrability condition on $\mathcal{P}_n, n \geq 1$, as in [Romano \(2004\)](#) for the case with i.i.d. random variables. Furthermore, the result of [Romano \(2004\)](#) does not easily extend to our setting, because, among others, he uses the fact that the long run variance is consistently estimable with i.i.d. random variables. (See the proof of Theorem 4, p.579, in [Romano \(2004\)](#).) As we shall see later, however, consistent estimation of long run variance is not possible in our setting, when the sample contains at least one large cluster.

4. Consistent estimation of variance

Recently, [Hansen and Lee \(2019\)](#) showed that it is necessary and sufficient for the weak law of large numbers to hold for the sample average of the clustered observations that

$$\lim_{n \rightarrow \infty} \sum_{m=1: n_m \geq 2}^{M_n} \left(\frac{n_m}{n} \right)^2 = 0.$$

In this section, we show that this condition is necessary and sufficient for consistent estimability of the long run variance. This implies that when there is a large cluster (i.e., which takes up an asymptotically nonnegligible fraction of the observations), the long run variance is not consistently estimable. This is a consequence of lack of knowledge of the dependence structure within the large cluster. It means that the usual asymptotic inference based on the asymptotic normal approximation of statistics is generally not applicable in this situation.

4.1. Consistent estimability

Let us introduce the notion of consistent estimability of a parameter. Let \mathcal{P}_n be the set of joint distributions of observed random variables, say, $\{X_1, \dots, X_n\}$. Given a parameter space $\Theta \subset \mathbb{R}^d$, we define our object of interest to be a map $\theta_n : \mathcal{P}_n \rightarrow \Theta$.

Definition 4.1. For any sequence of subsets $\mathcal{P}'_n \subset \mathcal{P}_n$, we say that θ_n is **consistently estimable in \mathcal{P}'_n** , if there exists an estimator $\hat{\theta}$ such that along any sequence $P_n \in \mathcal{P}'_n$,

$$P_n \{ \|\hat{\theta} - \theta_n(P_n)\| > \epsilon \} \rightarrow 0,$$

as $n \rightarrow \infty$, for each $\epsilon > 0$.

One can find a similar definition of consistent estimability in [LeCam and Schwartz \(1960\)](#). They provide necessary and sufficient conditions for a parameter to be consistently estimable when the data are i.i.d. See also Section 1.4 of [Ibragimov and Has'minskii \(1981\)](#) and Section 6.2 of [Pfanzagl \(1994\)](#).

Our setting is somewhat nonstandard, requiring a different technique to prove impossibility of consistent estimation. It is usually assumed that the probability model is indexed by a certain set, i.e., $\mathcal{P}_n = \{P_{n,h} : h \in \mathcal{H}\}$, where each $P_{n,h}$ is a probability measure indexed by h in some topological space \mathcal{H} that is independent of the sample size n . One can then redefine the parameter $\psi_n(h) = \theta_n(P_{n,h})$, $h \in \mathcal{H}$, i.e., as a map on \mathcal{H} . As long as ψ_n behaves ‘‘continuously’’ on \mathcal{H} , the parameter ψ_n can be shown to be consistently estimable. (See, e.g., Theorem 4.1 of [Ibragimov and Has'minskii \(1981\)](#) and Theorem 6.2.11 of [Pfanzagl \(1994\)](#).) Then the impossibility of consistent estimation stems from the discontinuity of ψ_n as a map on \mathcal{H} , which yields ‘‘non-identifiability’’ of the parameter ([Pötscher, 2002](#)).

However, we cannot apply this standard approach in our setting, because there is no natural space \mathcal{H} that indexes \mathcal{P}_n independently of n . The main reason is that we need to deal with a situation potentially with a large cluster with an unknown within-cluster dependence structure. This means that we need to require our probability model to accommodate a wide range of dependence structures for the entire sample. For example, suppose that there is only one large cluster, so that one does not know the dependence structure at all. This means, among other things, that our model needs to include various network dependence structures (such as those studied in [Kojevnikov et al., 2021](#)) for the joint distribution of the entire random vector $[X_1, \dots, X_n]$ whose dimension grows with the sample size n . One might consider parameterizing the probabilities in terms of the networks governing the dependence structure, but each network depends on the sample size n . To the best of our knowledge, there is no obvious way to topologize such a probability model and to define the continuity of the parameter θ_n on the probabilities, independently of sample size n .

Our approach relies on the following simple lemma that uses contiguity of probabilities at a primitive level. For any two sequences of probabilities P_n and P'_n , we say that P_n is **contiguous with respect to P'_n** if $P'_n(A_n) \rightarrow 0$ implies $P_n(A_n) \rightarrow 0$ for any sequence of Borel sets A_n , and write $P_n \triangleleft P'_n$. When $P_n \triangleleft P'_n$ and $P'_n \triangleleft P_n$, we say that P_n and P'_n are **mutually contiguous**, and write $P_n \triangleleft\triangleright P'_n$. Contiguity between probabilities was introduced by [LeCam \(1960\)](#) and is widely used, especially for deriving the limiting distribution of a test statistic under local alternatives. By tracing out the limiting distribution along a range of local alternatives, one obtains a limiting experiment which one can use to compute the asymptotic risk lower bound in statistical decision theory. (See, e.g., Chapter 6 of [van der Vaart \(1998\)](#).)

The following lemma summarizes our scheme of proving the impossibility of consistent estimability of σ^2_{LR} .

Lemma 4.1. Suppose that there exists a sequence $P_{n,0} \in \mathcal{P}_n$ such that $P_{n,1} \triangleleft P_{n,0}$ for every sequence $P_{n,1} \in \mathcal{P}_n$. Then, θ_n is consistently estimable in \mathcal{P}_n if and only if $\lim_{n \rightarrow \infty} \|\theta_n(P_{n,1}) - \theta_n(P_{n,0})\| = 0$ for any sequence $P_{n,1} \in \mathcal{P}_n$.

Later we use [Lemma 4.1](#) to prove the impossibility of consistent estimation of the long run variance, by selecting two Gaussian probabilities, $P_{n,0}$ and $P_{n,1}$, such that $P_{n,1} \triangleleft P_{n,0}$ and the values of the long run variance stay apart under $P_{n,0}$ and $P_{n,1}$ as $n \rightarrow \infty$. (See the discussion below [Theorem 4.1](#).)

The notion of consistent estimability in [Definition 4.1](#) coincides with consistent estimability *uniform in P* , i.e., the existence of an estimator $\hat{\theta}$ such that for each $\epsilon > 0$, as $n \rightarrow \infty$,

$$\sup_{P_n \in \mathcal{P}_n} P_n \{ \|\hat{\theta} - \theta_n(P_n)\| > \epsilon \} \rightarrow 0.$$

(See [Ibragimov and Has'minskii \(1981\)](#), p.31. See also [Pötscher \(2002\)](#) for discussion on asymptotics uniform in P .) When $\mathcal{P}_n = \{P_{n,h} : h \in \mathcal{H}\}$ for some index set \mathcal{H} which does not depend on the sample size n , uniform consistent estimability is stronger than pointwise consistent estimability which assumes the existence of an estimator $\hat{\theta}$ such that for each $h \in \mathcal{H}$ and for each $\epsilon > 0$,

$$P_{n,h} \{ \|\hat{\theta} - \theta_n(P_{n,h})\| > \epsilon \} \rightarrow 0,$$

as $n \rightarrow \infty$. However, as explained above, in our setting, there is no space \mathcal{H} that indexes \mathcal{P}_n and is independent of n . Hence, there is no natural notion of pointwise consistency in P in our set-up.

4.2. Consistent estimability of variance in Gaussian experiments

4.2.1. A necessary and sufficient condition for consistent estimability of variance

In our context, a major challenge is to show the contiguity condition (ii) of Lemma 4.1. A standard argument proving contiguity utilizes the local asymptotic normality or local asymptotic mixed normality results for a log-likelihood process. However, these latter results often use an i.i.d. or time-series set-up where the researcher knows the dependence structure, and hence, are not useful for our purpose here. For this reason, we focus on a Gaussian experiment, where we can explicitly compute the log-likelihood process in finite samples and investigate its asymptotic behavior as the dependence structure varies. In particular, we consider the following model for a fixed $\sigma^2 > 0$,

$$\mathcal{P}_{n,\mathcal{N}}(\sigma^2) = \{ \Phi(\bar{\mu}\mathbf{1}_n, \Sigma_n(\sigma^2, \delta)) : \delta \in (0, 1] \text{ and } \bar{\mu} \in \mathbf{R} \},$$

where $\mathbf{1}_n$ denotes the n -dimensional vector of ones. The set $\mathcal{P}_{n,\mathcal{N}}(\sigma^2)$ represents the set of LTIC Gaussian models, where each multivariate normal distributions with a common mean and the short run variance equal to σ^2 .

For the impossibility result below, we require that the probability model does not exclude this Gaussian experiment.

Theorem 4.1. *Suppose that $\mathcal{P}_{n,\mathcal{N}}(\sigma^2) \subset \mathcal{P}_n$ for each $n \geq 1$, for some $\sigma^2 > 0$ that is independent of n . Then, the long run variance σ_{LR}^2 is consistently estimable in \mathcal{P}_n if and only if*

$$\lim_{n \rightarrow \infty} \sum_{m=1}^{M_n} \frac{n_m^2}{n^2} \rightarrow 0, \tag{4.1}$$

as $n \rightarrow \infty$.

The sufficiency part of the theorem is straightforward. To see this, suppose for simplicity that there is no singleton cluster in the data. If (4.1) is satisfied, it means that the number of clusters M_n grows to infinity as $n \rightarrow \infty$. Then, we consider the following estimator:

$$\hat{\sigma}_{LR}^2 \equiv \frac{1}{n} \sum_{m=1}^{M_n} \sum_{i,j \in N_{n,m}} (X_{n,i} - \bar{X}_n)(X_{n,j} - \bar{X}_n).$$

In fact, for the sufficiency part, we do not require that $\mathcal{P}_{n,\mathcal{N}}(\sigma^2) \subset \mathcal{P}_n$.

The nontrivial part of the theorem is to show that the condition (4.1) is necessary for the consistent estimability of σ_{LR}^2 in $\mathcal{P}_{n,\mathcal{N}}$. Suppose that the condition of (4.1) fails, which implies that one has at least one nonnegligible cluster. Then, we show that σ_{LR}^2 is not consistently estimable. For this, we employ Lemma 4.1 after computing the log-likelihood process under cluster dependence. More specifically, suppose first that the observations consist of only a single large cluster. Then, we note that the model \mathcal{P}_n , due to the lack of knowledge on the dependence structure, does not exclude the LTIC Gaussian experiment: $\Phi(0, \Sigma_n(\sigma^2, \delta))$. Then we show that

$$\Phi(0, \Sigma_n(\sigma^2, 0)) \triangleleft \Phi(0, \Sigma_n(\sigma^2, \delta)),$$

whereas

$$\sigma_{LR}^2(\Phi(0, \Sigma_n(\sigma^2, \delta))) - \sigma_{LR}^2(\Phi(0, \Sigma_n(\sigma^2, 0))) \rightarrow c \neq 0,$$

as $n \rightarrow \infty$, for some nonzero constant c . Hence, by Lemma 4.1, σ_{LR}^2 cannot be consistently estimated in any probability model that does not exclude the LTIC Gaussian experiment. It is not hard to extend the same arguments to a setting where there are potentially multiple large clusters.

Theorem 4.1 then implies that if a nonnegligible portion of the sample belongs to non-singleton clusters, the long run variance σ_{LR}^2 is consistently estimable in $\mathcal{P}_{n,\mathcal{N}}$ if and only if the cluster structure consists of negligible clusters. We formalize this in the following corollary.

Corollary 4.1. *Suppose that the conditions of Theorem 4.1 hold, and $\liminf_{n \rightarrow \infty} n^*/n > 0$, where*

$$n^* = \sum_{m=1: n_m \geq 2}^{M_n} n_m.$$

Then, the long run variance σ_{LR}^2 is consistently estimable in \mathcal{P}_n if and only if the cluster structure \mathcal{M}_n consists of negligible clusters, i.e.,

$$\lim_{n \rightarrow \infty} \max_{1 \leq m \leq M_n} \frac{n_m}{n} = 0. \tag{4.2}$$

The condition $\liminf_{n \rightarrow \infty} n^*/n > 0$ requires that the fraction of random variables $X_{n,i}$ that do not belong to a singleton cluster is asymptotically nonnegligible. In this case, if the probability model in practice includes the Gaussian model $\mathcal{P}_{n,\mathcal{N}}$ as a subclass and there is at least one nonnegligible cluster, it is not possible to consistently estimate the long run variance. Certainly, this impossibility result carries over to a model where the long run variance σ_{LR}^2 is allowed to increase with the sample size n .

