Econometrics **March 31, 2024**

Topic 5: Cluster-Robust Standard Errors

by Sai Zhang

Key points: The validity of Two-Way Cluster-Robust (TWCR) standard errors

Disclaimer: *This note is compiled by Sai Zhang.*

5.1 One-Way Clustering

First, consider the case of one-way clustering. The linear model with one-way clustering

$$y_{ig} = \mathbf{x}_{ig}\boldsymbol{\beta} + u_{ig}$$

where i denotes the ith of the N individuals in the sample, j denotes the gth of the G clusters, assume that

- $\mathbb{E}\left[u_{ig} \mid \mathbf{x}_{ig}\right] = 0$
- error independence across clusters: for $i \neq j$

$$\mathbb{E}\left[u_{ig}u_{jg'} \mid \mathbf{x}_{ig}, \mathbf{x}_{jg'}\right] = 0 \tag{5.1}$$

  unless $g = g'$, that is, errors for individuals within the same cluster may be correlated.

Grouping observations by cluster, get

$$\mathbf{y}_g = \mathbf{X}_g\boldsymbol{\beta} + \mathbf{u}$$

where $\mathbf{X}_g$ has dimension $N_g \times K$ and $\mathbf{y}_g$ has dimension $N_g \times 1$, with N_g observations in cluster g. Stacking over cluster, get the matrix form of the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

with $\mathbf{y}, \mathbf{u}$ being $N \times 1$ vectors, $\mathbf{X}$ being an $N \times K$ matrix. OLS estimator gives

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \left(\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{X}_g\right)^{-1}\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{y}_g \tag{5.2}$$

then, by CLT, we have that $\sqrt{G}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Sigma})$ where the variance matrix of the limit normal distribution $\boldsymbol{\Sigma}$ is

$$\left(\lim_{G\to\infty}\frac{1}{G}\sum_{g=1}^{G}\mathbf{E}\left[\mathbf{X}_g'\mathbf{X}_g\right]\right)^{-1}\left(\lim_{G\to\infty}\frac{1}{G}\sum_{g=1}^{G}\mathbf{E}\left[\mathbf{X}_g'\mathbf{u}_g'\mathbf{u}_g\mathbf{X}_g\right]\right)\times\left(\lim_{G\to\infty}\frac{1}{G}\sum_{g=1}^{G}\mathbf{E}\left[\mathbf{X}_g'\mathbf{X}_g\right]\right)^{-1} \tag{5.3}$$

If the primary source of clustering is due to group-level common shocks, a useful approximation is that for the jth regressor, the default OLS variance estimate based on $s^2(\mathbf{X}'\mathbf{X})^{-1}$ should be inflated by $\tau_j \simeq 1 + \rho_{x_j}\rho_u\left(\overline{N}_g - 1\right)$, where

- s is the estimated standard deviation of the error

- ρ_{x_j} is a measure of within-cluster correlation of x_j
- ρ_u is the within-cluster error correlation
- $\overline{N}_g$ is the average cluster size

It's easy to see the τ_j can be large even with small ρ_u (Kloek, 1981; Scott and Holt, 1982; Moulton, 1990). If assume the model for the cluster error variance matrices $\mathbf{\Omega}_g = \mathbb{V}\left[\mathbf{u}_g \mid \mathbf{X}_g\right] = \mathbb{E}\left[\mathbf{u}_g\mathbf{u}_g' \mid \mathbf{X}_g\right]$, and there is a consistent estimate $\hat{\mathbf{\Omega}}_g$ of $\mathbf{\Omega}_g$, we can estimate $\mathbb{E}\left[\mathbf{X}_g'\mathbf{u}_g\mathbf{u}_g'\mathbf{X}_g\right] = \mathbb{E}\left[\mathbf{X}_g'\mathbf{\Omega}_g\mathbf{X}_g\right]$ via GLS.

Cluster-robust variance matrix estimate consider

$$\hat{\mathbb{V}}\left[\hat{\boldsymbol{\beta}}\right] = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{g=1}^{G}\mathbf{X}_g'\hat{\mathbf{u}}_g\hat{\mathbf{u}}_g'\mathbf{X}_g\right)(\mathbf{X}'\mathbf{X})^{-1} \tag{5.4}$$

where $\hat{\mathbf{u}}_g = \mathbf{y}_g - \mathbf{X}_g\hat{\boldsymbol{\beta}}$. This estimate is consistent if

$$G^{-1}\sum_{g=1}^{G}\mathbf{X}_g'\hat{\mathbf{u}}_g\hat{\mathbf{u}}_g'\mathbf{X}_g - G^{-1}\sum_{g=1}^{G}\mathbb{E}\left[\mathbf{X}_g'\mathbf{u}_g\mathbf{u}_g'\mathbf{X}_g\right] \xrightarrow{\text{P}} \mathbf{0}$$

as $G \to \infty$. An informal presentation of Eq.(5.4) is to rewrite the central matrix as

$$\hat{\mathbf{B}} = \sum_{g=1}^{G}\mathbf{X}_g'\hat{\mathbf{u}}_g\hat{\mathbf{u}}_g'\mathbf{X}_g = \mathbf{X}'\begin{bmatrix}\hat{\mathbf{u}}_1\hat{\mathbf{u}}_1' & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{u}}_2\hat{\mathbf{u}}_2' & & \vdots \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & & \hat{\mathbf{u}}_G\hat{\mathbf{u}}_G'\end{bmatrix}\mathbf{X} = \mathbf{X}'\left(\hat{\mathbf{u}}\hat{\mathbf{u}}' \otimes \mathbf{S}^G\right)\mathbf{X} \tag{5.5}$$

where $\otimes$ denotes element-wise multiplication. The (p, q)th element of this matrix is

$$\sum_{i=1}^{N}\sum_{j=1}^{N}x_{ia}x_{jb}\hat{u}_i\hat{u}_j \cdot \mathbf{1}\left(i, j \text{ in the same cluster}\right)$$

with $\hat{u}_i = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}$.

$\mathbf{S}^G$ is an $N \times N$ indicator matrix with $\mathbf{S}_{ij}^G = 1$ only if the ith and jth observation belong to the same cluster: it zeros out a large amount of $\hat{\mathbf{u}}\hat{\mathbf{u}}'$ (asymptotically equivalently, $\mathbf{u}\mathbf{u}'$), specifically, only $\sum_{g=1}^{G}N_g^2$ out of $N^2 = \left(\sum_{g=1}^{G}N_g\right)^2$ terms are not zero (sub-matrices on the diagonal). Asymptotically

- for fixed N_g, $\frac{1}{N^2}\sum_{g=1}^{G}N_g^2 \xrightarrow{G\to\infty} 0$
- for balanced clusters $N_g = N/G$, $\frac{1}{N^2}\sum_{g=1}^{G}N_g^2 = \frac{1}{G} \xrightarrow{G\to\infty} 0$

A strand of literature popularizes this method:

- Liang and Zeger (1986): in a generalized estimatin equations setting
- Arellano (1987): fixed effects estimator in linear panel models
- Hansen (2007): asymptotic theory for panel data where $T \to \infty$ in addition to $N \to \infty$ (or $N_g \to \infty$ in addition to $G \to \infty$ in the notation above).

5.2 Two-Way Clustering

Now, consider the case of two-way clustering,

$$y_{i,gh} = \mathbf{x}'_{i,gh}\boldsymbol{\beta} + u$$

where each observation may belong to **two** dimension of groups: group $g \in \{1, \cdots, G\}$ and $h \in \{1, \cdots, H\}$, and for $i \neq j$

$$\mathbb{E}\left[u_{i,gh}u_{j,g'h'} \mid \mathbf{x}_{i,gh}, \mathbf{j}, \mathbf{g}'\mathbf{h}'\right] = 0 \tag{5.6}$$

unless $g = g'$ or $h = h'$, that is, errors for individuals within the same group (along either g or h) may be correlated.

Cluster-robust variance matrix estimate extending the one-way clustering case, keep elements of $\hat{\mathbf{u}}\hat{\mathbf{u}}'$ where the ith and jth observations share a cluster in **any** dimension, then similar to Eq.(5.5)

$$\hat{\mathbf{B}} = \mathbf{X}'\left(\hat{\mathbf{u}}\hat{\mathbf{u}}' \otimes \mathbf{S}^{GH}\right)\mathbf{X} \tag{5.7}$$

here $\mathbf{S}^{GH}$ is an $N \times N$ indicator matrix with $\mathbf{S}^{GH}_{ij} = 1$ only if the ith and jth observation share any cluster, the (p, q)th element of this matrix is

$$\sum_{i=1}^{N}\sum_{i=1}^{N} x_{ia}x_{jb}\hat{u}_i\hat{u}_j \cdot \mathbf{1}\left(i, j \text{ share any cluster}\right)$$

$\hat{\mathbf{B}}$ can also be presented in one-way cluster-robust fashion:

$$\hat{\mathbf{B}} = \mathbf{X}'\left(\hat{\mathbf{u}}\hat{\mathbf{u}}' \otimes \mathbf{S}^{GH}\right)\mathbf{X} = \mathbf{X}'\left(\hat{\mathbf{u}}\hat{\mathbf{u}}' \otimes \mathbf{S}^{G}\right)\mathbf{X} + \mathbf{X}'\left(\hat{\mathbf{u}}\hat{\mathbf{u}}' \otimes \mathbf{S}^{H}\right)\mathbf{X} - \mathbf{X}'\left(\hat{\mathbf{u}}\hat{\mathbf{u}}' \otimes \mathbf{S}^{G\cap H}\right)\mathbf{X} \tag{5.8}$$

where $\mathbf{G}^{GH} = \mathbf{G}^{G} + \mathbf{G}^{H} - \mathbf{G}^{G\cap H}$, with

- $\mathbf{G}^{G}$: $\mathbf{G}^{G}_{ij} = 1$ only if the ith and jth observation belong to the same cluster $g \in \{1, 2, \cdots, G\}$
- $\mathbf{G}^{H}$: $\mathbf{G}^{H}_{ij} = 1$ only if the ith and jth observation belong to the same cluster $h \in \{1, 2, \cdots, H\}$
- $\mathbf{G}^{G\cap H}$: $\mathbf{G}^{G\cap H}_{ij} = 1$ only if the ith and jth observation belong to **both** the same cluster $g \in \{1, 2, \cdots, G\}$ and the same cluster $h \in \{1, 2, \cdots, H\}$

then, similar to one-way clustering case,

$$\begin{aligned}\hat{\mathbb{V}}\left[\hat{\boldsymbol{\beta}}\right] = &\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\hat{\mathbf{u}}\hat{\mathbf{u}}' \otimes \mathbf{S}^{G}\right)\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\ &+ \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\hat{\mathbf{u}}\hat{\mathbf{u}}' \otimes \mathbf{S}^{H}\right)\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\ &- \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\hat{\mathbf{u}}\hat{\mathbf{u}}' \otimes \mathbf{S}^{G\cap H}\right)\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\end{aligned} \tag{5.9}$$

that is,

$$\hat{\mathbb{V}}\left[\hat{\boldsymbol{\beta}}\right] = \hat{\mathbb{V}}^{G}\left[\hat{\boldsymbol{\beta}}\right] + \hat{\mathbb{V}}^{H}\left[\hat{\boldsymbol{\beta}}\right] - \hat{\mathbb{V}}^{G\cap H}\left[\hat{\boldsymbol{\beta}}\right] \tag{5.10}$$

each of Eq.(5.10) can be separately computed by OLS of $\mathbf{y}$ on $\mathbf{X}$, with variance matrix estimates $\hat{\mathbb{V}}$ based on

i clustering on $g \in \{1, 2, \cdots, G\}$
ii clustering on $h \in \{1, 2, \cdots, H\}$
iii clustering on $(g, h) \in \{(1, 1), \cdots, (G, H)\}$

Practical considerations It is required to know what *ways* will be potentially important for clustering, which can be tested via checking the dimension of correlations in the errors. There are several ways to test

- estimate sample covariances of $\mathbf{X}'\hat{\mathbf{u}}$ within dimensions, test the null that the **average** of such covariances is 0: rejecting this null is sufficient (not necessary) to reject the null of no clustering (White, 1980)
- for **small samples**, Eq. (5.4) is baised downwards. This is corrected (in Stata) by replacing $\hat{\mathbf{u}}_g$ with $\sqrt{c}\hat{\mathbf{u}}_g$, where $c = \frac{G}{G-1}\frac{N-1}{N-K} \simeq \frac{G}{G-1}$. For two-way clustering (Eq. 5.8), there are 2 ways of correction:
  - choose correction terms for each of the 3 components:

    $$c_1 = \frac{G}{G-1}\frac{N-1}{N-K}, c_2 = \frac{H}{H-1}\frac{N-1}{N-K}, c_3 = \frac{I}{I-1}\frac{N-1}{N-K}$$

    with I being the number of unique clusters determined by $G \cap H$
  - choose a constant terms for all components:

    $$c = \frac{J}{J-1}\frac{N-1}{N-K}$$

    with $J = \min(G, H)$
- **Var-cov matrix not positive-semidefinite**: $\hat{\mathbb{V}}\left[\hat{\boldsymbol{\beta}}\right]$ might have negative elements on the diagonal (Eq. 5.10), informly, this is more likely to arise when clustering is done over the same groups as the fixed effects. One way to address this issue is using *eigendecomposition* technique:

  $$\hat{\mathbb{V}}\left[\hat{\boldsymbol{\beta}}\right] = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}'$$

  where
  - $\mathbf{U}$ containing the eigenvectors of $\hat{\mathbf{V}}$
  - $\boldsymbol{\Lambda} = \text{diag}\left[\lambda_1, \cdots, \lambda_d\right]$ contains the eigenvalues of $\hat{\mathbf{V}}$

  then create $\boldsymbol{\Lambda}^+ = \text{diag}\left[\lambda_1^+, \cdots, \lambda_d^+\right]$ with $\lambda_j^+ = \max\left(0, \lambda_j\right)$ and use $\hat{\mathbf{V}}^+\left[\hat{\boldsymbol{\beta}}\right] = \mathbf{U}\boldsymbol{\Lambda}^+\mathbf{U}'$ as the estimate

5.3 Multiway Clustering

Cameron et al. (2011) extended the framework[1] to allow clustering in D dimensions, then we can do the following reframing

- G_d: the number of clusters in dimension $d \in \{1, 2, \cdots, D\}$
- $D-$vector $\boldsymbol{\delta}_i = \boldsymbol{\delta}(i)$, with funciton $\boldsymbol{\delta} : \{1, 2, \cdots, N\} \to \bigtimes_{d=1}^{D}\{1, 2, \cdots, G_d\}$ lists the cluster membership in each dimension of each observation

then we have

$$\mathbf{1}\left[i, j \text{ shares a clustser}\right] = 1 \Leftrightarrow \delta_{id} = \delta_{jd}$$

for some $d \in \{1, 2, \cdots, D\}$, where δ_{id} denotes the dth element of $\boldsymbol{\delta}_i$. Also

- $D-$vector $\mathbf{r}$: define the set

  $$R \equiv \{\mathbf{r} : r_d \in \{0, 1\}, d = 1, 2, \cdots, D, \mathbf{r} \neq \mathbf{0}\}$$

  elements of the set R can be used to index all cases where 2 observations share a cluster in at least one dimension. Define the function

  $$\mathbf{I}_\mathbf{r}(i, j) \equiv \mathbf{1}\left[r_d\delta_{id} = r_d\delta_{jd}, \forall d\right]$$

[1] Also proposed by Thompson (2011).

which indicates whether observations i and j have identical cluster menbership for **all** dimensions d s.t. $r_d = 1$. Then we have a *aggregate* identifier

$$\mathbf{I}(i,j) = 1 \Leftrightarrow \mathbf{I}_{\mathbf{r}}(i,j) = 1 \text{ for some } \mathbf{r} \in R$$

i.e., 2 observations share **at least** one dimension.

The define the $2^D - 1$ matrices

$$\tilde{\mathbf{B}}_{\mathbf{r}} \equiv \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbf{x}_i \mathbf{x}_j' \hat{u}_i \hat{u}_j \mathbf{I}_{\mathbf{r}}(i,j) \tag{5.11}$$

with $\mathbf{r} \in R$.

Var-cov matrix estimator consider, similarly, an estimator

$$\hat{\mathbb{V}}\left[\hat{\boldsymbol{\beta}}\right] = (\mathbf{X}'\mathbf{X})^{-1}\,\tilde{\mathbf{B}}\,(\mathbf{X}'\mathbf{X})^{-1} \equiv (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{\|\mathbf{r}\|=k,\mathbf{r}\in R} (-1)^{k+1}\tilde{\mathbf{B}}_r\right)(\mathbf{X}'\mathbf{X})^{-1} \tag{5.12}$$

where cases of clustering on an odd number of dimensions are added, those of clustering on an even number of dimensions are subtracted. Consider the case of $D = 3$,

$$\left(\tilde{\mathbf{B}}_{(1,0,0)} + \tilde{\mathbf{B}}_{(0,1,0)} + \tilde{\mathbf{B}}_{(0,0,1)}\right) - \left(\tilde{\mathbf{B}}_{(1,1,0)} + \tilde{\mathbf{B}}_{(1,0,1)} + \tilde{\mathbf{B}}_{(0,1,1)}\right) + \tilde{\mathbf{B}}_{(1,1,1)}$$

$\tilde{\mathbf{B}}$ is identical to $\hat{\mathbf{B}}$ defined analogically as in Eq.(5.8), since

- no observation pair with $\mathbf{I}(i,j) = 0$: this is immediate, since $\mathbf{I}(i,j) = 0 \Leftrightarrow \mathbf{I}_{\mathbf{r}}(i,j) = 0, \forall \mathbf{r}$
- the covariance term corresponding to each observation pair with $\mathbf{I}(i,j) = 1$ is included **exactly once** in $\tilde{\mathbf{B}}$: by inclusion-exclusion principle for set cardinality

$$\mathbf{I}(i,j) \Rightarrow \sum_{\|\mathbf{r}\|=k,\mathbf{r}\in R} (-1)^{k+1}\mathbf{I}_{\mathbf{r}}(i,j) = 1$$

Curse of dimensionality this could arise in a setting with **many dimensions** of clustering, and in which one or more dimensions have **few** clusters[2]. Cameron et al. (2011) suggested an ad-hoc rule of thumb for approximating sufficient numbers of clusters.

5.3.1 Non-linear Estimators

***m*-Estimators** Consider an m-estimator that solves

$$\sum_{i=1}^{N} \mathbf{h}_i\left(\hat{\boldsymbol{\theta}}\right) = \mathbf{0}$$

under standard assumptions, $\hat{\boldsymbol{\theta}}$ is asymptotically normal with estimated variance matrix

$$\hat{\mathbb{V}}\left[\hat{\boldsymbol{\theta}}\right] = \hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}'^{-1} \tag{5.13}$$

where $\hat{\mathbf{A}} = \sum_i \left.\frac{\partial \mathbf{h}_i}{\partial \boldsymbol{\theta}'}\right|_{\hat{\boldsymbol{\theta}}}$ and $\hat{\mathbf{B}}$ is an estimate of $\mathbb{V}\left[\sum_i \mathbf{h}_i\right]$.

[2]The square design (each dimension has the same number of clusters) with orthogonal dimensions has the **least** independence of observations.