Hence, by combining Theorem 3.2 with Theorem 4.1, we find that when we have several large clusters, the long run variance is not consistently estimable because the sample contains large clusters, but the mean can still be consistently \sqrt{n} -discriminated.

4.2.2. Implications for network dependent observations

One might wonder whether the result of Theorem 4.1 extends to the case where the observations exhibit a dependence structure other than cluster dependence. Below we give a partial answer for the case of a dependency graph. Dependency graphs were introduced by Stein (1986), and have been studied and used in statistics and econometrics. (See, e.g., Aronow and Samii (2017), Song (2018), Leung (2020) and Canen et al. (2020) and references therein.)

A graph (or network) is a pair $G_n = (N_n, E_n)$, where $N_n = \{1, \dots, n\}$ denotes the set of vertices and E_n the set of edges, where we denote $N(i) = \{j : ij \in E_n\}$ to mean the neighborhood of vertex i . (Here, we consider only simple, undirected graphs, i.e., $ii \notin E_n$ for all $i \in N_n$, and $ij \in E_n$ if and only if $ji \in E_n$.) We define

$$d_{\max} = \max_{i \in N_n} |N(i)| \quad \text{and} \quad d_{av} = \frac{1}{n} \sum_{i \in N_n} |N(i)|,$$

where d_{\max} is called the **maximum degree**, and d_{av} the **average degree** of the graph G_n . The maximum and average degrees are often used to capture the denseness of the graph. A subset of vertices in graph G_n is called a **clique** if any two distinct vertices in the subset are adjacent in G_n , and the number of vertices in the clique is called the size of the clique. The maximum clique size refers to the size of the clique that is largest in the graph G_n .

Recall that a graph $G_n = (N_n, E_n)$ on $N_n = \{1, \dots, n\}$ is called a **dependency graph** for $X_n = (X_{n,i})_{i \in N_n}$, if for any subset $A \subset N_n$, $(X_{n,i})_{i \in A}$ and $(X_{n,i})_{i \in N_n \setminus \overline{N_n}(A)}$ is independent, where $\overline{N_n}(A) = \{j : ij \in E_n \text{ for some } i \in A\} \cup \{i\}$. It is important to note that while the dependency graph imposes independence between $X_{n,i}$ and $X_{n,j}$ when they are not adjacent in the graph, it says nothing about dependence between them when they are adjacent. Thus, we allow in \mathcal{P}_n any degree of dependence (including independence) between $X_{n,i}$ and $X_{n,j}$ whenever i and j are adjacent in G_n . As the dependency graph becomes denser, this reflects our limited knowledge on the dependence structure, similarly to large clusters in the cluster dependence case.

Corollary 4.2. *Suppose that the conditions of Theorem 4.1 hold and that for each $n \geq 1$, there exists a graph $G_n = (N_n, E_n)$ which has maximum degree d_{\max} , average degree d_{av} , maximum clique size n_C , and each distribution $P_n \in \mathcal{P}_n$ of X_n has G_n as a dependency graph. Then, the following holds.*

- (i) If $\limsup_{n \rightarrow \infty} n_C/n > 0$, the long run variance σ_{LR}^2 is not consistently estimable.
- (ii) If $\lim_{n \rightarrow \infty} d_{\max}^2 d_{av}/n = 0$, the long run variance σ_{LR}^2 is consistently estimable.

The impossibility result in (i) has an important implication in many models with network dependent observations. As in the case of a dependency graph, many models of network dependence do not specify the strength of dependence between observations that are adjacent in the network (e.g., Kojevnikov et al., 2021). Weak dependence is usually imposed between observations that are far from each other in terms of the shortest path in the network. Hence, when there is a large clique in the network which constitutes a nonnegligible fraction of the entire sample in the limit as $n \rightarrow \infty$, Corollary 4.2(i) implies that the long run variance of the network dependent observations is not consistently estimable.

It is interesting to note that one cannot characterize a necessary and sufficient condition for the network solely in terms of its maximum degree. For example, if X_n is a multivariate normal random vector such that each component has a bounded variance and has a star graph as a dependency graph, the long run variance is consistently estimable. To see this, let $X_n = [X_{n,1}, \dots, X_{n,n}]^T$ be a centered multivariate random vector which has a graph G_n as a dependency graph. Let the graph G_n be a star graph with the unit 1 being its center.² In the context of multivariate normality, we can write

$$X_{n,1} = \sum_{i \in N_n \setminus \{1\}} \theta_i X_{n,i} + \varepsilon,$$

where the leading sum is the best linear projection, so that ε is independent of $X_{n,i}$'s, $i \in N_n \setminus \{1\}$, which are independent from each other (due to the dependency graph being a star graph). Since the variance of $X_{n,1}$ is bounded, we should have $\sum_{i \in N_n \setminus \{1\}} \theta_i^2 < C$, for all $n \geq 1$, for some $C > 0$. Note that we can identify

$$\theta_i = \text{Cov}(X_{n,i}, X_{n,1}) / \text{Var}(X_{n,i}).$$

Now, we can write

$$\sigma_{LR}^2 = \frac{1}{n} \sum_{i=1}^n \mathbf{E}[X_{n,i}^2] + \frac{1}{n} \sum_{i=1}^n \sum_{j=1, i \neq j}^n \mathbf{E}[X_{n,i} X_{n,j}].$$

The second term is written as

$$\frac{2}{n} \sum_{i=2}^n \mathbf{E}[X_{n,i} X_{n,1}] = 2\mathbf{E} \left[\left(\frac{1}{n} \sum_{i=2}^n X_{n,i} \right) X_{n,1} \right] = o(1),$$

² The star graph as a dependency graph is different from an additive common shock model such as $X_{n,i} = C_n + \varepsilon_i$, where C_n is a common shock, and ε_i 's are cross-sectionally independent idiosyncratic shock. In this case, the dependency graph is a *complete graph*, because every pair of random variables is correlated through the common shock. Hence, the center in the star graph as a dependency graph cannot be a source like a common shock. It is more plausible to imagine the center to be an aggregated outcome of independent sources. In this case, by simply eliminating the star, one obtains independent random variables.

as $n \rightarrow \infty$, because the normalized sum in the parenthesis converges to zero in moments. Hence, we can simply take

$$\hat{\sigma}_{LR}^2 = \frac{1}{n} \sum_{i=1}^n X_{n,i}^2$$

to be an estimator of the long run variance. It is not hard to see that $\hat{\sigma}_{LR}^2$ is consistent for σ_{LR}^2 . This example shows that one cannot express the condition for the consistent estimability solely in terms of the maximum degree of the dependency graph.

5. Implications

5.1. A linear regression model with cluster-dependent errors

Let us consider the following regression model with cluster-dependent errors (see, e.g., Cameron et al. (2008), Djogbenou et al. (2019) and Hansen and Lee (2019) and references therein):

$$y = X\beta + u,$$

where $y = [y_1^\top, \dots, y_{M_n}^\top]^\top$, $X = [X_1^\top, \dots, X_{M_n}^\top]^\top$, and $u = [u_1^\top, \dots, u_{M_n}^\top]^\top$, with $E[u_m | X] = 0$ for each $m = 1, \dots, M_n$, and each cluster m has n_m observations (so that y_m and u_m are n_m dimensional column vectors, and X_m is an $n_m \times k$ matrix.) We assume that u_1, \dots, u_{M_n} are independent, but for each m , the dependence structure of u_m is not known. We do not exclude the possibility that the error term follows a normal distribution.

Then, the OLS estimator of β is given by

$$\hat{\beta} = (X^\top X)^{-1} X^\top y.$$

The sandwich form of the variance matrix of $\hat{\beta}$ is given by

$$V = (X^\top X)^{-1} \left(\sum_{m=1}^{M_n} X_m^\top E[u_m u_m^\top | X] X_m \right) (X^\top X)^{-1}.$$

Once we obtain a consistent estimator \hat{V} of V , we can construct a standard error of the j th entry of β , i.e., β_j , as $\hat{\sigma}_j^2 = [\hat{V}]_{jj}$, the j th diagonal of \hat{V} . From the asymptotic normal inference applied to a t -statistic for β_j , we obtain the following confidence interval for β_j :

$$\left[\hat{\beta}_j - \frac{z_{\alpha/2} \hat{\sigma}_j}{\sqrt{n}}, \hat{\beta}_j + \frac{z_{\alpha/2} \hat{\sigma}_j}{\sqrt{n}} \right]. \tag{5.1}$$

As for the consistent estimator \hat{V} , Djogbenou et al. (2019) considered the following estimator:

$$\hat{V} = d(X^\top X)^{-1} \left(\sum_{m=1}^{M_n} X_m^\top \hat{u}_m \hat{u}_m^\top X_m \right) (X^\top X)^{-1},$$

where $\hat{u}_m = y_m - X_m^\top \hat{\beta}$, and d is a sequence such that $d \rightarrow 1$. They established the consistency of this estimator under a set of conditions, and showed that their conditions are not compatible with a setting in which one of the clusters is large, i.e., its size is proportional to the entire sample.

Our result implies that such an estimator \hat{V} is not uniformly consistent when there is at least one large cluster. In fact, our result is much stronger than this. It shows that it is not possible to construct a uniformly consistent estimator of V in such a case. Hence, in this case, we cannot construct a confidence interval of the form (5.1) that is uniformly asymptotically valid. When a nonnegligible fraction of observations belong to a (non-singleton) cluster — which is the case with most cluster-dependence settings, the necessary and sufficient condition for the uniformly consistent estimability of V is that each cluster is asymptotically negligible in the sense of (4.2).

5.2. Difference-in-differences with spillovers

Let us explore the implications of the impossibility results in the context of a difference-in-differences approach to causal inference. (See Section 6.5 of Imbens and Wooldridge (2009) for an overview of this method. See also Roth et al. (2022) for an overview including recent advances in the literature.) Suppose that there are n individuals who are subject to a treatment and the researcher observes their outcomes before and after the treatment. We let $Y_{i,t}(1)$ and $Y_{i,t}(0)$ denote the potential outcomes at time $t = 0, 1$ for the treated state and the control state, respectively. As standard in the literature, we assume that in time 0, no individual is treated, and $Y_{i,0} = Y_{i,0}(0)$, which is observed. The observed outcome $Y_{i,1}$ at time 1 is defined by

$$Y_{i,1} = D_i Y_{i,1}(1) + (1 - D_i) Y_{i,1}(0),$$

where D_i is the indicator of treatment for i that happens between times 0 and 1. Our parameter of interest is the average treatment effect on the treated:

$$ATT = E[Y_{i,1}(1) - Y_{i,1}(0) | D_i = 1].$$

Suppose that we have observations $\{(Y_{i,1}, Y_{i,0}, D_i)\}_{i=1}^n$, where $Y_{i,0}$ is the outcome for person i at time 0. Furthermore, we assume that the researcher knows the probability $p = P\{D_i = 1\}$. (The impossibility result we mention below carries over to the case where p is not known.)

Let us introduce the standard parallel trend assumption used in the literature:

$$E[Y_{i,1}(1) - Y_{i,1}(0) | D_i = 1] = E[Y_{i,1}(1) - Y_{i,1}(0) | D_i = 0].$$

Under this assumption, we can identify

$$ATT = E[\Delta Y_i | D_i = 1] - E[\Delta Y_i | D_i = 0],$$

where $\Delta Y_i = Y_{i,1} - Y_{i,0}$. We can obtain a sample analog estimator by

$$\widehat{ATT} = \frac{1}{n} \sum_{i=1}^n \frac{D_i \Delta Y_i}{p} - \frac{1}{n} \sum_{i=1}^n \frac{(1 - D_i) \Delta Y_i}{1 - p}.$$

We consider settings where the observed outcomes are cross-sectionally dependent. Our interest is in constructing a confidence interval for ATT that is uniformly asymptotically valid. Below we consider two situations, one with treatment spillover and the other with spillover of treatment effects. We explore implications of our impossibility results in these situations.

5.2.1. Treatment spillover

Suppose that there is a spillover of the treatments so that D_i 's are correlated across i , along some network among people. For example, one can think of a situation in a social program where two people i and j are neighbors and participating in the program by i can induce the participation by j . Suppose that the researcher does not have information on the neighborhoods of the subjects. Then, this creates dependence among $Y_{i,1}$'s along a dependence structure that is unknown to the researcher. Then, our impossibility result shows that ATT cannot be consistently discriminated.

In practice, the treatment assignment is often done at the cluster level, where the potential outcomes $Y_{i,1}(1)$ and $Y_{i,1}(0)$ may exhibit arbitrary dependence within each cluster. (See Section 5 of Roth et al. (2022) for examples and references studying such a setting.)

Suppose that we have at least two large clusters such that $(Y_{i,1}(1), Y_{i,0}(0), D_i)$ are independent across the clusters but arbitrarily correlated within each cluster. The researcher might attempt to test the null hypothesis of $ATT = 0$ by considering the usual t statistic for testing the null hypothesis of $ATT = 0$ such that

$$t = \frac{\sqrt{n}(\widehat{ATT} - ATT)}{\hat{\sigma}},$$

where $\hat{\sigma}^2$ is a consistent estimator of the variance of $\sqrt{n}(\widehat{ATT} - ATT)$, and the critical values are taken from the standard normal table. Our impossibility result implies that it is not possible to consistently estimate the variance of $\sqrt{n}(\widehat{ATT} - ATT)$, when there is at least one large cluster, and hence, such a t -test is not uniformly asymptotically valid. For the same reason, we cannot construct a confidence interval of the following familiar form:

$$\left[\widehat{ATT} - \frac{z_{\alpha/2} \hat{\sigma}}{\sqrt{n}}, \widehat{ATT} + \frac{z_{\alpha/2} \hat{\sigma}}{\sqrt{n}} \right] \tag{5.2}$$

such that the confidence interval is uniformly asymptotically valid. (See Section 5 of Roth et al. (2022) for various approaches.³)

5.2.2. Spillover of treatment effects

Suppose that the treatments D_i themselves do not exhibit any spillover, but the cross-sectional dependence of $(Y_{i,1}(1), Y_{i,0}(0))$ arises due to the spillover of the treatment effects, for example, the treatment of a person i influences the outcome of the person j in the next period. Such a setting has been studied in the recent literature (see Aronow and Samii (2017), Leung (2020) and He and Song (2022) and references therein.)

Suppose that the spillover of the treatment effects arises along some network among people, and yet the researcher does not have any information on the network. Then, our impossibility result implies that we cannot consistently discriminate ATT in such a situation.

Suppose that the researcher observes a group structure where the spillover does not arise between groups, so that $(Y_{i,1}(1), Y_{i,0}(0), D_i)$ are independent across groups. If each within-group sum of $(Y_{i,1}(1), Y_{i,0}(0), D_i)$ satisfies the central limit theorem, our result shows that the ATT can be consistently \sqrt{n} -discriminated. However, when there is at least one large group, there does not exist a consistent estimator of the variance of $\sqrt{n}(\widehat{ATT} - ATT)$. Hence, similarly as before, we cannot construct a uniformly asymptotically valid t -test for the null hypothesis of $ATT = 0$ using the usual t statistic and standard normal critical values, and cannot construct a uniformly asymptotically valid confidence interval of the form (5.2) based on a normal approximation.

³ To the best of our knowledge, there is no formal result that proposes a uniformly asymptotically valid confidence interval for ATT in this setting. However, we expect that the bootstrap approach of Canay et al. (2021) can be used to construct a uniformly valid confidence interval under mild additional conditions.

6. Conclusion

In this paper, we show two impossibility results on the inference on the mean when the random variables follow a cluster dependence structure. The first result is the impossibility of consistent discrimination of the mean when there is only one cluster. However, when there are at least two nonnegligible clusters and the asymptotic normal approximation of the location-scale normalized sample mean is possible, consistent \sqrt{n} -discrimination of the mean is restored. The second result is the impossibility of the consistent estimation of the long-run variance, when there is at least one large cluster. This result shows that the usual asymptotic inference method based on a t -statistic is not applicable in such a setting.

Our impossibility result demonstrates that the researcher’s knowledge of the cross-sectional dependence structure plays a crucial role when performing asymptotic inference using cross-sectionally dependent data.

Appendix. Mathematical proofs

A.1. Preliminary results

For the proof of the main results, we first prove auxiliary lemmas. As a first step, we provide an explicit form of a log-likelihood process in Gaussian experiments in Lemma A.2. For this, we use the following auxiliary lemma.

Lemma A.1. *Let $\Sigma_0 = USU^T$ be the spectral decomposition of an $n \times n$, symmetric positive definite matrix Σ_0 and let Σ_1 be an $n \times n$ matrix defined as*

$$\Sigma_1 = \Sigma_0 + UAU^T$$

for some symmetric positive semidefinite matrix A . Let $B\Lambda B^T$ be the spectral decomposition of $S^{-1/2}AS^{-1/2}$. Suppose that $|\lambda_i| < 1$ for all $i = 1, \dots, n$, where λ_i denote the i -th diagonal entry of Λ .

Then the following results hold.

(i)

$$\log(|\Sigma_1|^{-1/2}) - \log(|\Sigma_0|^{-1/2}) = -\frac{1}{2} \sum_{i=1}^n \log(1 + \lambda_i).$$

(ii) For any vectors $a, b \in \mathbf{R}^n$,

$$a^T(\Sigma_1^{-1} - \Sigma_0^{-1})b = -\sum_{i=1}^n \frac{\lambda_i}{1 + \lambda_i} \tilde{a}_i \tilde{b}_i,$$

where \tilde{a}_i and \tilde{b}_i are the i -th entries of \tilde{a} and \tilde{b} , with

$$\tilde{a} = B^T S^{-1/2} U^T a \quad \text{and} \quad \tilde{b} = B^T S^{-1/2} U^T b.$$

Proof. Let $Q = US^{1/2}$ such that $\Sigma_0 = QBB^TQ^T$ and $\Sigma_{n,1} = QB(I + \Lambda)B^TQ^T$. Thus,

$$|\Sigma_1| = |\Sigma_0| \cdot |I + \Lambda| = |\Sigma_0| \cdot \prod_{i=1}^n (1 + \lambda_i),$$

and

$$\Sigma_0^{-1} - \Sigma_1^{-1} = (B^T Q^T)^{-1} (I - (I + \Lambda)^{-1}) (QB)^{-1} = US^{-1/2} B \tilde{\Lambda} B^T S^{-1/2} U^T,$$

where

$$\tilde{\Lambda} = \text{diag} \left(\frac{\lambda_1}{1 + \lambda_1}, \dots, \frac{\lambda_n}{1 + \lambda_n} \right). \quad \blacksquare$$

The following lemma provides an explicit form of a general log-likelihood process for Gaussian measures. Recall that $\Phi(\mu, \Sigma)$ denotes the multivariate normal distribution with mean vector μ and covariance matrix Σ .

Lemma A.2. *Let $\Sigma_0, \Sigma_1, \Lambda, B, S$ and U be the matrices in Lemma A.1. Then, for all $x, \mu_1, \mu_0 \in \mathbf{R}^n$,*

$$\log \frac{d\Phi(\mu_1, \Sigma_1)}{d\Phi(\mu_0, \Sigma_0)}(x) = -\sum_{i=1}^n \log q_i + \frac{1}{2} \sum_{i=1}^n \frac{1}{q_i^2} (Z_i(x)(q_i + 1) - \tilde{\mu}_i) (Z_i(x)(q_i - 1) + \tilde{\mu}_i),$$

where $q_i = \sqrt{1 + \lambda_i}$, λ_i is the i -th diagonal entry of Λ , $Z_i(x)$ is the i -th entry of $Z(x)$ and $\tilde{\mu}_i$ is the i -th entry of $\tilde{\mu}$ with

$$Z(x) = B^T S^{-1/2} U^T (x - \mu_0) \quad \text{and} \quad \tilde{\mu} = B^T S^{-1/2} U^T (\mu_1 - \mu_0).$$

Proof. We write

$$\begin{aligned} \log \frac{d\Phi(\mu_1, \Sigma_1)}{d\Phi(\mu_0, \Sigma_0)}(x) &= \log (|\Sigma_1|^{-1/2}) - \log (|\Sigma_0|^{-1/2}) \\ &\quad - \frac{1}{2} \left((x - \mu_1)^\top \Sigma_1^{-1} (x - \mu_1) - (x - \mu_0)^\top \Sigma_0^{-1} (x - \mu_0) \right). \end{aligned} \tag{A.1}$$

We apply Lemma A.1(i) to the first term on the right hand side. As for the last term we let $\mu_\Delta = \mu_1 - \mu_0$, and $x_* = x - \mu_0$. Note that

$$\begin{aligned} \mu_\Delta^\top \Sigma_0^{-1} \mu_\Delta &= \mu_\Delta^\top U S^{-1} U \mu_\Delta = \mu_\Delta^\top U S^{-1/2} S^{-1/2} U^\top \mu_\Delta \\ &= \mu_\Delta^\top U S^{-1/2} B B^\top S^{-1/2} U^\top \mu_\Delta = \tilde{\mu}^\top \tilde{\mu}. \end{aligned}$$

Similarly, $\mu_\Delta^\top \Sigma_0^{-1} x_* = \tilde{\mu}^\top Z(x)$. We rewrite the last term in (A.1) as

$$\begin{aligned} & - \frac{1}{2} x_*^\top (\Sigma_1^{-1} - \Sigma_0^{-1}) x_* - \frac{1}{2} (\mu_\Delta^\top (\Sigma_1^{-1} - \Sigma_0^{-1}) \mu_\Delta - 2 \mu_\Delta^\top (\Sigma_1^{-1} - \Sigma_0^{-1}) x_*) \\ & \quad - \frac{1}{2} (\mu_\Delta^\top \Sigma_0^{-1} \mu_\Delta - 2 \mu_\Delta^\top \Sigma_0^{-1} x_*) \\ &= \frac{1}{2} \sum_{i=1}^n \frac{\lambda_i}{1 + \lambda_i} Z_i^2(x) + \frac{1}{2} \sum_{i=1}^n \frac{\lambda_i}{1 + \lambda_i} (\tilde{\mu}_i^2 - 2 \tilde{\mu}_i Z_i(x)) - \frac{1}{2} \sum_{i=1}^n (\tilde{\mu}_i^2 - 2 \tilde{\mu}_i Z_i(x)) \\ &= \frac{1}{2} \sum_{i=1}^n \left(\left(\frac{\lambda_i}{1 + \lambda_i} - 1 \right) (Z_i(x) - \tilde{\mu}_i)^2 + Z_i^2(x) \right), \end{aligned}$$

(by applying Lemma A.1(ii)). By rearranging terms, we rewrite the last sum as

$$\frac{1}{2} \sum_{i=1}^n \frac{1}{q_i^2} (Z_i(x)(q_i + 1) - \tilde{\mu}_i) (Z_i(x)(q_i - 1) + \tilde{\mu}_i).$$

Combining this with an earlier result, we obtain the desired result. ■

Lemma A.3. Let $\Sigma_n = \sigma^2 ((1 - \delta)I_n + \delta \mathbf{1}_n \mathbf{1}_n^\top)$, where δ is such that $n\delta \in (-1, 1)$. Then, for any $\tilde{\mu} \in \mathbf{R}$ and $x \in \mathbf{R}^n$, we have

$$\begin{aligned} \log \frac{d\Phi(\tilde{\mu} \mathbf{1}_n, \Sigma_n)}{d\Phi(0, I_n)}(x) &= -\log \sqrt{\sigma^2(1 + (n - 1)\delta)} - (n - 1) \log \sqrt{\sigma^2(1 - \delta)} \\ &\quad + \frac{\sigma^2(1 + (n - 1)\delta) - 1}{2\sigma^2(1 + (n - 1)\delta)} Z_1^2(x) + \frac{\sigma^2(1 - \delta) - 1}{2\sigma^2(1 - \delta)} \sum_{k=2}^n Z_k^2(x) \\ &\quad + \frac{\sqrt{n\tilde{\mu}}}{\sigma^2(1 + (n - 1)\delta)} Z_1(x) - \frac{n\tilde{\mu}^2}{2\sigma^2(1 + (n - 1)\delta)}, \end{aligned}$$

where $Z_k(x)$ is the k th entry of $Z(x) \equiv B^\top x$, and $B = [b_1, \dots, b_n]$ is an $n \times n$ orthogonal matrix such that $b_1 = n^{-1/2} \mathbf{1}$, $b_k^\top \mathbf{1} = 0$ for all $k = 2, \dots, n$, $b_k^\top b_\ell = 0$ for all $k \neq \ell = 2, \dots, n$.