- **one-way clustering** $\hat{\mathbf{B}} = \sum_{g=1}^{G} \hat{\mathbf{h}}_g \hat{\mathbf{h}}_g'$ where $\hat{\mathbf{h}}_g = \sum_{i=1}^{N_g} \hat{\mathbf{h}}_{ig}$, clustering may not lead to parameter inconsistency, depending on whether $\mathbb{E}[\mathbf{h}_i(\boldsymbol{\theta})] = \mathbf{0}$ with clustering
  - **population-averaged approach**: assum $\mathbf{E}\left[y_{ig} \mid \mathbf{x}_{ig}\right] = \Phi\left(\mathbf{x}_{ig}'\boldsymbol{\beta}\right)$
  - **random effects approach**: let $y_{ig} = 1$ if $y_{ig}^* > 0$ where $y_{ig}^* = \mathbf{x}_{ig}'\boldsymbol{\beta} + \epsilon_g + \epsilon_{ig}$, where
    * idiosyncratic error $\epsilon_{ig} \sim \mathcal{N}(0,1)$
    * cluster-specific error $\epsilon_g \sim \mathcal{N}(0, \sigma_g^2)$

    then we have the alternative moment condition

$$\mathbb{E}\left[y_{ig} \mid \mathbf{x}_{ig}\right] = \Phi\left(\frac{\mathbf{x}_{ig}'\boldsymbol{\beta}}{\sqrt{1+\sigma_g^2}}\right)$$

- **multiway clustering** replacing $\hat{u}_i \mathbf{x}_i$ in Eq.(5.11) with $\hat{\mathbf{h}}_i$, then we have

$$\hat{\mathbb{V}}\left[\hat{\boldsymbol{\theta}}\right] = \hat{\mathbf{A}}^{-1}\tilde{\mathbf{B}}\hat{\mathbf{A}}'^{-1}$$

  where

$$\hat{\mathbf{A}} = \sum_i \left.\frac{\partial \mathbf{h}_i}{\partial \boldsymbol{\theta}'}\right|_{\hat{\boldsymbol{\theta}}} \qquad \tilde{\mathbf{B}} = \sum_{\|\mathbf{r}\|=k, \mathbf{r}\in R} (-1)^{k+1}\tilde{\mathbf{B}}_r \qquad \tilde{\mathbf{B}}_r \equiv \sum_{i=1}^{N}\sum_{j=1}^{N} \hat{\mathbf{h}}_i \hat{\mathbf{h}}'_j \mathbb{I}_{\mathbf{r}}(i,j)$$

  with $\mathbf{r} \in R$[3].

GMM estimation Consider an example of over-identified models: linear two stage least squares with more instruments than endogenous regressors, we have

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}} \left(\sum_{i=1}^{N} \mathbf{h}_i(\boldsymbol{\theta})\right)' \mathbf{W} \left(\sum_{i=1}^{N} \mathbf{h}_i(\boldsymbol{\theta})\right)$$

where $\mathbf{W}$ is a symmetric positive definite weighting matrix. Under standard regularity conditions, $\hat{\boldsymbol{\theta}}$ is asymptotically normal, with estimated variance matrix

$$\hat{\mathbb{V}}\left[\hat{\boldsymbol{\theta}}\right] = \left(\hat{\mathbf{A}}'\mathbf{W}\hat{\mathbf{A}}\right)^{-1}\hat{\mathbf{A}}'\mathbf{W}\tilde{\mathbf{B}}\mathbf{W}\hat{\mathbf{A}}\left(\hat{\mathbf{A}}'\mathbf{W}\hat{\mathbf{A}}\right)^{-1}$$

again, $\hat{\mathbf{A}} = \sum_i \left.\frac{\partial \mathbf{h}_i}{\partial \boldsymbol{\theta}'}\right|_{\hat{\boldsymbol{\theta}}}$, and $\tilde{\mathbf{B}}$ is an estimate of $\mathbb{V}\left[\sum_i \mathbf{h}_i\right]$.

5.4 Menzel (2021): Asymptotic Gaussianity

One key of TWCR inference is the asymptotic Gaussianity, Menzel (2021) pointed out the potential non-Gaussianity of the limit distribution. Still, consider a random array (Y_{it}) indexed by two dimensions by $i = 1, \cdots, N$ and $t = 1, \cdots, T$. Clusters are sampled independently at random from an infinite population, but otherwise **unrestricted** in dependence within each row $\mathbf{Y}_{i\cdot} := (Y_{i1} \cdots, Y_{iT})$ and within each column $\mathbf{Y}_{\cdot t} := (Y_{1t}, \cdots, Y_{Nt})$.

[3]This multiway clustering can be implemented using several one-way clustered bootstraps. Each of the one-way cluster robust matrices is estimated by a pairs cluster bootstrap that resamples with replacement from the appropriate cluster dimension. They are then combined as if they had been estimated analytically (Cameron et al., 2011).

5.4.1 Distribution of Sample Average

First, consider

$$\overline{Y}_{NT} \coloneqq \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} Y_{it}$$

and approximate the asymptotic distribution regardless of whether, or what type of, cluster-dependence is present.

3 scenarios of the array (Y_{it})

- **no cluster-dependence**: (Y_{it}) are mutually independent, CLT at a rate of $(NT)^{-1/2}$ applies (under regularity conditions)
- **correlation within clusters**: the convergence rate of (Y_{it}) is determined by the number of relevant clusters
- **non-separable models of heterogeneity (dependence with clusters, even uncorrelated)**[4]: The asymptotic behavior is non-standard

Consider 2 examples:

- **Additive factor model**

  $$Y_{it} = \mu + \alpha_i + \gamma_t + \epsilon_{it}$$

  where μ is a constant, and $\alpha_i, \gamma_i, \epsilon_{it}$ are zero-mean i.i.d. random variables for $i = 1, \cdots, N$ and $t = 1, \cdots, T$ with bounded second moments, and $N = T$. Based on a standard central limit theory, we have

  – in the non-degenerate case with $\text{Var}(\alpha_i) > 0$ or $\gamma_t > 0$, the sample distribution

  $$\sqrt{N}\left(\overline{Y}_{NT} - \mathbb{E}\left[Y_{it}\right]\right) \xrightarrow{d} \mathcal{N}\left(0, \text{Var}(\alpha_i) + \text{Var}(\gamma_t)\right)$$

  – in the degenerate case of no clustering with $\text{Var}(\alpha_i) = \text{Var}(\gamma_t) = 0$, the sample distribution

  $$\sqrt{NT}\left(\overline{Y}_{NT} - \mathbb{E}\left[Y_{it}\right]\right) \xrightarrow{d} \mathcal{N}\left(0, \text{Var}(\epsilon_{it})\right)$$

  if marginal distributions of $\alpha_i, \gamma_t, \epsilon_{it}$ are known, we can simulate from the joint distribution of (Y_{it}) by sampling the individual components at random, a bootstrap procedure would be consistent. If **unknown**, consider estimators

  $$\begin{aligned}
  \hat{\alpha}_i &\coloneqq \frac{1}{T}\sum_{t=1}^{T}\left(Y_{it} - \overline{Y}_{NT}\right) = \alpha_i + \frac{1}{T}\sum_{t=1}^{T}\left(\epsilon_{it} - \overline{\epsilon}_{NT}\right) \\
  \hat{\gamma}_t &\coloneqq \frac{1}{N}\sum_{i=1}^{N}\left(Y_{it} - \overline{Y}_{NT}\right) = \gamma_t + \frac{1}{N}\sum_{i=1}^{N}\left(\epsilon_{it} - \overline{\epsilon}_{NT}\right) \\
  \hat{\epsilon}_{it} &\coloneqq Y_{it} - \overline{Y}_{NT} - \hat{\alpha}_i - \hat{\gamma}_t
  \end{aligned}$$

  then use these empirical distributions for estimation and form a bootstrap sample

  $$Y^*_{it} \coloneqq \overline{Y}_{NT} + \alpha^*_i + \gamma^*_t + \epsilon^*_{it}$$

[4] This is specific to clustering in 2 or more dimensions.

by drawing from these estimators and obtain $\overline{Y}^*_{NT} \coloneqq \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} Y^*_{it}$, and verify the conditional variances of the bootstrap distribution given the sample:

$$\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\alpha}_i - \frac{1}{N}\sum_{j=1}^{N}\hat{\alpha}_j\right)^2 - \left[\text{Var}(\alpha_i) + \frac{\text{Var}(\epsilon_{it})}{T}\right] \xrightarrow{p} 0$$

$$\frac{1}{T}\sum_{i=1}^{N}\left(\hat{\gamma}_t - \frac{1}{T}\sum_{s=1}^{T}\hat{\gamma}_t\right)^2 - \left[\text{Var}(\gamma_t) + \frac{\text{Var}(\epsilon_{it})}{N}\right] \xrightarrow{p} 0$$

then the bootstrap distribution is

– in the non-degenerate case,

$$\sqrt{N}\left(\overline{Y}^*_{NT} - \overline{Y}_{NT}\right) \xrightarrow{d} \mathcal{N}\left(0, \text{Var}(\alpha_i) + \text{Var}(\gamma_t)\right)$$

the estimation error $\hat{\alpha}_i$ does **NOT** affect the asymptotic variance.

– in the degenerate case,

$$\sqrt{NT}\left(\overline{Y}^*_{NT} - \overline{Y}_{NT}\right) \xrightarrow{d} \mathcal{N}\left(0, 3\text{Var}(\epsilon_{it})\right)$$

asymptotically overestimates the variance of the sampling distribution, leading to inconsistency of this naive bootstrapping procedure.

- **Non-Gaussian limit distribution**

$$Y_{it} = \alpha_i\gamma_t + \epsilon_{it}$$

where $\alpha_i, \gamma_t, \epsilon_{it}$ are independently distributed with $\mathbb{E}[\epsilon_{it}] = 0, \text{Var}(\alpha_i) = \sigma^2_\alpha, \text{Var}(\gamma_t) = \sigma^2_\gamma, \text{Var}(\epsilon)_{it} = \sigma^2_\epsilon$. **If** $\mathbb{E}[\alpha_i] = \mathbb{E}\left[\gamma_t\right] = 0$, then CLT and Continuous Mapping Theorem (CMT) imply

$$\begin{aligned}\sqrt{NT}\cdot\overline{Y}_{NT} =& \frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\alpha_i\gamma_t + \epsilon_{it}\right)\\ =& \left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\alpha_i\right)\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\gamma_t\right) + \frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\epsilon_{it}\\ \xrightarrow{d}& \sigma_\alpha\sigma_\gamma Z_1 Z_2 + \sigma_\epsilon Z_3\end{aligned}$$

then even without correlation within clusters, non-separable heterogeneity can still generate dependence in 2nd or higher moments in the limiting distribution[5].

5.4.2 Menzel (2021)'s Bootstrap procedure

5.4.2.1 Notation

For the array $(Y_{it})_{i,t}$, denote

- $\mathbb{P}$: joint distribution of $(Y_{it})_{i,t}$

[5] 2 major issues arise:
- The limiting distribution needs **not** be Gaussian: plug-in asymptotic inference based on the normal distribution is invalid
- It only comes from two-or-more-dimension cluster dependence, not single-dimension cluster dependence.