Proof. We apply Lemma A.2 with $S = U = I_n$,

$$A = (\sigma^2(1 - \delta) - 1) I_n + \sigma^2 \delta \mathbf{1}_n \mathbf{1}_n^\top.$$

Note that the spectral decomposition of A is given by $B \Lambda B^\top$, where Λ is the diagonal matrix with the diagonal elements $\lambda_1, \dots, \lambda_n$ given as $\lambda_1 = (\sigma^2 - 1) + \sigma^2(n - 1)\delta$, $\lambda_2 = \dots = \lambda_n = -\sigma^2\delta$, and the orthogonal matrix B as given the lemma. The desired result follows from Lemma A.2. ■

Lemma A.3 yields the following result for the case with cluster dependence. From here on, we make the dimension of the matrices and vectors explicit. Let I_{n_m} be the n_m -dimensional identity matrix and $\mathbf{1}_{n_m}$ denote the n_m -dimensional column vector of ones.

Corollary A.1. Let Σ_n be the block diagonal matrix whose m -th block, $m = 1, \dots, M_n$, is given by

$$\Sigma_{n,m} = \sigma^2 \left((1 - \delta_{n,m}) I_{n_m} + \delta_{n,m} \mathbf{1}_{n_m} \mathbf{1}_{n_m}^\top \right),$$

for some $\delta_{n,m} \in \mathbf{R}$ such that $n_m \delta_{n,m} \in (-1, 1)$.

Then, for any $\tilde{\mu}_n \in \mathbf{R}$, and $x = [x_1, \dots, x_n]^\top \in \mathbf{R}^n$,

$$\begin{aligned} \log \frac{d\Phi(\tilde{\mu}_n \mathbf{1}_n, \Sigma_n)}{d\Phi(0, I_n)}(x) &= - \sum_{m=1}^{M_n} \log \sqrt{\sigma^2(1 + (n_m - 1)\delta_{n,m})} - \sum_{m=1}^{M_n} (n_m - 1) \log \sqrt{\sigma^2(1 - \delta_{n,m})} \\ &\quad + \sum_{m=1}^{M_n} \frac{\sigma^2(1 + (n_m - 1)\delta_{n,m}) - 1}{2\sigma^2(1 + (n_m - 1)\delta_{n,m})} Z_{m,i_m}^2(x) \end{aligned}$$

$$\begin{aligned}
 & + \sum_{m=1}^{M_n} \frac{\sigma^2(1 - \delta_{n,m}) - 1}{2\sigma^2(1 - \delta_{n,m})} \sum_{i \in N_{n,m} \setminus \{i_m\}} Z_{m,i}^2(x) \\
 & + \sum_{m=1}^{M_n} \frac{n\bar{\mu}_n/\sqrt{n_m}}{\sigma^2(1 + (n_m - 1)\delta_{n,m})} Z_{n,i_m}(x) - \frac{1}{2} \sum_{m=1}^{M_n} \frac{n^2\bar{\mu}_n^2/\sqrt{n_m}}{\sigma^2(1 + (n_m - 1)\delta_{n,m})},
 \end{aligned}$$

where $Z_{m,i}(x)$, $i \in N_{n,m}$, are the entries of $Z_m(x) = B_m^\top x_{n,m}$, B_m is an $n_m \times n_m$ orthogonal matrix, i_m denotes the first index in $N_{n,m}$, and $x_{n,m} = [x_i]_{i \in N_{n,m}}$.

Lemma A.4. Suppose that $f : \mathbf{R} \rightarrow \mathbf{R}^+$ is a continuously differentiable function such that for some $\delta > 0$,

$$\sup_{x \in [-\delta, \delta]} \left| \frac{d \log f(x)}{dx} \right| |\bar{x}| < 1,$$

where $\bar{x} \in \mathbf{R}$ is such that

$$\sup_{x \in [-\delta, \delta]} f(x) = f(\bar{x}).$$

Then,

$$f(\bar{x}) \leq \left(1 - \sup_{x \in [-\delta, \delta]} \left| \frac{d \log f(x)}{dx} \right| |\bar{x}| \right)^{-1} f(0).$$

Proof. Using the Mean Value Theorem,

$$\begin{aligned}
 \exp \log f(x) &= \exp \log f(0) + \frac{d \log f(x^*(x))}{dx} (\exp \log f(x^*(x))) x \\
 &\leq \exp \log f(0) + \sup_{\bar{x} \in [-\delta, \delta]} \left| \frac{d \log f(\bar{x})}{dx} \right| (\exp \log f(\bar{x})) |x|,
 \end{aligned}$$

where $x^*(x)$ is a point on the line segment between 0 and x . Evaluating the inequality at $x = \bar{x}$ gives us the desired result. ■

Recall the definition of n^* in Corollary 4.1:

$$n^* = \sum_{m=1: n_m \geq 2}^{M_n} n_m.$$

The number $n - n^*$ represents the number of random variables, $X_{n,i}$, that are known to be mutually independent. Each variable outside this set belongs to a non-singleton cluster.

Lemma A.5. $\lim_{n \rightarrow \infty} \sum_{m=1}^{M_n} (n_m/n)^2 = 0$ if and only if

- (a) $\lim_{n \rightarrow \infty} n^*/n = 0$, or
- (b) $\lim_{n \rightarrow \infty} \sum_{m=1: n_m \geq 2}^{M_n} (n_m/n^*)^2 = 0$.

Proof. For each $n \geq 1$, we have either

$$\sum_{m=1}^{M_n} \frac{n_m^2}{n^2} = \begin{cases} \frac{1}{n}, & \text{if } n^* = 0, \\ \left(\frac{n^*}{n}\right)^2 \sum_{m=1: n_m \geq 2}^{M_n} \left(\frac{n_m}{n^*}\right)^2 + \left(\frac{1}{n}\right)^2 (n - n^*), & \text{if } n^* > 0. \end{cases}$$

Since $(1/n)^2(n - n^*) = o(1)$, $\lim_{n \rightarrow \infty} \sum_{m=1}^{M_n} (n_m/n)^2 = 0$ if and only if (a) or (b) holds. ■

Lemma A.6. Suppose that $n \geq 2$, and Σ_n is a block diagonal matrix along a cluster structure \mathcal{M}_n , where the m -th block, denoted by $\Sigma_{n,m}$ is given by

$$\Sigma_{n,m} = \sigma^2(1 - \delta_{n,m})I_{n_m} + \sigma^2\delta_{n,m}\mathbf{1}_{n_m}\mathbf{1}_{n_m}^\top,$$

where

$$\delta_{n,m} = \frac{\bar{\delta}}{n^*}, \tag{A.2}$$

for some $\bar{\delta} \in [-a, a]$, with $0 < a < 1/2$, if $n_m \geq 2$, and $\sigma^2 > 0$ is independent of n . Then, the following holds for any random vector $X_n \in \mathbf{R}^n$ which follows $\Phi(0, \sigma^2 I_n)$.

- (i) $(d\Phi(0, \Sigma_n)/d\Phi(0, \sigma^2 I_n))(X_n)$ is uniformly integrable.
- (ii) $\log((d\Phi(0, \Sigma_n)/d\Phi(0, \sigma^2 I_n))(X_n))$ is uniformly tight.

Proof. For brevity, we focus on the case $\sigma^2 = 1$. By Corollary A.1,

$$\log \frac{d\Phi(0, \Sigma_n)}{d\Phi(0, I_n)}(X_n) = A_n + R_n,$$

where

$$A_n = - \sum_{m=1}^{M_n} \log \sqrt{1 + (n_m - 1)\delta_{n,m}} + \sum_{m=1}^{M_n} \frac{(n_m - 1)\delta_{n,m} Z_{m,i_m}^2}{2(1 + (n_m - 1)\delta_{n,m})}, \text{ and}$$

$$R_n = - \sum_{m=1}^{M_n} (n_m - 1) \log \sqrt{1 - \delta_{n,m}} - \sum_{m=1}^{M_n} \frac{\delta_{n,m}}{2(1 - \delta_{n,m})} \sum_{i \in N_{n,m} \setminus \{i_m\}} Z_{m,i}^2,$$

and i_m denotes the first i in block m .

(i) Let us take small $\epsilon > 0$ such that

$$a < \frac{2}{(1 + \epsilon)(4 + \epsilon)}. \tag{A.3}$$

We write (under $\Phi(0, I_n)$)

$$\mathbf{E} \left[\left(\frac{d\Phi(0, \Sigma_n)}{d\Phi(0, I_n)} \right)^{(1+\epsilon)} \right] = \mathbf{E} [\exp((1 + \epsilon)A_n)] \mathbf{E} [\exp((1 + \epsilon)R_n)],$$

since A_n and R_n are independent. Let $t_m = (n_m - 1)/n^*$, and write

$$\begin{aligned} \mathbf{E} [\exp((1 + \epsilon)A_n)] &= \prod_{m=1}^{M_n} \left(1 + t_m \bar{\delta} \right)^{-\frac{1+\epsilon}{2}} \mathbf{E} \left[\exp \left(\frac{t_m(1 + \epsilon)\bar{\delta}}{2(1 + t_m\bar{\delta})} Z_{m,1}^2 \right) \right] \\ &\leq \prod_{m=1}^{M_n} \left(1 + t_m \bar{\delta} \right)^{-\frac{1+\epsilon}{2}} \left(1 - t_m(1 + \epsilon)\bar{\delta} \right)^{-1/2} \equiv f_n(\bar{\delta}), \text{ say.} \end{aligned}$$

Note that

$$\frac{d \log f_n(\bar{\delta})}{d \bar{\delta}} = \frac{(1 + \epsilon)(2 + \epsilon)\bar{\delta}}{2} \sum_{m=1}^{M_n} \frac{t_m^2}{(1 + t_m\bar{\delta})(1 - t_m\bar{\delta}(1 + \epsilon))} > 0,$$

because $t_m \leq 1$ and $\bar{\delta}(1 + \epsilon) < 1$ by (A.3). This means that $f_n(\bar{\delta})$ is increasing in $\bar{\delta} \in [-a, a]$ and achieves its maximum at $\bar{\delta} = a$. Hence,

$$\begin{aligned} \left| \frac{d \log f_n(\bar{\delta})}{d \bar{\delta}} \right| a &\leq \left| \frac{d \log f_n(\bar{\delta})}{d \bar{\delta}} \right| \leq \frac{(1 + \epsilon)(2 + \epsilon)\bar{\delta}}{2(1 - (1 + \epsilon)\bar{\delta})} \sum_{m=1}^{M_n} t_m^2 \\ &\leq \frac{(1 + \epsilon)(2 + \epsilon)a}{2(1 - (1 + \epsilon)a)} < 1, \end{aligned}$$

because $\sum_{m=1}^{M_n} t_m^2 \leq 1$ and we chose ϵ such that (A.3) holds. By Lemma A.4, we have

$$f_n(\bar{\delta}) \leq f_n(a) \leq \left(1 - \frac{(1 + \epsilon)(2 + \epsilon)a}{2(1 - (1 + \epsilon)a)} \right)^{-1}.$$

The bound does not depend on n , and hence,

$$\sup_{n \geq 1} \mathbf{E} [\exp((1 + \epsilon)A_n)] < \infty.$$

Now, we turn to $\mathbf{E} [\exp((1 + \epsilon)R_n)]$. We can write

$$R_n = -\frac{1}{2} n^* \log \left(1 - \frac{\bar{\delta}}{n^*} \right) - \frac{1}{2} \frac{\bar{\delta}/n^*}{1 - \bar{\delta}/n^*} \sum_{m=1: n_m \neq 1}^{M_n} \sum_{i \in N_{n,m} \setminus \{i_m\}} Z_{m,i}^2. \tag{A.4}$$

Using this expression, we rewrite

$$\begin{aligned} \mathbf{E} [\exp((1 + \epsilon)R_n)] &= \left(1 - \frac{\bar{\delta}}{n^*} \right)^{-\frac{n^*(1+\epsilon)}{2}} \prod_{m=1: n_m \neq 1}^{M_n} \prod_{i \in N_{n,m} \setminus \{i_m\}} \mathbf{E} \left[\exp \left(-\frac{1}{2} \frac{\bar{\delta}(1 + \epsilon)/n^*}{1 - \bar{\delta}/n^*} Z_{m,i}^2 \right) \right] \\ &= \left(1 - \frac{\bar{\delta}}{n^*} \right)^{-\frac{n^*(1+\epsilon)}{2}} \prod_{m=1: n_m \neq 1}^{M_n} \prod_{i \in N_{n,m} \setminus \{i_m\}} \left(1 + \frac{\bar{\delta}(1 + \epsilon)/n^*}{1 - \bar{\delta}/n^*} \right)^{-\frac{1}{2}} \end{aligned}$$

$$\leq \left(1 - \frac{\bar{\delta}}{n^*}\right)^{-\frac{n^*(1+\epsilon)}{2}} \leq \left(1 - \frac{a}{n^*}\right)^{-\frac{n^*(1+\epsilon)}{2}}.$$

The last bound is a sequence converging to $\exp(a(1 + \epsilon)/2)$ as $n^* \rightarrow \infty$, and hence, is a bounded sequence. Thus, we conclude that

$$\sup_{n \geq 1} \mathbf{E} [\exp((1 + \epsilon)R_n)] < \infty.$$

This proves that

$$\sup_{n \geq 1} \mathbf{E} \left[\left(\frac{d\Phi(0, \Sigma_n)}{d\Phi(0, I_n)} \right)^{(1+\epsilon)} (X_n) \right] < \infty.$$

Hence, the proof of (i) is complete.

(ii) We rewrite

$$A_n = -\frac{1}{2} \sum_{m=1}^{M_n} \left(\log(1 + (n_m - 1)\delta_{n,m}) - \frac{(n_m - 1)\delta_{n,m}}{1 + (n_m - 1)\delta_{n,m}} \right) + \sum_{m=1}^{M_n} \frac{(n_m - 1)\delta_{n,m}(Z_{m,i_m}^2 - 1)}{2(1 + (n_m - 1)\delta_{n,m})}. \tag{A.5}$$

For any $x \in [0, 1]$, we have

$$\frac{x^2}{8} \leq \frac{1}{2} \frac{x^2}{(1+x)^2} \leq \log(1+x) - \frac{x}{1+x} \leq \frac{1}{2} \frac{x^2}{1+x} \leq \frac{x^2}{2}.$$

Hence,

$$\left| \frac{1}{2} \sum_{m=1}^{M_n} \left(\log(1 + (n_m - 1)\delta_{n,m}) - \frac{(n_m - 1)\delta_{n,m}}{1 + (n_m - 1)\delta_{n,m}} \right) \right| \leq \frac{\bar{\delta}^2}{2} \sum_{m=1}^{M_n} \frac{(n_m - 1)^2}{2n^{*2}} \leq \frac{\bar{\delta}^2}{2}.$$

It suffices to show the uniform tightness of the second sum in (A.5). Under $\Phi(0, I_n)$, it has mean zero, and

$$\begin{aligned} \text{Var} \left(\sum_{m=1: n_m \geq 2}^{M_n} \frac{\bar{\delta}(n_m - 1)/n^*}{1 + (\bar{\delta}(n_m - 1)/n^*)} (Z_{m,i_m}^2 - 1) \right) &\leq \sum_{m=1}^{M_n} \frac{(\bar{\delta}(n_m - 1)/n^*)^2}{(1 + (\bar{\delta}(n_m - 1)/n^*))^2} \text{Var} (Z_{m,i_m}^2) \\ &\leq 2\bar{\delta}^2 \sum_{m=1: n_m \geq 2}^{M_n} \left(\frac{n_m - 1}{n^*} \right)^2 \leq 2\bar{\delta}^2, \end{aligned}$$

because $\text{Var}(Z_{m,i_m}^2) = 2$. Therefore, A_n is uniformly tight.

As for R_n , we recall (A.4), and can follow similar arguments to show that R_n is uniformly tight as well. ■

A.2. Consistent discrimination of the mean

We let

$$\Sigma(\sigma^2, \delta) = \sigma^2 (I_n + \delta \mathbf{1}_n \mathbf{1}_n^\top - \delta I_n),$$

and for any $\bar{\mu} \in \mathbf{R}$, we write $\Phi(\bar{\mu} \mathbf{1}_n, \Sigma(\sigma^2, \delta))$ simply as $\Phi(\bar{\mu}, \sigma^2, \delta)$. Let us recall some basic notions of optimality of tests (Lehmann and Romano, 2005). Given a model \mathcal{P}_n which is partitioned as $\mathcal{P}_{n,0} \cup \mathcal{P}_{n,1}$, a test ϕ_n is said to be a **UMP (uniformly most powerful) test** of $\mathcal{P}_{n,0}$ against $\mathcal{P}_{n,1}$ at level $\alpha \in (0, 1)$, if under any $P_{n,0} \in \mathcal{P}_{n,0}$,

$$\mathbf{E}[\phi_n] \leq \alpha,$$

and for any alternative test ϕ'_n such that $\mathbf{E}[\phi'_n] \leq \alpha$ under any $P_{n,0} \in \mathcal{P}_{n,0}$, we have

$$\mathbf{E}[\phi'_n] - \mathbf{E}[\phi_n] \leq 0,$$

under any $P_{n,1} \in \mathcal{P}_{n,1}$.

A sequence of tests ϕ_n is said to be an **AUMP (asymptotically uniformly most powerful) test** of $\mathcal{P}_{n,0}$ against $\mathcal{P}_{n,1}$ at level $\alpha \in (0, 1)$, if under any sequence $P_{n,0} \in \mathcal{P}_{n,0}$,

$$\limsup_{n \rightarrow \infty} \mathbf{E}[\phi_n] \leq \alpha,$$

and for any alternative test ϕ'_n such that $\limsup_{n \rightarrow \infty} \mathbf{E}[\phi'_n] \leq \alpha$ under any sequence $P_{n,0} \in \mathcal{P}_{n,0}$, we have

$$\limsup_{n \rightarrow \infty} \mathbf{E}[\phi'_n] - \mathbf{E}[\phi_n] \leq 0,$$

under any sequence $P_{n,1} \in \mathcal{P}_{n,1}$.

Lemma A.7. Suppose that $A \subset \mathbf{R}$ is an open interval, and $\{\varphi_\alpha\}_{\alpha \in A}$ is a class of tests of $\mathcal{P}_{n,0}$ against $\mathcal{P}_{n,1}$, such that for each $\alpha \in A$, the test φ_α is UMP at level α , and for any $\tilde{\alpha} = \alpha + o(1)$ as $n \rightarrow \infty$,

$$\mathbf{E}[\varphi_\alpha] = \mathbf{E}[\varphi_{\tilde{\alpha}}] + o(1), \tag{A.6}$$

under any sequence $P_n \in \mathcal{P}_{n,0} \cup \mathcal{P}_{n,1}$. Then, φ_α is AUMP at level α .

Proof. Choose any test $\tilde{\varphi}$ such that under any sequence $P_n \in \mathcal{P}_{n,0}$,

$$E[\tilde{\varphi}] \leq \alpha + \epsilon_n,$$

for some sequence $\epsilon_n \rightarrow 0$, as $n \rightarrow \infty$. Fix one such sequence $P_n \in \mathcal{P}_{n,0}$, together with the sequence ϵ_n , and let $\tilde{\alpha} = \alpha + \epsilon_n$. Now, select a large enough n such that $\tilde{\alpha} \in A$ and choose any $P'_n \in \mathcal{P}_{n,1}$. Then, since $\varphi_{\tilde{\alpha}}$ is UMP at level $\tilde{\alpha}$, we have

$$E_{P'_n}[\varphi_{\tilde{\alpha}}] \geq E_{P'_n}[\tilde{\varphi}]$$

under P'_n . By (A.6), we can see that φ_{α} is AUMP at level α . ■

Proof of Theorem 3.1. Suppose that $M_n = 1$. First, consider the case where $\mathcal{P}_n = \mathcal{P}'_{n,\mathcal{N}}$, with

$$\mathcal{P}'_{n,\mathcal{N}} = \{ \Phi(\mu, \sigma^2, \delta) : (n-1)\delta \in (0, 1), \sigma^2 > 0, \mu \geq 0 \}.$$

Later, we generalize the result to the case where \mathcal{P}_n contains the above probability model. Define

$$\mathcal{P}_{n,0} = \{ \Phi(0, \sigma^2, \delta) : (n-1)\delta \in (0, 1), \sigma^2 > 0 \}$$

and let $\mathcal{P}_{n,1} = \mathcal{P}_n \setminus \mathcal{P}_{n,0}$. In light of Lemma A.7, it suffices to construct a class of tests $\{ \varphi_{\alpha} \}_{\alpha \in (0, 1/2)}$ of $\mathcal{P}_{n,0}$ against $\mathcal{P}_{n,1}$ such that

- (a) it satisfies (A.6) in Lemma A.7,
- (b) the test φ_{α} has power bounded by a constant below 1 uniformly over n , and
- (c) each test φ_{α} is a UMP test of $\mathcal{P}_{n,0}$ against $\mathcal{P}_{n,1}$.

Let us first construct such a test and show that (a)–(c) are satisfied. Define

$$V_n = \frac{Z_{n,1}}{|Z_{n,1}|} = \text{sgn}(Z_{n,1}).$$

For each $\alpha \in (0, 1/2)$, let

$$\phi_{\alpha}(V_n) = \begin{cases} 1, & \text{if } V_n > C_0 \\ \gamma_0, & \text{if } V_n = C_0, \\ 0, & \text{if } V_n < C_0, \end{cases}$$

for some $C_0 > 0$ and $\gamma_0 \in [0, 1]$. Let $Z = (Z_{n,1} - E[Z_{n,1}]) / \sqrt{\text{Var}(Z_{n,1})}$. Then the size control requires that under the null hypothesis,

$$E[\phi_{\alpha}(V_n)] = P\{Z > C_0 | Z|\} + P\{Z = C_0 | Z|\} \gamma_0 = \alpha.$$

Since $\alpha \in (0, 1/2)$, we must have $C_0 = 1$ and $\gamma_0 = 2\alpha$.

Let us first show that this test satisfies the condition (a). For any $\tilde{\alpha}$ such that $\tilde{\alpha} = \alpha + o(1)$, and under any sequence $P_n \in \mathcal{P}_n$,

$$E[\phi_{\tilde{\alpha}}(V_n)] = 2\tilde{\alpha}P_n\{Z_{n,1} \geq 0\} = 2\alpha P_n\{Z_{n,1} \geq 0\} + o(1) = E[\phi_{\alpha}(V_n)] + o(1).$$

Hence, the class of tests $\{ \phi_{\alpha}(V_n) \}_{\alpha \in (0, 1/2)}$ satisfies the condition (A.6).

As for the condition (b), note that under any alternative hypothesis in $\mathcal{P}_{n,1}$, we have

$$E[\varphi_{\alpha}(V_n)] = P_n\{Z_{n,1} > |Z_{n,1}|\} + 2\alpha P_n\{Z_{n,1} = |Z_{n,1}|\} = 2\alpha P_n\{Z_{n,1} \geq 0\} \leq 2\alpha.$$

Hence, the test does not have power exceeding $2\alpha < 1$.