- $\mathbb{P}_{NT}$: drifting DGP indexed by N, T
- $\mathbb{P}^*_{NT}$: bootstrap distribution for $\left(Y^*_{it}\right)$ given the realizations $(Y_{it} : i = 1, \cdots, N; t = 1, \cdots, T)$
- respective distributions $\mathbb{E}, \mathbb{E}_{NT}, \mathbb{E}^*_{NT}$

5.4.2.2 Inference: Sample Mean

First, consider the assumption of *separate exchangeability*

Assumption 5.4.1: Separate Exchangeability

- A **separately exchangeable** array is an infinite array $(Y_{it})_{i,t}$ such that for any integers $\tilde{N}, \tilde{T}$ and permutations $\pi_1 : \{1, \cdots, \tilde{N}\} \to \{1, \cdots, \tilde{N}\}$ and $\pi_2 : \{1, \cdots, \tilde{T}\} \to \{1, \cdots, \tilde{T}\}$, we have

$$\left(Y_{\pi_1(i),\pi_2(t)}\right)_{i,t} \overset{d}{=} (Y_{it})_{i,t}$$

  such an array is called **dissociated** if for any $N_0, T_0 \geq 1$, $(Y_{it})_{i=1,t=1}^{i=N_0,t=T_0}$ is independent of $(Y_{it})_{i>N_0,t>T_0}$.
- For dyadic data, consider the alternative assumption **jointly exchangeable** arrays $\left(Y_{ij}\right)_{i,j}$ satisfying

$$\left(Y_{\pi(i),\pi(j)}\right)_{i,j} \overset{d}{=} \left(Y_{ij}\right)_{i,j}$$

  for any permutation π on $\{1, \cdots, \tilde{N}\}$, in addition, $\left(Y_{ij}\right)_{i,j=1}^{N_0}$ is independent of $\left(Y_{ij}\right)_{i,j>N_0}$

This assumption can be interpreted as rows (and columns) corresponding to units that are drawn independently from a common population, where we then observe the joint outcome for every row-column pair, consider the requirements in the following applications

- **DiD/matched data**: the units corresponding to either dimension of the sample to represent independent draws from a common, infinite population
- **non-exhaustively matched data**: only observe joint outcomes for a posibly self-selected subset of unit pairs, sample selection should be (jointly or separately) exchangeable
- **U-/V-statistics**: the kernel $Y_{i_1,\cdots,i_D} \coloneqq h\left(X_{i_1}, \cdots, X_{i_D}\right)$ evaluated at i.i.d. observations $X_1, \cdots, X_N$ forms a dissociated, jointly exchangeable array
- **Network**: unlabeled[6] data implies finite exchangeability, the sampled graph has joint (*infinite*) exchangeability if it is a subgraph of an infinite graph

Directly from Assumption 5.4.1, any dissociated separately exchangeable array can be represented as

$$Y_{it} = f\left(\alpha_i, \gamma_t, \epsilon_{it}\right)$$

for some function $f(\cdot)$ where $\alpha_1, \cdots, \alpha_N$, $\gamma_1, \cdots, \gamma_T$, $\epsilon_{11}, \cdots, \epsilon_{NT}$ are mutually independent, uniformly distributed random variables.

Projection now, decompose the array $(Y_{it})_{i,t}$ as

$$Y_{it} = b + a_i + g_t + w_{it} \qquad \mathbb{E}\left[w_{it} \mid a_i, g_t\right] = 0$$

[6] *Unlabeled*: nodel identifiers do not carry any significance for the statistical model.

where a_i and g_t are mean-zero and mutually independent, s.t. the joint distribution of Y_{it} can then be expanded as

$$\begin{aligned} Y_{it} =& \mathbb{E}[Y_{it}] + (\mathbb{E}[Y_{it} \mid \alpha_i] - \mathbb{E}[Y_{it}]) + (\mathbb{E}[Y_{it} \mid \gamma_t] - \mathbb{E}[Y_{it}]) \\ &+ (\mathbb{E}[Y_{it} \mid \alpha_i, \gamma_t] - \mathbb{E}[Y_{it} \mid \alpha_i] - \mathbb{E}[Y_{it} \mid \gamma_t] + \mathbb{E}[Y_{it}]) + (Y_{it} - \mathbb{E}[Y_{it} \mid \alpha_i, \gamma_t]) \\ =:& b + a_i + g_t + v_{it} + e_{it} \end{aligned}$$

with

- $e_{it} = Y_{it} - \mathbb{E}[Y_{it} \mid \alpha_i, \gamma_t]$
- $a_i = \mathbb{E}[Y_{it} \mid \alpha_i] - \mathbb{E}[Y_{it}]$, $g_t = \mathbb{E}[Y_{it} \mid \gamma_t] - \mathbb{E}[Y_{it}]$
- $v_{it} = \mathbb{E}[Y_{it} \mid \alpha_i, \gamma_t] - \mathbb{E}[Y_{it} \mid \alpha_i] - \mathbb{E}[Y_{it} \mid \gamma_t] + \mathbb{E}[Y_{it}]$
- $b = \mathbb{E}[Y_{it}]$

here,

- temporal and cross-sectional units were drawn independently: $a_1, \cdots, a_N$ and $g_1, \cdots, g_T$ are independent of each other.
- by construction, $\mathbb{E}[e_{it} \mid a_i, g_t, v_{it}] = 0$, $\mathbb{E}[v_{it} \mid a_i] = \mathbb{E}[v_{it} \mid g_t] = 0$
- e_{it}, (a_i, g_t) and v_{it} are **uncorrelated**

then, rewrite the sample mean as

$$\begin{aligned} \hat{Y}_{NT} &= b + \overline{a}_N + \overline{g}_T + \overline{v}_{NT} + \overline{e}_{NT} \\ &:= b + \frac{1}{N}\sum_{i=1}^{N} a_i + \frac{1}{T}\sum_{t=1}^{T} g_t + \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} v_{it} + \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} e_{it} \end{aligned}$$

and the unconditional variances of the projections with

$$\sigma_a^2 := \mathrm{Var}(a_i) \qquad \sigma_g^2 := \mathrm{Var}(g_t) \qquad \sigma_v^2 := \mathrm{Var}(v_{it}) \qquad \sigma_e^2 := \mathrm{Var}(e_{it})$$

let $w_{it} := v_{it} + e_{it}$, and denote its variance by $\sigma_w^2 = \mathrm{Var}(w_{it})$. Then, assume integrability

Assumption 5.4.2: Integrability

Let $Y_{it} = f(\alpha_i, \gamma_t, \epsilon_{it})$, where $\alpha_i, \gamma_t, \epsilon_{it}$ are random arrays with elements i.i.d. drawn from $[0, 1]$ uniform distribution, assume

- a_i/σ_a, g_t/σ_g, v_{it}/σ_v, e_{it}/σ_e are well-defined and have bounded moments up to the order $4 + \delta$ for some $\delta > 0$, whenever the respective variances $\sigma_a^2, \sigma_g^2, \sigma_v^2, \sigma_e^2$ are non-zero.
- $\sigma_a^2 + \sigma_g^2 > 0$, or $\sigma_v^2 + \sigma_e^2 > 0$

Low-rank approximation Consider the row/column projection

$$\overline{v}_{NT} \equiv \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} (\mathbb{E}[Y_{it} \mid \alpha_i, \gamma_t] - \mathbb{E}[Y_{it} \mid \alpha_i] - \mathbb{E}[Y_{it} \mid \gamma_t] + \mathbb{E}[Y_{it}]) =: \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} v(\alpha_i, \gamma_t)$$

as a generalized U-statistic with a kerel $v(\alpha, \gamma)$ evaluated at the samples $\alpha_1, \cdots, \alpha_N$ and $\gamma_1, \cdots, \gamma_t$. There are 2 major issues w.r.t. characterizing the distribution of $\overline{Y}_{NT}$

- the presence of the projection error e_{it}
- the factors α_i, γ_t are not observable

Define,

$$v(\alpha,\gamma) \coloneqq \mathbb{E}\left[Y_{it} \mid \alpha_i = \alpha, \gamma_t = \gamma\right] - \mathbb{E}\left[Y_{it} \mid \alpha_i = \alpha\right] - \mathbb{E}\left[Y_{it} \mid \gamma_t = \gamma\right] + \mathbb{E}\left[Y_{it}\right]$$

under Assumption 5.4.2, we have compact integral operators

$$S(u)(g) = \int v(a,g)u(a)F_\alpha(\mathrm{d}a) \qquad S^*(u)(a) = \int v(a,g)u(g)F_\gamma(\mathrm{d}g)$$

where F_α, F_γ are the marginal distributions corresponding to the joint $F_{\alpha\gamma}$ of α_i, γ_t. Then the low-rank approximation is

$$v(\alpha,\gamma) = \sum_{k=1}^{\infty} c_k \phi_k(\alpha)\psi_k(\gamma) \tag{5.14}$$

under the $L_2(F_{\alpha\gamma})$ norm on the space of smooth functions of $(\alpha,\gamma) \in [0,1]^2$. Here

- $(c_k)_{k\geq 1}$: a sequence of singular values, $\lim |c_k| \to 0$
- $\left(\phi_k(\cdot)\right)_{k\geq 1}$ and $\left(\psi_k(\cdot)\right)_{k\geq 1}$: orthonormal bases for $L_2\left([0,1], F_\alpha\right)$ and $L_2\left([0,1], F_\gamma\right)$:
  - By construction:

$$\mathbb{E}\left[v(a,\gamma_t)\right] = \mathbb{E}\left[v(\alpha_i,g)\right] = 0, \forall a, g \in [0,1] \Rightarrow \mathbb{E}\left[\phi_k(\alpha_i)\right] = \mathbb{E}\left[\psi_k(\gamma_t)\right] = 0, \forall k = 1,2,\cdots$$

  - the basis functions are orthonormal and α_i and γ_t are independent, then $\forall K < \infty$

$$\mathrm{Cov}\left[\left(\phi_1(\alpha_i), \psi_1(\gamma_t), \cdots, \phi_K(\alpha_i), \psi_K(\gamma_t)\right)\right]$$

    is the $2K$-dimensional identity matrix
  - $\left(\phi_1(\alpha_i), \cdots, \phi_K(\alpha_i)\right)$ can be correlated with a_i: $\sigma_{ak} \coloneqq \mathrm{Cov}\left(a_i, \phi_k(\alpha_i)\right)$
  - $\left(\psi_1(\gamma_t), \cdots, \psi_K(\gamma_t)\right)$ can be correlated with g_t: $\sigma_{gk} \coloneqq \mathrm{Cov}\left(g_t, \psi_k(\gamma_t)\right)$

with this representation of Eq.(5.14), we have[7]

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} v(\alpha_i,\gamma_t) = \sum_{k=1}^{\infty} c_k \left(\frac{1}{N}\sum_{i=1}^{N}\phi_k(\alpha_i)\right)\left(\frac{1}{T}\sum_{t=1}^{T}\psi_k(\gamma_t)\right)$$

and the second-order projection term can also be represented as a function of **countably many** sample averages of **i.i.d. mean-zero** random variables.