Finally, we show that the condition (c) is satisfied. Let

$$\begin{aligned} \mathcal{L}(X_n; \mu, \sigma^2, \delta_n) &= -\log \sqrt{\sigma^2(1 + (n-1)\delta_n)} - (n-1) \log \sqrt{\sigma^2(1 - \delta_n)} \\ &\quad + \frac{\sigma^2(1 + (n-1)\delta_n) - 1}{2\sigma^2(1 + (n-1)\delta_n)} Z_{n,1}^2 + \frac{\sigma^2(1 - \delta_n) - 1}{2\sigma^2(1 - \delta_n)} \sum_{k=2}^n Z_{n,k}^2 \\ &\quad + \sqrt{n}\mu Z_{n,1} - \frac{n\mu^2 \sigma^2(1 + (n-1)\delta_n)}{2}. \end{aligned} \tag{A.7}$$

Hence, $\mathcal{L}(X_n; \mu, \sigma^2, \delta_n)$ is the same as $\log(d\Phi(\mu \mathbf{1}_n, \Sigma(\sigma^2, \delta_n)) / d\Phi(0, I_n))(X_n)$ in Lemma A.3, except that the coefficient of $Z_{n,1}$ is $\sqrt{n}\mu$ and the last term is different. Define a probability measure $P_n(\mu, \delta_n)$ as follows: for any Borel B ,

$$P_n(\mu, \delta_n)(B) = \int_B \exp(\mathcal{L}(x; \mu, \delta_n)) d\Phi(0, I_n)(x).$$

Similarly as before, we define

$$\tilde{P}_n = \{ P_n(\mu, \delta) : (n-1)\delta \in (0, 1), \mu \geq 0 \}, \quad \tilde{P}_{n,0} = \{ P_n(0, \delta) : (n-1)\delta \in (0, 1) \},$$

and let $\tilde{P}_{n,1} = \tilde{P}_n \setminus \tilde{P}_{n,0}$. It is not hard to see that

$$\tilde{P}_{n,0} = \mathcal{P}_{n,0} \quad \text{and} \quad \tilde{P}_{n,1} = \mathcal{P}_{n,1}.$$

Therefore, a UMP test of $\tilde{\mathcal{P}}_{n,0}$ against $\tilde{\mathcal{P}}_{n,1}$ is also a UMP test of $\mathcal{P}_{n,0}$ against $\mathcal{P}_{n,1}$. It suffices for condition (c) to show that the test φ_α is a UMP test of $\tilde{\mathcal{P}}_{n,0}$ against $\tilde{\mathcal{P}}_{n,1}$. From (A.7), the sufficient statistics for $\tilde{\mathcal{P}}_n$ in the case of $M_n = 1$ are given by

$$\left(Z_{n,1}, Z_{n,1}^2, \sum_{k=2}^n Z_{n,k}^2 \right),$$

where $Z_{n,k}$'s are as in Lemma A.3. For any $t \geq 0$,

$$\begin{aligned} P \left\{ V_n = 1, Z_{n,1}^2 \leq t \right\} &= P \left\{ V_n = 1, -\sqrt{t} \leq Z_{n,1} \leq \sqrt{t} \right\} = P \left\{ Z_{n,1} > 0, -\sqrt{t} \leq Z_{n,1} \leq \sqrt{t} \right\} \\ &= P \left\{ 0 \leq Z_{n,1} \leq \sqrt{t} \right\} = \frac{1}{2} P \left\{ -\sqrt{t} \leq Z_{n,1} \leq \sqrt{t} \right\} = P \left\{ V_n = 1 \right\} P \left\{ Z_{n,1}^2 \leq t \right\}, \end{aligned}$$

under the null hypothesis. Hence, V_n and $Z_{n,1}^2$ are independent under any probability in $\tilde{\mathcal{P}}_{n,0}$. Furthermore, under any probability in $\tilde{\mathcal{P}}_{n,0}$,

$$\begin{aligned} E \left[Z_n Z_n^\top \right] &= B_n^\top E \left[Z_n Z_n^\top \right] B_n \\ &= B_n^\top \Sigma_n B_n = B_n^\top (I_n + \delta_n \mathbf{1}_n \mathbf{1}_n^\top - \delta_n I_n) B_n = (1 - \delta_n) I_n + B_n^\top \mathbf{1}_n \mathbf{1}_n B_n. \end{aligned}$$

Note that $B_n^\top \mathbf{1}_n \mathbf{1}_n B_n$ is a matrix whose $(1, 1)$ -th entry is $b_1^\top \mathbf{1}_n \mathbf{1}_n^\top b_1 = 1$ and all the other entries are zeros. Hence, $Z_{n,k}$'s are independent across k 's under any probability in $\tilde{\mathcal{P}}_{n,0}$. Therefore, V_n and $(Z_{n,1}^2, \sum_{k=2}^n Z_{n,k}^2)$ are independent under any probability in $\tilde{\mathcal{P}}_{n,0}$. By Theorem 5.1.1 of Lehmann and Romano (2005), the randomized test $\varphi_\alpha(V_n)$ is an α -level UMP test.

Next, consider the case where $\mathcal{P}'_{n,\mathcal{N}} \subset \mathcal{P}_n$. Take a sequence of tests φ_n such that for any sequence of probabilities $P_n \in \mathcal{P}_{n,0}$, $\limsup_{n \rightarrow \infty} E[\varphi_n(X_n)] \leq \alpha$. Now, we take a sequence $P_{n,1} \in \mathcal{P}_{n,1} \cap \mathcal{P}'_{n,\mathcal{N}}$. For any $\alpha \in (0, 1/2)$, the test $\varphi_\alpha(V_n)$ is a UMP test at level α of the null hypothesis $\mathcal{P}_{n,0} \cap \mathcal{P}'_{n,\mathcal{N}}$ against $\mathcal{P}_{n,1} \cap \mathcal{P}'_{n,\mathcal{N}}$. Note that for any sequence of probabilities $P_n \in \mathcal{P}_{n,0} \cap \mathcal{P}'_{n,\mathcal{N}}$, $\limsup_{n \rightarrow \infty} E[\varphi_n(X_n)] \leq \alpha$. Hence, if we take $\epsilon > 0$ such that $\alpha + \epsilon \in (0, 1/2)$, there exists $n_0 \geq 1$ such that for all $n \geq n_0$, $E[\varphi_n(X_n)] \leq \alpha + \epsilon$. For all such n , under any $P_n \in \mathcal{P}_{n,1} \cap \mathcal{P}'_{n,\mathcal{N}}$, we have

$$E[\varphi_n(X_n)] \leq E[\varphi_{\alpha+\epsilon}(V_n)] \leq 2(\alpha + \epsilon).$$

Hence, we find that along any sequence $P_n \in \mathcal{P}_{n,1} \cap \mathcal{P}'_{n,\mathcal{N}}$, we have

$$\liminf_{n \rightarrow \infty} E[\varphi_n(X_n)] \leq 2(\alpha + \epsilon) < 1.$$

Thus, the proof is complete. ■

The following lemma is used for the proof of Theorem 3.2. For $w \in [0, 1]$, define $\xi_1 = wZ_1$ and $\xi_2 = (1-w)Z_2$, where $Z_i \sim N(0, 1)$, independent across $i = 1, 2$. Define

$$T(w) = \frac{(\xi_1 + \xi_2)/\sqrt{2}}{\sqrt{(\xi_1 - \xi_2)^2 + (\xi_2 - \xi_1)^2}},$$

where $\bar{\xi} = (\xi_1 + \xi_2)/2$. Let us take $t \geq 0$, and define

$$p(w; t) = P\{T(w) \leq t\}.$$

Lemma A.8.

- (i) For all $t \geq 1$ and all $w \in [0, 1]$, $p(w; t) \geq p(1/2; t)$.
- (ii) For all $0 \leq t < 1$ and all $w \in [0, 1]$, $p(w; t) \leq p(1/2; t)$.

Proof. First, we write

$$T(w) = \frac{\xi_1 + \xi_2}{\sqrt{(\xi_1 - \xi_2)^2}}.$$

For (i) and (ii), since $[\xi_1, \xi_2]$ is symmetrically distributed around the origin, it suffices to show that $p(\cdot; t)$ is increasing on $[1/2, 1]$ for all $t \geq 1$, and $p(\cdot; t)$ is decreasing on $[1/2, 1]$ for all $0 \leq t < 1$. Let A denote the event that $\xi_1 + \xi_2 < 0$. Then, on the event A , for all $w \geq 0$ and all $t \geq 0$, we have $T(w) \leq t$. Hence,

$$P(\{T(w) \leq t\} \cap A) = P(A) = 0.5.$$

On the event A^c , $T(w) \leq t$ if and only if

$$w^2 Z_1^2 + (1-w)^2 Z_2^2 + 2w(1-w)Z_1 Z_2 \leq (w^2 Z_1^2 + (1-w)^2 Z_2^2 - 2w(1-w)Z_1 Z_2)t^2$$

if and only if

$$0 \leq (w^2 Z_1^2 + (1-w)^2 Z_2^2)(t^2 - 1) - 2(t^2 + 1)w(1-w)Z_1 Z_2 = f(w; Z_1, Z_2),$$

where

$$f(w; Z_1, Z_2) = (wZ_1 + (1 - w)Z_2)^2(t^2 - 1) - 4t^2w(1 - w)Z_1Z_2.$$

Take $t \geq 1$. Let B be the event $Z_1Z_2 \leq 0$. Certainly, if $Z_1Z_2 \leq 0$, $f(w; Z_1, Z_2) \geq 0$ for all $w \in [1/2, 1]$. Hence,

$$\begin{aligned} P(\{T(w) \leq t\} \cap A^c) &= P(\{T(w) \leq t\} \cap A^c \cap B) + P(\{T(w) \leq t\} \cap A^c \cap B^c) \\ &= P(A^c \cap B) + P(\{T(w) \leq t\} \cap A^c \cap B^c). \end{aligned}$$

We show that $f(w; Z_1, Z_2)$ is increasing in w on the event A^c . We take the derivative $f'(w; Z_1, Z_2)$ with respect to w :

$$f'(w; Z_1, Z_2) = 2(wZ_1 - (1 - w)Z_2)(t^2 - 1)(Z_1 - Z_2) - 4t^2(1 - 2w)Z_1Z_2.$$

The function is linear in w . First, we take $w = 1$. Then, on the event B^c ,

$$\begin{aligned} f'(1; Z_1, Z_2) &= 2(t^2 - 1)Z_1(Z_1 - Z_2) + 4t^2Z_1Z_2 = 2(t^2 - 1)Z_1^2 + (4t^2 - 2(t^2 - 1))Z_1Z_2 \\ &= 2(t^2 - 1)Z_1^2 + (2t^2 + 2)Z_1Z_2 \geq 0. \end{aligned}$$

Second, we take $w = 1/2$. Then,

$$f'(1/2; Z_1, Z_2) = (t^2 - 1)(Z_1 - Z_2)^2.$$

Hence, for all Z_1, Z_2 such that $Z_1Z_2 > 0$, $f(w; Z_1, Z_2)$ is increasing on $[1/2, 1]$ for all $t \geq 1$. Therefore, whenever $t \geq 1$, $P\{T(w) \leq t\}$ is increasing on $[1/2, 1]$.

Take $0 \leq t < 1$. If $Z_1Z_2 > 0$, then $f(w; Z_1, Z_2) < 0$. Hence,

$$\begin{aligned} P(\{T(w) \leq t\} \cap A) &= P(\{T(w) \leq t\} \cap A^c \cap B) + P(\{T(w) \leq t\} \cap A^c \cap B^c) \\ &= P(\{T(w) \leq t\} \cap A^c \cap B). \end{aligned}$$

On the event B , when $w = 1$,

$$f'(1; Z_1, Z_2) = 2(t^2 - 1)Z_1^2 + (2t^2 + 2)Z_1Z_2 \leq 0,$$

and when $w = 1/2$,

$$f'(1/2; Z_1, Z_2) = (t^2 - 1)(Z_1 - Z_2)^2 \leq 0.$$

Hence, for all Z_1, Z_2 such that $Z_1Z_2 \leq 0$, $f(w; Z_1, Z_2)$ is decreasing on $[1/2, 1]$ for all $0 \leq t < 1$. Therefore, whenever $0 \leq t < 1$, $P\{T(w) \leq t\}$ is decreasing on $[1/2, 1]$. ■

Proof of Theorem 3.2. Suppose that we have at least two nonnegligible clusters, i.e., $M_n \geq 2$ for all but finite number of n 's. Consider testing the null hypothesis of $E[X_{n,i}] = 0$ against $E[X_{n,i}] > 0$. Without loss of generality, we enumerate $\mathcal{M}'_n = \{N_1, \dots, N_{M'_n}\}$, $M'_n \geq 2$. Now, we construct a test that consistently \sqrt{n} -discriminates the mean. Define

$$\xi_m = \frac{1}{\sqrt{n_m}} \sum_{i \in N_{n,m}} X_{n,i},$$

and

$$U'_n = \frac{1}{2} \sum_{m=1}^2 \xi_m \quad \text{and} \quad T'_n = \sum_{m=1}^2 \left(\xi_m - \frac{1}{2} \sum_{m=1}^2 \xi_m \right)^2.$$

We take

$$V'_n = \frac{\sqrt{2U'_n}}{\sqrt{T'_n}}.$$

Let $c_{1-\alpha}$ be the $1 - \alpha$ quantile of the t -distribution with degree of freedom 1. Define

$$\varphi_n = 1\{V'_n > \max\{c_{0.75}, c_{1-\alpha}\}\}.$$

Note that $c_{0.75} = 1$.