Assumption 5.4.3: Eigenfucntions and coefficients in the spectral represention (5.14)

The function $v(\alpha,\gamma) \coloneqq \mathbb{E}\left[Y_{it} \mid \alpha_i = \alpha, \gamma_t = \gamma\right] - \mathbb{E}\left[Y_{it} \mid \alpha_i = \alpha\right] - \mathbb{E}\left[Y_{it} \mid \gamma_t = \gamma\right] + \mathbb{E}\left[Y_{it}\right]$ admits a spectral representation

$$v(\alpha,\gamma) = \sum_{k=1}^{\infty} c_k \phi_k(\alpha)\psi_k(\gamma)$$

under the $L_2(F_{\alpha\gamma})$ norm. And

- the singular values are uniformly bounded by a square summable null sequence $\overline{c}_k$: $c_k \leq \overline{c}_k, \forall k = 1,2,\cdots$, where $\sum_{k=1}^{\infty} c_k^2 < \infty$

[7] The limiting distribution of this term is not Gaussian, but can be represented as a linear combination of independent chi-squared random variables. This type of distributions is known as Wiener/Gaussian chaos.

- $\forall k = 1, 2, \cdots$, the first 3 moments of the eigenfunctions $\phi_k(\alpha_i)$ and $\psi_k(\gamma_t)$ are bounded by a constant $B > 0$

To summarize the two assumptions

- Assumption 5.4.1 guarantees the pointwise consistency of the bootstrap
- Assumption 5.4.3 gives the uniform consistency of the bootstrap: it imposes common bounds on moments and singular values and restricts the set of joint distribution F to a **uniformity** class[8].

5.4.2.3 Bootstrap procedure

For the sample mean $\overline{Y}_{NT} - \mathbb{E}[Y_{it}]$, the limiting distribution depends on the scale parameters:

- If observations are independent across rows and columns: $\sqrt{NT}\left(\overline{Y}_{NT} - \mathbb{E}[Y_{it}]\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma_e^2\right)$
- If $N = T$, within-cluster covariances are bounded from 0 in **at least one dimension**: $\sqrt{N}\left(\overline{Y}_{NT} - \mathbb{E}[Y_{it}]\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma_a^2 + \sigma_g^2\right)$

The bootstrap procedure should then be adaptive for both degenerate and non-degenerate cases. For the expansion

$$\begin{aligned} Y_{it} =& \mathbb{E}[Y_{it}] + \left(\mathbb{E}[Y_{it} \mid \alpha_i] - \mathbb{E}[Y_{it}]\right) + \left(\mathbb{E}\left[Y_{it} \mid \gamma_t\right] - \mathbb{E}[Y_{it}]\right) \\ &+ \left(\mathbb{E}\left[Y_{it} \mid \alpha_i, \gamma_t\right] - \mathbb{E}[Y_{it} \mid \alpha_i] - \mathbb{E}\left[Y_{it} \mid \gamma_t\right] + \mathbb{E}[Y_{it}]\right) + \left(Y_{it} - \mathbb{E}\left[Y_{it} \mid \alpha_i, \gamma_t\right]\right) \\ =:& b + a_i + g_t + v_{it} + e_{it} \end{aligned} \tag{5.15}$$

the sample analogs are:

$$\hat{a}_i := \frac{1}{T}\sum_{t=1}^{T} Y_{it} - \overline{Y}_{NT} \qquad \hat{g}_t := \frac{1}{N}\sum_{i=1}^{N} Y_{it} - \overline{Y}_{NT} \qquad \hat{w}_{it} := Y_{it} - \hat{a}_i - \hat{g}_t - \overline{Y}_{NT}$$

Evaluating bootstrap performance it is crucial at what rates these estimators are consistent depending on the extent of clustering in the true DGP. The variance of the projection terms are:

$$\mathrm{Var}(\hat{a}_i) = \sigma_a^2 + \frac{\sigma_w^2}{T} \qquad \mathrm{Var}\left(\hat{g}_t\right) = \sigma_g^2 + \frac{\sigma_w^2}{N}$$

s.t. the **convolution error** depending on σ_w^2 dominates in the degenerate case. Therefore, to correct for the contribution of the row/column averages of w_{it}, consider the scalar for the distribution of $\hat{a}_i, \hat{g}_t$ by

$$\lambda_a = \frac{T\sigma_a^2}{T\sigma_a^2 + \sigma_w^2} \qquad \lambda_g = \frac{N\sigma_g^2}{N\sigma_g^2 + \sigma_w^2}$$

[8]Here, the sequence $\mathbf{c} := \left(\bar{(c)}_k\right)_{k\geq 0}$ controls the magnitude of the error from a finite-dimensional approximation to $v(\alpha, \gamma)$.

Component variance estimator let

$$\hat{s}_a^2 := \frac{1}{N-1}\sum_{i=1}^{N}\left(\hat{a}_i - \overline{Y}_{NT}\right)^2$$
$$\hat{s}_g^2 := \frac{1}{T-1}\sum_{t=1}^{T}\left(\hat{g}_t - \overline{Y}_{NT}\right)^2$$
$$\hat{s}_w^2 := \frac{1}{NT-N-T}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(Y_{it} - \hat{a}_i - \hat{g}_t - \overline{Y}_{NT}\right)^2$$

then form the estimators as

$$\hat{\sigma}_a^2 = \max\left\{0, \hat{s}_a^2 - \frac{1}{T}\hat{s}_w^2\right\} \qquad \hat{\sigma}_g^2 = \max\left\{0, \hat{s}_g^2 - \frac{1}{N}\hat{s}_w^2\right\} \qquad \hat{\sigma}_w^2 := \hat{s}_w^2 \tag{5.16}$$

the rates of convergence for these estimators are given in the following lemma:

Lemma 5.4.4: Stochastic Order of Variance Estimators

Under Assumption 5.4.1,

$$\hat{\sigma}_a^2 - \sigma_a^2 = O_p\left(\frac{1}{\sqrt{N}}\left(\sigma_a + \frac{\sigma_e}{\sqrt{T}}\right)^2 + \frac{\sigma_v^2}{T}\right)$$
$$\hat{\sigma}_g^2 - \sigma_g^2 = O_p\left(\frac{1}{\sqrt{T}}\left(\sigma_g + \frac{\sigma_e}{\sqrt{N}}\right)^2 + \frac{\sigma_v^2}{N}\right)$$
$$\hat{\sigma}_w^2 - \sigma_w^2 = O_p\left(\frac{\sigma_e^2}{\sqrt{NT}} + \left(\frac{1}{N} + \frac{1}{T}\right)\sigma_v^2\right)$$

and there exist **no estimators** for $\sigma_a^2, \sigma_g^2, \sigma_w^2$ that converge at rates faster than these rates. Specifically, σ_a^2 can **NOT** be estimated at a rate faster than T^{-1} even when $\sigma_a^2 = 0$[a].

[a] See the appendix of Menzel (2021) for the proof.

Hence, a bootstrap procedure can use a consistent pre-test for the presence of cluster dependence in the **first moment**, with the model selectors

$$\hat{D}_a(\kappa) := \mathbf{1}\left\{T\hat{\sigma}_a^2 \geq \kappa\right\} \qquad \hat{D}_g(\kappa) := \mathbf{1}\left\{N\hat{\sigma}_g^2 \geq \kappa\right\}$$

$\forall \kappa \geq 0$. And for some κ_a, κ_g, let

$$\hat{\lambda}_a := \frac{\hat{D}_a(\kappa_a)T\hat{\sigma}_a^2}{\hat{D}_a(\kappa_a)T\hat{\sigma}_a^2 + \hat{\sigma}_w^2} \qquad \hat{\lambda}_g := \frac{\hat{D}_g(\kappa_g)T\hat{\sigma}_g^2}{\hat{D}_g(\kappa_g)N\hat{\sigma}_g^2 + \hat{\sigma}_w}$$

and estimate the asymptotic variance of the sample mean as

$$\hat{S}_{NT,sel}^2 := \hat{D}_a(\kappa_a)T\hat{\sigma}_a^2 + \hat{D}_g(\kappa_g)N\hat{\sigma}_g^2 + \hat{\sigma}_w^2 \tag{5.17}$$

Bootstrap procedures Menzel (2021) proposed the following resampling algorithm to estimate the sampling distribution for exhaustive sampling with cluster dependence in two dimensions

Algorithm 5.4.5: Resampling Algorithm

(a) For the b-th bootstrap iteration, draw

$$a^*_{i,b} := \hat{a}_{k^*_b(i)} \qquad\qquad g^*_{t,b} := \hat{g}_{s^*_b(t)}$$

where $k^*_b(i)$ and $s^*_b(t)$ are i.i.d. draws from the discrete uniform distribution on the index sets $\{1, \cdots, N\}$ and $\{1, \cdots, T\}$ respectively

(b) Generate

$$w^*_{it,b} := \omega_{1i,b}\omega_{2t,b}\hat{w}_{k^*_b(i)s^*_b(t)}$$

where $\omega_{1i,b}, \omega_{2t,b}$ are i.i.d. random variables with $\mathbb{E}[\omega] = 0$, $\mathbb{E}\left[\omega^2\right] = \mathbb{E}\left[\omega^3\right] = 1$[a]

(c) Generate a bootstrap sample of draws

$$Y^*_{it,b} = \overline{Y}_{NT} + \sqrt{\hat{\lambda}_a}a^*_{i,b} + \sqrt{\hat{\lambda}_g}g^*_{t,b} + w^*_{it,b}$$

and get the bootstrapped statistic

$$\overline{Y}^*_{NT,b} := \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} Y^*_{it,b}$$

(d) Repeat this procedure, get a sample of B replications and approximate the conditional distribution of $\overline{Y}^*_{NT}$ given the sample with the empirical distribution over the bootstrap draws $\overline{Y}^*_{NT,1}, \cdots, \overline{Y}^*_{NT,B}$

[a] Typical choices of $\omega_{1i,b}, \omega_{2t,b}$ are the Gamma distribution (with shape = 4, scale = 1/2).

For the **pivotal boostrap**, the last step uses instead the empirical distribution of the studentized bootstrap draws to approximate the distribution of

$$\sqrt{NT}\left(\overline{Y}^*_{NT} - \overline{Y}_{NT}\right)/\hat{S}^*_{NT,sel}$$

where $\hat{S}^*_{NT,sel}$ is the bootstrap analog of the variance estimator $\hat{S}_{NT,sel}$.

Definition 5.4.6: Bootstrap Procedures

Consider 3 versions of the bootstrap procedure based on 5.4.5:

- **BS-N** (bootstrap *without* model selection): apply steps (a) - (d), and set $\kappa_a = \kappa_g = 0$
- **BS-S** (bootstrap *with* model selection): apply steps (a) - (d), and set κ_a, κ_g according to increasing sequences $\kappa_g, \kappa_a \to \infty$ s.t. $\kappa_a/T \to 0$ and $\kappa_g/N \to 0$
- **BS-C** (*conservative* bootstrap): addition to the settings of **BS-S**, set

$$\hat{\lambda}_a := \frac{\hat{q}_a}{\hat{q}_a + \hat{\sigma}^2_w}\frac{\hat{q}_a}{T\hat{\sigma}^2_a} \qquad\qquad \hat{\lambda}_g := \frac{\hat{q}_g}{\hat{q}_g + \hat{\sigma}^2_w}\frac{\hat{q}_g}{N\hat{\sigma}^2_g}$$

where

$$\hat{q}_a := \max\left\{T\hat{\sigma}^2_a, \kappa_a\right\} \qquad\qquad \hat{q}_g := \max\left\{N\hat{\sigma}^2_g, \kappa_g\right\}$$

Consistency of the bootstrap procedures

- **BS-N** (bootstrap *with* model selection): **pointwise** consistent in $\sigma_a^2, \sigma_g^2, \sigma_w^2$
- **BS-S** (bootstrap *without* model selection): **uniformly** consistent if the limiting distribution is Gaussian
- **BS-C** (*conservative* bootstrap): **consistent** in the nondegerate case $\sigma_a^2 + \sigma_g^2 > 0$, but asymptotically **conservative** for the degenerate cases

To establish the consistency, define the **adaptive rate** r_{NT} as[9]

$$r_{NT}^{-2} \coloneqq N^{-1}\sigma_a^2 + T^{-1}\sigma_g^2 + (NT)^{-1}\sigma_w^2 \equiv \text{Var}(\overline{Y}_{NT})$$

then consider the limiting distribution with the respective limits of normalized sequences:

$$\begin{aligned} q_{a,NT} &\coloneqq r_{NT}^2 N^{-1}\sigma_a^2 \qquad q_{g,NT} \coloneqq r_{NT}^2 T^{-1}\sigma_g^2 \qquad q_{e,NT} \coloneqq r_{NT}^2 (NT)^{-1}\sigma_e^2 \qquad q_{v,NT} \coloneqq r_{NT}^2 (NT)^{-1}\sigma_v^2 \\ q_{ak,NT} &\coloneqq r_{NT}^2 N^{-1}\sigma_{ak} \qquad q_{gk,NT} \coloneqq r_{NT}^2 T^{-1}\sigma_{gk} \end{aligned} \tag{5.18}$$

for $k = 1, 2, \cdots$. Let $\varrho_{NT} \coloneqq r_{NT}\,(NT)^{-1/2}$, then

$$q_{a,NT} + q_{g,NT} + q_{e,NT} + q_{v,NT} = 1$$

stacking the sequences as the vector

$$\mathbf{q}_{NT} \coloneqq \left(q_{e,NT}, q_{a,NT}, q_{g,NT}, q_{a1,NT}, q_{g1,NT}, q_{a2,NT}, q_{g2,NT}, \cdots\right)$$

and the singular values for the spectral decomposition (5.14):

$$\begin{aligned} \mathbf{c}_{NT} &\coloneqq (c_{1,NT}, c_{2,NT}, \cdots) \in l^2 && \text{for } \mathbb{E}_{NT}\left[Y_{it} \mid \alpha_i, \gamma_t\right] \\ \mathbf{c} &\coloneqq (c_1, c_2, \cdots) \in l^2 && \text{for } \mathbb{E}\left[Y_{it} \mid \alpha_i, \gamma_t\right] \end{aligned}$$

then for convergent sequences $\mathbf{q}_{NT}, \mathbf{c}_{NT}, \mathbf{c}$, denote the limits

$$\begin{aligned} q_a &\coloneqq \lim_{N,T} q_{a,NT} \qquad q_g \coloneqq \lim_{N,T} q_{g,NT} \qquad q_e \coloneqq \lim_{N,T} q_{e,NT} \qquad q_v \coloneqq \lim_{N,T} q_{v,NT} \\ \mathbf{q} &\coloneqq \lim_{N,T} \mathbf{q}_{NT} \qquad \mathbf{c} \coloneqq \lim_{N,T} \mathbf{c}_{NT} \qquad \varrho \coloneqq \lim_{N,T} \varrho_{NT} \end{aligned}$$

for any fixed values of $\mathbf{q}, \mathbf{c}, \varrho \in [0, 1]$, define

$$\mathcal{L}_0\left(\mathbf{q}, \mathbf{c}, \varrho\right) \coloneqq \left(\sqrt{q_e} Z^e + \sqrt{q_a} Z^a + \sqrt{q_g} Z^g\right) + \varrho V \tag{5.19}$$

where

$$V \coloneqq \sum_{k=1}^{\infty} c_k Z_k^{\psi} Z_k^{\phi}$$

and $Z^e, Z_k^{\psi}, Z_k^{\phi}$ are i.i.d. standard normal random variables, Z^a, Z_g are standard normal random variables with

$$\text{Cov}\left(Z^a, Z_k^{\phi}\right) = \frac{q_{ak}}{\sqrt{q_a}} \qquad \text{Cov}\left(Z^g, Z_k^{\psi}\right) = \frac{q_{gk}}{\sqrt{q_g}} \qquad \text{Cov}\left(Z^a, Z^g\right) = \text{Cov}\left(Z^a, Z_k^{\psi}\right) = \text{Cov}\left(Z^g, Z_k^{\psi}\right) = 0$$

Then, the CLT for sampling distribution is established as

[9] Following Eq. (5.15), $\text{Var}(\overline{Y}_{NT}) = \text{Var}(b + \overline{a}_N + \overline{g}_T + \overline{v}_{NT} + \overline{e}_{NT})$.

Theorem 5.4.7: CLT for Sampling Distribution

Under Assumption 5.4.2,

(**a**) along *any* convergent sequence $\mathbf{q}_{NT} \to \mathbf{q}$ and fixed $\mathbf{c} = (c_1, c_2, \cdots)$, we have

$$\left\| \mathbb{P}\left(r_{NT}\left(\overline{Y}_{NT} - \mathbb{E}[Y_{it}] \right) \right) - \mathcal{L}_0\left(\mathbf{q}, \mathbf{c}, \varrho\right) \right\|_\infty \to 0$$

where $\varrho \coloneqq \lim_{N,T} \varrho_{NT}$, and $\|\cdot\|_{infty}$ denotes the Kolmogorov metric; the limiting distribution $\mathcal{L}_0\left(\mathbf{q}, \mathbf{c}, \varrho\right)$ is continuous[a].

(**b**) if in addition, Assumption 5.4.3 holds, (**a**) is robust under drifting sequences $\mathbf{c}_{NT} \to \mathbf{c}$[b]

[a] The convergence is pointwise w.r.t. the conditional mean function $\mathbb{E}\left[Y_{it} \mid \alpha_i = \alpha, \gamma_t = \gamma\right]$

[b] The convergence is uniform within the class of distributions satisfying Assumption 5.4.3

Estimating the asymptotic distribution Lemma 5.4.4 establishes the consistency of the estimation for the components vairances $\sigma_a^2, \sigma_g^2, \sigma_w^2$, but are they **fast** enough?