We first show that this test controls the size of the test at α asymptotically under the null hypothesis. Define an infeasible test statistic

$$V''_n = \frac{\sqrt{2U''_n}}{\sqrt{T''_n}},$$

where

$$U''_n = \frac{1}{2} \sum_{m=1}^2 \frac{\xi_m}{\sigma_{n,m}} \quad \text{and} \quad T''_n = \sum_{m=1}^2 \left(\frac{\xi_m}{\sigma_{n,m}} - \frac{1}{2} \sum_{m=1}^2 \frac{\xi_m}{\sigma_{n,m}} \right)^2.$$

Here $\sigma_{n,m}^2$ is the variance of ξ_m under the null hypothesis. Then V_n'' converges in distribution to the t -distribution with 1 degree of freedom under the null hypothesis. By Lemma A.8, if $0 < \alpha < 0.25$ so that $c_{1-\alpha} > 1$,

$$\mathbf{E}[\varphi_n] = P\{V_n' > c_{1-\alpha}\} \leq P\{V_n'' > c_{1-\alpha}\} = \alpha + o(1),$$

and if $\alpha \geq 0.25$ so that $0 \leq c_{1-\alpha} \leq 1$,

$$\mathbf{E}[\varphi_n] = 1 - P\{V_n' \leq c_{0.75}\} \leq 1 - P\{V_n'' \leq c_{0.75}\} = 0.25 \leq \alpha + o(1),$$

as $n \rightarrow \infty$. Hence, the size of the test φ_n is bounded by α asymptotically.

Suppose that we are under the local alternatives such that $\mathbf{E}[X_n] = \bar{\mu}_n \sigma_{LR} \mathbf{1}_n / \sqrt{n}$, for some sequence $\bar{\mu}_n \rightarrow \infty$. Define

$$\bar{U}_n = U_n' - \frac{1}{2} \sum_{m=1}^2 \sqrt{\frac{n_m}{n}} \bar{\mu}_n \sigma_{LR}.$$

Note that

$$\begin{aligned} P\{V_n' > c_{1-\alpha}\} &= P\left\{ \frac{\sqrt{2}\bar{U}_n}{\sigma_{LR}} > c_{1-\alpha} \sqrt{\frac{T_n'}{\sigma_{LR}^2}} - \frac{1}{\sqrt{2}} \sum_{m=1}^2 \sqrt{\frac{n_m}{n}} \bar{\mu}_n \right\} \\ &\geq P\left\{ \frac{\sqrt{2}\bar{U}_n}{\sigma_{LR}} > c_{1-\alpha} \sqrt{\frac{T_n'}{\sigma_{LR}^2}} - \frac{1}{\sqrt{2}} \sum_{m=1}^2 \frac{n_m}{n} \bar{\mu}_n \right\} \end{aligned} \tag{A.8}$$

because $n_m/n \leq 1$. Note that

$$\sigma_{LR}^2 = \sum_{m=1}^{M_n} \frac{n_m}{n} \sigma_{n,m}^2 \geq \sum_{m=1}^2 \frac{n_m}{n} \sigma_{n,m}^2.$$

Since

$$0 < \liminf_{n \rightarrow \infty} \min_{m=1,2} \frac{n_m}{n} \leq \limsup_{n \rightarrow \infty} \min_{m=1,2} \frac{n_m}{n} \leq 1,$$

there exist $\epsilon > 0$ and $n_0 > 0$ such that for all $n \geq n_0$,

$$\sigma_{LR}^2 \geq \epsilon \sum_{m=1}^2 \sigma_{n,m}^2 \geq \epsilon \max_{m=1,2} \sigma_{n,m}^2.$$

Therefore,

$$\frac{|\bar{U}_n|}{\sigma_{LR}} \leq \frac{\max_{m=1,2} \sigma_{n,m}}{2\sigma_{LR}} \sum_{m=1}^2 \frac{|\xi_m - \mathbf{E}[\xi_m]|}{\sigma_{n,m}} \leq \frac{1}{2\sqrt{\epsilon}} \sum_{m=1}^2 \frac{|\xi_m - \mathbf{E}[\xi_m]|}{\sigma_{n,m}}.$$

Since $(\xi_m - \mathbf{E}[\xi_m])/\sigma_{n,m}$ converges in distribution to $N(0, 1)$ under any sequence $P_n \in \mathcal{P}_n$ as $n \rightarrow \infty$ by (3.1), we have $\bar{U}_n/\sigma_{LR} = O_P(1)$. Similarly, we can show that $T_n'/\sigma_{LR}^2 = O_P(1)$. Hence, the last probability in (A.8) converges to one as $\bar{\mu}_n \rightarrow \infty$, proving that the mean is consistently \sqrt{n} -discriminated. ■

A.3. Impossibility of consistent estimation of long run variance

Proof of Lemma 4.1. We first show sufficiency. Suppose that $\|\theta_n(P_{n,1}) - \theta_n(P_{n,0})\| = o(1)$ for any sequence $P_{n,1} \in \mathcal{P}_n$. Then we take $\hat{\theta} = \theta_n(P_{n,0})$, so that $\|\hat{\theta} - \theta_n(P_{n,1})\| \rightarrow 0$, along $P_{n,1} \in \mathcal{P}_n$, as $n \rightarrow \infty$. Hence, sufficiency follows.

Conversely, suppose that θ_n is consistently estimable in \mathcal{P}_n , so that there exists an estimator, say, $\tilde{\theta}$, such that $\tilde{\theta} - \theta_n(P_{n,1}) = o_P(1)$ along any $P_{n,1} \in \mathcal{P}_n$. Since $P_{n,0} \in \mathcal{P}_n$, this means that $\tilde{\theta} - \theta_n(P_{n,0}) = o_P(1)$ under $P_{n,0}$. Since $P_{n,1} \triangleleft P_{n,0}$, $\tilde{\theta} - \theta_n(P_{n,0}) = o_P(1)$ under any $P_{n,1} \in \mathcal{P}_n$. We choose any $P_{n,1} \in \mathcal{P}_n$ and write

$$\tilde{\theta} - \theta_n(P_{n,1}) = \tilde{\theta} - \theta_n(P_{n,0}) + \theta_n(P_{n,0}) - \theta_n(P_{n,1}).$$

The difference on the left hand side and the first difference on the right hand side are $o_P(1)$ under $P_{n,1}$. This implies that $\|\theta_n(P_{n,0}) - \theta_n(P_{n,1})\| = o(1)$. ■

Proof of Theorem 4.1. Let us first show sufficiency. Suppose that either (a) or (b) in Lemma A.5 holds. Let us take

$$\hat{\sigma}_{LR}^2 = \frac{1}{n} \sum_{m=1}^{M_n} \sum_{i,j \in N_{n,m}} (X_{n,i} - \bar{X}_n)(X_{n,j} - \bar{X}_n).$$

Since $\sigma_{LR}^2 \leq c$ for all $n \geq 1$, we have

$$\bar{X}_n = \mathbf{E}[X_{n,i}] + o_P(1).$$

Hence,

$$\hat{\sigma}_{LR}^2 = \frac{1}{n} \sum_{m=1}^{M_n} \sum_{i,j \in N_{n,m}} (X_{n,i} - \mathbf{E}[X_{n,i}])(X_{n,j} - \mathbf{E}[X_{n,i}]) + o_P(1).$$

Note that

$$\sum_{m=1}^{M_n} \left(\frac{n_m}{n}\right)^2 \leq \sum_{m=1}^{M_n} \frac{n_m}{n} \leq 1. \tag{A.9}$$

Choose $P_n \in \mathcal{P}_n$. Then, under P_n ,

$$\begin{aligned} \mathbf{E} \left[(\hat{\sigma}_{LR}^2 - \sigma_{LR}^2)^2 \right] &= \frac{1}{n^2} \sum_{m=1}^{M_n} \sum_{i,j \in N_{n,m}} \sum_{i',j' \in N_{n,m}} \text{Cov} (X_{n,i} X_{n,j}, X_{n,i'} X_{n,j'}) \\ &\leq \frac{C'}{n^2} \sum_{m=1}^{M_n} n_m^2 = \frac{C'}{n^2} \left((n - n^*) + \sum_{m=1:n_m \geq 2}^{M_n} n_m^2 \right) \\ &= C' \left(\frac{n - n^*}{n^2} + C' \left(\frac{n^*}{n} \right)^2 \sum_{m=1:n_m \geq 2}^{M_n} \left(\frac{n_m}{n^*} \right)^2 \right), \end{aligned}$$

where $C' > 0$ is a constant that does not depend on n . By (A.9), the last term is $o(1)$, if either of the conditions (a) and (b) in Lemma A.5. Therefore, σ_{LR}^2 is consistently estimable in \mathcal{P}_n .

Now, let us show necessity. Suppose that both (a) and (b) in Lemma A.5 are violated. That is,

$$\limsup_{n \rightarrow \infty} \frac{n^*}{n} > 0 \quad \text{and} \quad \limsup_{n \rightarrow \infty} \sum_{m \geq 1: n_m \geq 2}^{M_n} \left(\frac{n_m}{n^*} \right)^2 > 0.$$

We fix σ^2 and show that σ_{LR}^2 is not consistently estimable in $\mathcal{P}_{n,\mathcal{N}}(\sigma^2)$. We choose $\delta_{n,m} > 0$ for each cluster m and n such that (A.2) holds for some $\bar{\delta} \in [-a, a] \setminus \{0\}$, $a < 1/2$, if $n_m \geq 2$. By (A.9), there exists a subsequence $\{n_k\} \subset \{n\}$ such that

$$\sum_{m=1}^{M_{n_k}} \left(\frac{n_{m,n_k}}{n_k} \right)^2 \rightarrow \tilde{c}_1 > 0 \quad \text{and} \quad \frac{n_k^*}{n_k} \rightarrow \tilde{c}_2, \tag{A.10}$$

for some constants $\tilde{c}_1, \tilde{c}_2 \in (0, 1]$. For simplicity, we fix this subsequence, and denote n_k by n .

We let Σ_n be the block diagonal $n \times n$ matrix whose m -th block is given by $\sigma^2((1 - \delta_{n,m})I_{n_m} + \delta_{n,m} \mathbf{1}_{n_m} \mathbf{1}_{n_m}^\top)$. We show that $\Phi(0, \Sigma_n) \prec \Phi(0, \sigma^2 I_n)$. First, we observe that by Lemma A.6(ii), $\log(d\Phi(0, \Sigma_n)/d\Phi(0, \sigma^2 I_n))(X_n)$ is uniformly tight under $\Phi(0, \sigma^2 I_n)$. Furthermore, we find that by Prohorov’s Theorem, there exists a subsequence $\{n_k\}$ of $\{n\}$ such that the sequence $\log(d\Phi(0, \Sigma_{n_k})/d\Phi(0, \sigma^2 I_{n_k}))(X_n)$ weakly converges. Let W be a random variable whose distribution is identical to the weak limit. By the Continuous Mapping Theorem, we have

$$\frac{d\Phi(0, \Sigma_{n_k})}{d\Phi(0, \sigma^2 I_{n_k})}(X_{n_k}) \rightarrow_d e^W,$$

along the subsequence $\{n_k\}$. Note that $\mathbf{E}[(d\Phi(0, \Sigma_{n_k})/d\Phi(0, \sigma^2 I_{n_k}))(X_{n_k})] = 1$, where the expectation is under $\Phi(0, \sigma^2 I_{n_k})$. By Lemma A.6(i), $(d\Phi(0, \Sigma_{n_k})/d\Phi(0, \sigma^2 I_{n_k}))(X_{n_k})$ is uniformly integrable under $\Phi(0, \sigma^2 I_{n_k})$. Hence, we find that $\mathbf{E}e^W = 1$. By Le Cam’s First Lemma (e.g., van der Vaart, 1998, Lemma 6.4), we conclude that $\Phi(0, \Sigma_{n_k}) \prec \Phi(0, \sigma^2 I_{n_k})$.