Proposition 5.4.8: Estimability of Asymptotic Distribution

Let $\hat{\mathcal{L}}_{NT}$ denote an arbitrary estimator for $\mathcal{L}_0$ based on an array of size N, T form the unknown distribution, then $\exists \delta > 0$ s.t.

$$\liminf_{N,T\to\infty} \sup_{f\in\mathcal{F}} \mathbb{P}_{f,NT}\left(\left\| \hat{\mathcal{L}}_{NT} - \mathcal{L}_0\left(\mathbf{q}_{NT}(f), \mathbf{c}_{NT}(f), \varrho_{NT}(f)\right) \right\|_\infty > \delta \right) > 0$$

where

- $\mathcal{F}$: the class of functions $f(\alpha, \gamma, \epsilon)$ corresponding to distributions of Y_{it} satisfying Assump. 5.4.2 and 5.4.3, for i.i.d. uniform $\alpha_i, \gamma_t, \epsilon_{it}$[a]
- $\mathbb{P}_{f,NT}(\cdot)$: probabilities for events w.r.t. an array of size N, T, generated according to f
- $\mathbf{q}_{NT}(f) \coloneqq \left(q_{e,NT}(f), q_{a,NT}(f), \cdots\right)$: the vector of normalized variances from Eq. 5.18

[a] From the Aldous-Hoover representation

Proposition 5.4.8 states that there exists **no estimator** [10] for the asymptotic distribution that achieves consistency uniformly over the space of distributions satisfying Assumption 5.4.2 and 5.4.3:

- Under Theorem 5.4.7, the sample mean $\overline{Y}_{NT}$ converges to a continuous limiting distribution $\mathcal{L}_0(\mathbf{q}, \mathbf{c}, \varrho)$ along sequences $f_{NT} \in \mathcal{F}$ with proper limits for $\mathbf{q}_{NT}, \mathbf{c}_{NT}$

[10] Consider the counterexample for this impossibility: for the model

$$Y_{it} = \alpha_i \gamma_t$$

where α_i, γ_t are mutually independent with i.i.d. factors $\alpha_i \sim \mathcal{N}(0,1), \gamma_t \sim \mathcal{N}(\mu_\gamma, 1)$. This model satisfies Assump.5.4.2, hence Thm.5.4.7 gives convergence results. However, for this model

$$a_i \coloneqq \mathbb{E}\left[Y_{it} \mid \alpha_i\right] = \alpha_i \mu_\gamma \qquad g_t \coloneqq \mathbb{E}\left[Y_{it} \mid \gamma_t\right] = \gamma_t \mathbb{E}\left[\alpha_i\right] \equiv 0$$
$$v_{it} = \alpha_i(\gamma_t - \mu_\gamma) \qquad \sigma_a^2 = \mu_\gamma^2 \qquad \sigma_v^2 = 1$$

here, μ_γ can **not** be estimated from the original data at a rate faster than $T^{-1/2}$, the fastest possible rate at which μ_γ can be estimated from observing $\gamma_1, \cdots, \gamma_T$ directly. Therefore, no test can consistently distinguish the model $\mu_\gamma = 0$ (asymptotic variance σ_v^2) from a drifting sequence $\tilde{\mu}_{T,\gamma} \coloneqq T^{-1/2} m_\gamma$ (asymptotic variance $m_\gamma^2 + \sigma_v^2$).

Bootstrap Consistency Consider the bootstrap analog of $\hat{S}_{NT,sel}$ in Eq. 5.17

$$\hat{S}^{2*}_{NT,sel} := \hat{D}_a(\kappa_a)T\hat{\sigma}^{2*}_a + \hat{D}_g(\kappa_g)N\hat{\sigma}^{2*}_g + \hat{\sigma}^{2*}_w$$

where $\hat{D}_a(\kappa_a), \hat{D}_g(\kappa_g)$ are fixed at the sample values, κ_a, κ_g are chosen according to whether the bootstrap is **with** or **without** model selection. Consider 2 versions based on the studentized sample mean:

- **non-pivotal** bootstrap: approximating the distribution of **the sample mean** $r_{NT}\left(\overline{Y}_{NT} - \mathbb{E}\left[Y_{it}\right]\right)$ with the bootstrap distribution $r_{NT}\left(\overline{Y}^*_{NT} - \overline{Y}_{NT}\right)$
- **pivotal** bootstrap: approximating the distribtuion of the **studentized sample mean** $\frac{(NT)^{1/2}}{\hat{S}_{NT,sel}}\left(\overline{Y}_{NT} - \mathbb{E}\left[Y_{it}\right]\right)$ with the boostrap distribution $\frac{(NT)^{1/2}}{\hat{S}^*_{NT,sel}}\left(\overline{Y}^*_{NT} - \overline{Y}_{NT}\right)$

And we can establish the consistency

Theorem 5.4.9: Bootstrap Consistency

Under Assumption 5.4.2,

(a) the bootstrap **with model selection** satisfies

$$\left\|\mathbb{P}^*_{NT}\left(r_{NT}\left(\overline{Y}^*_{NT} - \overline{Y}_{NT}\right)\right) - \mathbb{P}_{NT}\left(r_{NT}\left(\overline{Y}_{NT} - \mathbb{E}\left[\overline{Y}_{it}\right]\right)\right)\right\|_\infty \xrightarrow{\text{a.s.}} 0 \tag{5.20}$$

and its pivotal analog

$$\left\|\mathbb{P}^*_{NT}\left(\sqrt{NT}\frac{\overline{Y}^*_{NT} - \overline{Y}_{NT}}{\hat{S}^*_{NT,sel}}\right) - \mathbb{P}_{NT}\left(\sqrt{NT}\frac{\overline{Y}_{NT} - \mathbb{E}\left[Y_{it}\right]}{\hat{S}_{NT,sel}}\right)\right\|_\infty \xrightarrow{\text{a.s.}} 0 \tag{5.21}$$

pointwise for any fixed $\sigma^2_a, \sigma^2_g, \sigma^2_e, \sigma^2_v$

(b) the bootstrap **without model selection** satisfies Eq.5.20 and Eq.5.21 **uniformly** if $q_v = 0$

(c) the **conservative** bootstrap satisfies

$$\left\|\mathbb{P}^*_{NT}\left(r_{NT}\left(\overline{Y}^*_{NT} - \overline{Y}_{NT}\right)\right) - \mathcal{L}_0\left(\overline{\mathbf{q}}, \mathbf{c}, \varrho\right)\right\|_\infty \xrightarrow{\text{p}} 0 \tag{5.22}$$

and its pivotal analog

$$\left\|\mathbb{P}^*_{NT}\left(\sqrt{NT}\frac{\overline{Y}^*_{NT} - \overline{Y}_{NT}}{\hat{S}^*_{NT,sel}}\right) - \mathcal{L}_0\left(\overline{\mathbf{q}}, \mathbf{c}, \varrho\right)\right\|_\infty \xrightarrow{\text{p}} 0 \tag{5.23}$$

uniformly over the **entire parameter space**, where $\overline{q} = \left(q_c, \overline{q}_a, \overline{q}_g, 0, 0, \cdots\right)$, with $\overline{q}_a := \max\left\{\kappa_a/T, q_a\right\}$ and $\overline{q}_g := \max\left\{\kappa_g/T, q_g\right\}$, which increases as $N, T \to \infty$.

Theorem 5.4.7 gives that

- **bootstrap with model selection**: pointwise valid asymptotically
- **bootstrap without model selection**: valid uniformly w.r.t. clustering in means, but **inconsistent** if $q_v > 0$
- **conservative bootstrap**: uniformly valid without any qualifications. In degenerate cases $\left(q_e + q_v > 0\right)$, the scale of the estimated asymptotic distribution **diverges** at a rate $\kappa_a/T + \kappa_g/N$

Notice that $\mathcal{L}_0(\overline{\mathbf{q}}, \mathbf{c}, \varrho)$ in Thm. 5.4.9 is a mean-preserving spread of $\mathcal{L}_0(\mathbf{q}, \mathbf{c}, \varrho)$ in Thm. 5.4.7, hence estimates of percentiles from the conservative bootstrap are **biased outwards** away from 0, leading to asymptotic conservative CIs.

Refinements Using standard results on Edgeworth expansions, get

Proposition 5.4.10: Refinements

Under Assumption 5.4.2 for any $0 < \delta < \infty$, and the distributions of a_i and g_t satisfy Cramer's condition[a]

$$\limsup_{\|t\| \to \infty} \left| \mathbb{E}\left[\exp(i\mathbf{t}'\mathbf{X})\right] \right| < 1$$

then if $\sigma_a^2 + \sigma_g^2 \geq C$ for some $C > 0$, we have

$$\left\| \mathbb{P}_{NT}^* \left(\sqrt{NT} \frac{\overline{Y}_{NT}^* - \overline{Y}_{NT}}{\hat{S}_{NT,sel}^*} - \mathbb{P}_{NT} \left(\sqrt{NT} \frac{\overline{Y}_{NT} - \mathbb{E}[Y_{it}]}{\hat{S}_{NT,sel}} \right) \right) \right\|_\infty = O_p\left(r_{NT}^{-2} \vee (NT)^{-1/2} \right)$$

for all three versions of the bootstrap.

[a] Cramer's condition states that $\mathbf{X}$ has a non-degenerate, absolutely continuous component.

5.4.2.4 Inference in Regression Models

Consider the linear projection model

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + u_{it} \tag{5.24}$$

with the dependent variable y_{it} and the vector of k regressors $\mathbf{x}_{it} \in \mathbb{R}^k$. Consider LS estimator

$$\hat{\boldsymbol{\beta}}_{LS} := (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\left(\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{x}_{it} u_{it} \right)$$

assume $(\mathbf{x}_{it} u_{it})_{i,t}$ constitute a dissociated, separately exchangeable array, then we can have the Aldous-Hoover representation

$$F_{it} := \mathbf{x}_{it} u_{it} = f\left(\alpha_i, \gamma_t, \epsilon_{it}\right)$$

then denote

$$\begin{aligned} \mathbf{a}_i &:= \mathbb{E}\left[\mathbf{x}_{it} u_{it} \mid \alpha_i\right] & \mathbf{g}_t &:= \mathbb{E}\left[\mathbf{x}_{it} u_{it} \mid \gamma_t\right] \\ \mathbf{v}_{it} &:= \mathbb{E}\left[\mathbf{x}_{it} u_{it} \mid \alpha_i, \gamma_t\right] - \mathbf{a}_i - \mathbf{g}_t & \mathbf{e}_{it} &:= \mathbf{x}_{it} u_{it} - \mathbb{E}\left[\mathbf{x}_{it} u_{it} \mid \alpha_i, \gamma_t\right] \\ \mathbf{w}_{it} &:= \mathbf{x}_{it} u_{it} - \mathbf{a}_i - \mathbf{g}_t = \mathbf{v}_{it} + \mathbf{e}_{it} \end{aligned}$$

and the unconditional component variances as $\sigma_{al}^2, \sigma_{gl}^2, \sigma_{vl}^2, \sigma_{el}^2, \sigma_{wl}^2 = \sigma_{vl}^2 + \sigma_{el}^2$. The empirical analog of this decomposition is given by

$$\begin{aligned} \hat{\mathbf{a}}_i &:= \frac{1}{T} \sum_{t=1}^{T} \mathbf{x}_{it} \hat{u}_{it} & \hat{\mathbf{g}}_t &:= \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_{it} \hat{u}_{it} \\ \hat{\mathbf{w}}_{it} &:= \mathbf{x}_{it} \hat{u}_{it} - \hat{\mathbf{a}}_i - \hat{\mathbf{g}}_t \end{aligned}$$