On the other hand, note that the difference between the long-run variances under $\Phi(0, \Sigma_n)$ and under $\Phi(0, \sigma^2 I_n)$ is given by

$$\frac{\sigma^2}{n} \sum_{m=1:n_m \geq 2}^{M_n} \mathbf{1}_{n_m}^\top A_{n,m} \mathbf{1}_{n_m} = \sigma^2 \bar{\delta} \sum_{m=1:n_m \geq 2}^{M_n} \frac{n_m(n_m - 1)}{nn^*} = \sigma^2 \bar{\delta} \sum_{m=1}^{M_n} \frac{n_m^2}{nn^*} - \sigma^2 \bar{\delta} \sum_{m=1:n_m \geq 2}^{M_n} \frac{n_m}{nn^*},$$

where $A_{n,m} = \delta_{n,m} \mathbf{1}_{n_m} \mathbf{1}_{n_m}^\top - \delta_{n,m} I_{n_m}$. We rewrite the last term as

$$\begin{aligned} \sigma^2 \bar{\delta} \sum_{m=1:n_m \geq 2}^{M_n} \frac{n_m^2}{nn^*} - \sigma^2 \bar{\delta} \sum_{m=1:n_m \geq 2}^{M_n} \frac{n_m}{nn^*} &= \bar{\delta} \sum_{m=1:n_m \geq 2}^{M_n} \frac{n_m^2}{nn^*} + o(1) \\ &= \frac{\sigma^2 \bar{\delta} n^*}{n} \sum_{m=1:n_m \geq 2}^{M_n} \left(\frac{n_m}{n^*} \right)^2 + o(1) \rightarrow \sigma^2 \bar{\delta} \tilde{c}_1 \tilde{c}_2 \neq 0, \end{aligned}$$

as $n \rightarrow \infty$, where the last convergence is due to (A.10). By Lemma 4.1, we conclude that σ_{LR}^2 is not consistently estimable in $\mathcal{P}_{n,\mathcal{N}}(\sigma^2)$. Hence, it is not consistently estimable in \mathcal{P}_n either. ■

Proof of Corollary 4.1. Let us show sufficiency. First suppose that \mathcal{M}_n consists of negligible clusters, so that $\max_{1 \leq m \leq M_n} n_m/n \rightarrow 0$, as $n \rightarrow \infty$. From (A.9), this implies that

$$\sum_{m=1: n_m \geq 2}^{M_n} \left(\frac{n_m}{n^*}\right)^2 \leq \left(\frac{n}{n^*}\right) \max_{1 \leq m \leq M_n} \frac{n_m}{n} \rightarrow 0,$$

as $n \rightarrow \infty$. Therefore, σ_{LR}^2 is consistently estimable in \mathcal{P}_n by Theorem 4.1.

Conversely, suppose that for some $\epsilon > 0$,

$$\limsup_{n \rightarrow \infty} \max_{1 \leq m \leq M_n} \frac{n_m}{n} > \epsilon.$$

Then there exist subsequences $\{n_k\} \subset \{n\}$ and $\{n_{m(n_k)}\} \subset \{n_{m(n)}\}_{n \geq 1}$, $m(n) \in \{1, \dots, M_n\}$, such that

$$\lim_{k \rightarrow \infty} \frac{n_{m(n_k)}}{n_k} > \epsilon.$$

This implies that

$$\limsup_{n \rightarrow \infty} \sum_{m=1: n_m \geq 2}^{M_n} \left(\frac{n_m}{n^*}\right)^2 > 0.$$

Hence, σ_{LR}^2 is not consistently estimable in \mathcal{P}_n by Theorem 4.1. ■

Proof of Corollary 4.2. (i) Let $N_o \subset \{1, \dots, n\}$ be the set of nodes in a clique with size n_C . Take \mathcal{M}_n to be the cluster structure such that there is only one non-singleton cluster that is N_o . It suffices to show that σ_{LR}^2 is not consistently estimable in $\mathcal{P}_{n, \mathcal{N}}(\sigma^2)$ with the cluster structure \mathcal{M}_n . Note that

$$\sum_{m=1: n_m \geq 2}^{M_n} \left(\frac{n_m}{n^*}\right)^2 = 1,$$

because we have only one non-singleton cluster in \mathcal{M}_n . By Theorem 4.1, σ_{LR}^2 is not consistently estimable in $\mathcal{P}_{n, \mathcal{N}}(\sigma^2)$ with any fixed $\sigma^2 > 0$.

(ii) Let us define $\bar{N}(i) = \{j \in N_n : ij \in \bar{E}_n\}$, where $\bar{E}_n = E_n \cup \{ii : i \in N_n\}$, and consider

$$\hat{\sigma}_{LR}^2 = \frac{1}{n} \sum_{i=1}^n \sum_{j \in \bar{N}(i)} (X_{n,i} - \bar{X}_n)(X_{n,j} - \bar{X}_n).$$

By rearranging terms, we can write

$$\begin{aligned} \hat{\sigma}_{LR}^2 - \sigma_{LR}^2 &= \frac{1}{n} \sum_{i=1}^n \sum_{j: ij \in \bar{E}_n} (X_{n,i}X_{n,j} - \mathbf{E}[X_{n,i}X_{n,j}]) - 2(\bar{X}_n - \mathbf{E}[X_{n,i}]) \frac{1}{n} \sum_{j \in \bar{N}(i)} X_{n,j} \\ &\quad - 2\mathbf{E}[X_{n,i}] \frac{1}{n} \sum_{i=1}^n \sum_{j \in \bar{N}(i)} (X_{n,j} - \mathbf{E}X_{n,j}) + \bar{X}_n^2 - (\mathbf{E}[X_{n,i}])^2 \frac{1}{n} \sum_{i=1}^n |\bar{N}(i)|. \end{aligned} \tag{A.11}$$

We can write the squared L^2 norm of the leading term on the right hand side as

$$\frac{1}{n^2} \sum_{i_1=1}^n \sum_{i_2=1}^n \sum_{j_1: i_1j_1 \in \bar{E}_n} \sum_{j_2: i_2j_2 \in \bar{E}_n} \text{Cov}(X_{n,i_1}X_{n,j_1}, X_{n,i_2}X_{n,j_2}).$$

If $\{i_1, j_1\}$ and $\{i_2, j_2\}$ are not adjacent in G_n , the covariance above is zero by the dependency graph assumption. The number of the terms in the above sum such that $\{i_1, j_1\}$ and $\{i_2, j_2\}$ are adjacent in G_n is of the order $O(nd_{mx}^2 d_{av}) = o(n^2)$. The last rate comes from our assumption that $\lim_{n \rightarrow \infty} d_{mx}^2 d_{av}/n = 0$. Therefore, the leading term on the right hand side of (A.11) is $o_p(1)$. Similarly, we can show that the remainder terms are $o_p(1)$. Hence, $\hat{\sigma}_{LR}^2$ is a consistent estimator of σ_{LR}^2 . ■

References

Aronow, P., Samii, C., 2017. Estimating average causal effects under general interference, with application to a social network experiment. *Ann. Appl. Stat.* 11 (4), 1912–1947.
 Bahadur, R.R., Savage, L.J., 1956. The nonexistence of certain statistical procedures in nonparametric problems. *Ann. Math. Stat.* 27, 1115–1122.
 Bakirov, N.K., Szekeley, G.J., 2005. Student’s t -test for Gaussian scale mixtures. *Zap. Nauchnyh Semin. POMI* 328, 5–19.
 Bertanha, M., Moreira, M.J., 2020. Impossible inference in econometrics: Theory and applications. *J. Econometrics* 218, 247–270.
 Bester, A.C., Conley, T.G., Hansen, C.B., 2011. Inference with dependent data using cluster covariance estimators. *J. Econometrics* 165, 137–151.
 Cameron, A.C., Gelbach, J.B., Miller, D.L., 2008. Bootstrap based improvements for inference with clustered errors. *Rev. Econ. Stat.* 90, 414–427.
 Canay, I.A., Romano, J.P., Shaikh, A.M., 2017. Randomization tests under an approximate symmetry assumption. *Econometrica* 85, 1013–1030.
 Canay, I.A., Santos, A., Shaikh, A.M., 2021. The wild bootstrap with a “small” number of “large” clusters. *Rev. Econ. Stat.* 103, 346–363.
 Canen, N., Schwartz, J., Song, K., 2020. Estimating local interactions among many agents who observe their neighbors. *Quant. Econ.* 11, 346–363.
 Conley, T.G., Taber, C.R., 2011. Inference with “difference in differences” with a small number of policy changes. *Rev. Econ. Stat.* 93, 113–125.

- Djogbenou, A.A., MacKinnon, J.G., Nielsen, M.Ø., 2019. Asymptotic theory and wild bootstrap inference with clustered errors. *J. Econometrics* 212, 393–412.
- Donald, S.G., Lang, K., 2007. Inference with difference-in-difference and other panel data. *Rev. Econ. Stat.* 89, 221–233.
- Dufour, J.M., 1997. Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica* 65, 1365–1387.
- Hagemann, A., 2019. Placebo inference on treatment effects when the number of clusters is small. *J. Econometrics* 213, 190–209.
- Hansen, B.E., Lee, S., 2019. Asymptotic theory for clustered samples. *J. Econometrics* 210, 268–290.
- He, X., Song, K., 2022. Measuring diffusion over a large network. [arXiv:1812.04195v4](https://arxiv.org/abs/1812.04195v4) [stat.ME].
- Ibragimov, I.A., Has'minskii, R.Z., 1981. *Statistical Estimation: Asymptotic Theory*. Springer Science+Business Media, New York.
- Ibragimov, R., Müller, U.K., 2010. T-statistic based correlation and heterogeneity robust inference. *J. Bus. Econom. Statist.* 28, 453–468.
- Ibragimov, R., Müller, U.K., 2016. Inference with few heterogeneous clusters. *Rev. Econ. Stat.* 98, 83–96.
- Imbens, G.W., Wooldridge, J.M., 2009. Recent developments in the econometrics of program evaluation. *J. Econ. Lit.* 47, 5–86.
- Kiefer, N.M., Vogelsang, T.J., 2002. Heteroskedasticity-autocorrelation robust standard errors using the Bartlett Kernel without truncation. *Econometrica* 70, 2093–2095.
- Kojevnikov, D., Marmer, V., Song, K., 2021. Limit theorems for network dependent random variables. *J. Econometrics* 222, 882–908.
- LeCam, L., 1960. Locally asymptotically normal families of distributions. *Univ. Calif. Publ. Stat.* 3, 37–98.
- LeCam, L., Schwartz, L., 1960. A necessary and sufficient condition for the existence of consistent estimates. *Ann. Math. Stat.* 31, 140–150.
- Lehmann, E.L., Romano, J.P., 2005. *Testing Statistical Hypotheses*. Springer, New York.
- Leung, M.P., 2020. Treatment and spillover effects under network interference. *Rev. Econ. Stat.* 102, 368–380.
- Leung, M.P., 2021. Dependence-robust inference using resampled statistics. *J. Appl. Econometrics* 37, 270–285.
- MacKinnon, J.G., Webb, M.D., 2017. Wild bootstrap inference for wildly different cluster sizes. *J. Appl. Econometrics* 32, 233–254.
- MacKinnon, J.G., Webb, M.D., 2020. Randomization inference for difference-in-differences with few treated clusters. *J. Econometrics* 218, 435–450.
- Menzel, K., 2021. Bootstrap with cluster-dependence in two or more dimensions. *Econometrica* 89, 2143–2188.
- Pfanzagl, J., 1994. *Parametric Statistical Theory*. De Gruyter, Berlin.
- Pötscher, B.M., 2002. Lower risk bounds and properties of confidence sets for ill-posed estimation problems with applications to spectral density and persistence estimation, unit roots, and estimation of long memory parameters. *Econometrica* 70, 1035–1065.
- Romano, J.P., 2004. On non-parametric testing, the uniform behaviour of the *t*-test, and related problems. *Scand. J. Stat.* 31, 567–584.
- Roth, J., Sant'Anna, P.H.C., Bilinski, A., Poe, J., 2022. What's trending in difference-in-differences? A synthesis of the recent econometrics literature. [arXiv:2201.01194v2](https://arxiv.org/abs/2201.01194v2) [econ.EM].
- Song, K., 2016. Ordering-free inference from locally dependent data. [arXiv:1604.00447v1](https://arxiv.org/abs/1604.00447v1) [stat.ME].
- Song, K., 2018. Measuring the graph concordance of locally dependent observations. *Rev. Econ. Stat.* 100, 535–549.
- Stein, C., 1986. Approximate computation of expectations. *Lect. Notes-Monogr. Ser.* 7, i–164.
- Sun, Y., 2014. Fixed-smoothing asymptotics in a two-step generalized method of moments framework. *Econometrica* 82, 2327–2370.
- van der Vaart, A.W., 1998. *Asymptotic Statistics*. Cambridge University Press, New York, USA.