for each $l = 1, \cdots, k$, then construct

Bootstrap procedure for regression

- for the bth bootstrap iteration, draw $\mathbf{a}^*_{i,b} := \hat{\mathbf{a}}_{k^*_b(i)}$ and $\mathbf{g}^*_{t,b} := \hat{\mathbf{g}}_{s^*_b(i)}$ where $k^*_b(i)$ and $s^*_b(t)$ are i.i.d. draws from the discrete uniform distribution on the index sets $\{1, \cdots, N\}$ and $\{1, \cdots, T\}$, respectively
- generate $\mathbf{w}^*_{it,b} := \omega_{1i,b}\omega_{2t,b}\hat{\mathbf{w}}_{k^*_b(i)s^*_b(i)}$, where $\omega_{1i,b}, \omega_{2t,b}$ are i.i.d. random variables with $\mathbb{E}[\omega] = 0, \mathbb{E}[\omega^2] = \mathbb{E}[\omega^3] = 1$
- simulate values of $\mathbf{z}^*_{it,b} = \left(z^*_{it1,b}, \cdots, z^*_{itk,b}\right)'$, where the lth component is given by

$$z^*_{itl,b} := \sqrt{\hat{\lambda}_{al}}a^*_{il,b} + \sqrt{\hat{\lambda}_{gl}}g^*_{tl,b} + w^*_{itl,b}$$

  where the scalars are $\hat{\lambda}_{al} := \frac{\hat{D}_{al}(\kappa_a)T\hat{\sigma}^2_{al}}{\hat{D}_{al}(\kappa_a)T\hat{\sigma}^2_{al}+\hat{\sigma}^2_{wl}}$ and $\hat{\lambda}_{gl} := \frac{\hat{D}_{gl}(\kappa_g)N\hat{\sigma}^2_{gl}}{\hat{D}_{gl}(\kappa_g)N\hat{\sigma}^2_{gl}+\hat{\sigma}^2_{wl}}$
- then compute

$$\hat{\boldsymbol{\beta}}^*_{LS,b} = \hat{\boldsymbol{\beta}}_{LS} + (\mathbf{X}'\mathbf{X})^{-1}\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{z}^*_{it,b}\right)$$

  for each bootstrap sample.

Next, we can approximate the asymptotic distribution of $r_{NT}\left(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}\right)$ with the simulated distribution of $r_{NT}\left(\hat{\boldsymbol{\beta}}^*_{LS,b} - \hat{\boldsymbol{\beta}}_{LS}\right)$.

Assumption 5.4.11: Regression

Assume the model in Eq.(5.24) with $\mathbf{x}_{it}u_{it} = f(\alpha_i, \gamma_t, \epsilon_{it})$ and $\alpha_i, \gamma_t, \epsilon_{it}$ are i.i.d. uniform on $[0, 1]$. And

(a) $\mathbf{X}$ has full column rank

(b) $\forall l = 1, \cdots, k$ and some $\delta > 0$, the $(4+\delta)$th absolute moments of x_{itl} are bounded, and the $(4+\delta)$th conditional moments of each component $\frac{a_{il}}{\sqrt{\text{Var}(a_{il}|\mathbf{X})}}$, $\frac{g_{tl}}{\sqrt{\text{Var}(g_{tl}|\mathbf{X})}}$, $\frac{v_{itl}}{\sqrt{\text{Var}(v_{itl}|\mathbf{X})}}$ and $\frac{e_{itl}}{\sqrt{\text{Var}(e_{itl}|\mathbf{X})}}$ given $\mathbf{X}$ are bounded whenever the conditional variance of either component is strictly positive.

(c) unconditional variance: $\text{Var}(a_{il}) + \text{Var}(g_{tl}) > 0$ or $\text{Var}(w_{itl}) > 0$ for each $l = 1, \cdots, k$.

(d) For each component of $\mathbf{z}_{it} = \mathbf{x}_{it}u_{it}$, there exists a spectral representation satisfying Assumption 5.4.3

then analogous to the sample mean inference, we have

Proposition 5.4.12: Regression Inference

Under Assumption 5.4.11, then

- $\hat{\boldsymbol{\beta}}_{LS}$ is consistent at the r_{NT} rate
- The bootstrap with model selection satisfies Eq.(5.20) and (5.21) pointwise as $\sigma^2_{al}, \sigma^2_{gl}, \sigma^2_{el}, \sigma^2_{vl}$ are held fixed for all $l = 1, \cdots, k$
- The bootstrap without mode selection satisfies Eq.(5.20) and (5.21) uniformly if $q_{vl} = 0$ for all $l = 1, \cdots, k$
- The conservative bootstrap satisfies Eq.(5.22) and (5.23) uniformly over the entire parameter space

Asymptotic Gaussian of the LS estimator for conditional asymptotic normality of bilinear forms $V_k := \mathbf{Z}_{1k}'\mathbf{X}\mathbf{Z}_{2k}$ of random vectors $\mathbf{Z}_{1k}, \mathbf{Z}_{2k}$ given the matrix $\mathbf{X}$. Under the conditions of this paper, V_k is asymptotically Gaussian if $\ddot{\mathbf{x}}_{it}, \ddot{\mathbf{x}}_{js}$ are mean-independent for any $(j,s) \neq (i,t)$[11].

5.5 Latest Development

5.5.1 LLN and CLT for Exchangeable Arrays

Davezies et al. (2021) establish uniform LLN and CLT to show consistency and asymptotic normality of **nonlinear** estimators under weak regularity conditions.

5.5.1.1 Set up

Notations For any $A \subset \mathbb{R}$, $B \subset \mathbb{R}^k$ for some $k \geq 2$, then let

$$A^+ = A \cap (0, \infty)$$
$$\overline{B} = \left\{ b = (b_1, \cdots, b_k) \in B : \forall (i,j) \in \{1, \cdots, k\}^2, i \neq j, b_i \neq b_j \right\}$$

and let

- $\mathbb{I}_k = \overline{\mathbb{N}^{+k}}$ denote the set of k-tuples of $\mathbb{N}^+$ **without** repetition
- for any $n \in \mathbb{N}^+$, let $\mathbb{I}_{n,k} = \overline{\{1, \cdots, n\}^k}$
- for any $\mathbf{i} = (i_1, \cdots, i_k), \mathbf{j} = (j_1, \cdots, j_k)$ in $\mathbb{N}^k$, let $\mathbf{i} \odot \mathbf{j} = (i_1 j_1, \cdots, i_k j_k)$, and denote the distinct elements of $\mathbf{i}$ as $\{\mathbf{i}\}$
- for any $r \in \{1, \cdots, k\}$, let
$$\mathcal{E}_r = \left\{ (e_1, \cdots, e_k) \in \{0,1\}^k : \sum_{j=1}^k e_j = r \right\}$$
- for any $A \subset \mathbb{N}^+$, let $\mathfrak{S}(A)$ denote the set of permutations on A, then for any $\mathbf{i} = (i_1, \cdots, i_k) \in \mathbb{N}^{+k}$ and $\pi \in \mathfrak{S}(\mathbb{N}^+)$, let $\pi(\mathbf{i}) = (\pi(i_1), \cdots, \pi(i_k))$

Polyadic data For random variables $Y_{\mathbf{i}}$ indexed by $\mathbf{i} \in \mathbb{I}_k$[12], it's assumed that the random variables are generated according to a **jointly exchangeable** and **dissociated** array:

Assumption 5.5.1: Jointly Exchangeable and Dissociated Arrays

For any $\pi \in \mathfrak{S}(\mathbb{N}^+)$,

$$(Y_{\mathbf{i}})_{\mathbf{i} \in \mathbb{I}_k} \overset{\mathrm{d}}{=} \left(Y_{\pi(\mathbf{i})}\right)_{\mathbf{i} \in \mathbb{I}_k}$$

and for any disjoint subsets of $\mathbb{N}^+$, A, B, with $\min(|A|, |B|) \geq k$, $(Y_{\mathbf{i}})_{\mathbf{i} \in \overline{A^k}}$ is **independent** of $(Y_{\mathbf{i}})_{\mathbf{i} \in \overline{B^k}}$

The assumption implies that

[11] For difference-in-differences designs with a regressor $x_{it1} := \mathbf{1}\{t \geq T_i\}$ for unit-specific intervention date T_i, or when $\mathbf{x}_{it} := \mathbf{x}(\xi_i, \zeta_t)$ are a non-additive function of row- and column-level attributes ξ_i and ζ_t, respectively, these conditions need not hold in general.

[12] Some examples are: Y_{i_1,i_2} corresponds to export flows from country i_1 to i_2, or whether there is a link between node i_1 and i_2 in a network. $\{i_1, \cdots, i_k\}$ can also correspond to the different dimensions of clustering.

- **jointly exchangeability**: the joint distribution of the data remains identical under any possible permutation of labels, i.e., labeling conveys no information
- **dissociation**: the variables are indepednent if they have **no unit** in common, that is Y_{i_1,i_2} must be independent of Y_{j_1,j_2} if $\{i_1,i_2\} \cap \{j_1,j_2\} = \varnothing$

the dependence structure under such assumptions are

Lemma 5.5.2: Key Dependence Structure

Assumption 5.5.1 holds **if and only if** there exists i.i.d. variables $(U_J)_{J\subset\mathbb{N}^+,1\leq|J|\leq k}$ and a measurable function τ s.t. almost surely

$$Y_{\mathbf{i}} = \tau\left(\left(U_{\{\mathbf{i}\odot\mathbf{e}\}^+}\right)_{\mathbf{e}\in\bigcup_{r=1}^k \mathcal{E}_r}\right), \forall \mathbf{i} \in \mathbb{I}_k$$

this result is referred to as the AHK representation (Aldous 1981, Hoover 1979, Kallenberg 1989)[a].

[a] Consider dyadic data ($k = 2$), then for every $i_1 < i_2$ (the ranking is precise), $Y_{i_1,i_2} = \tau\left(U_{i_1}, U_{i_2}, U_{\{i_1,i_2\}}\right)$, that is, the outcome Y depends of factors specific to i_1 and i_2, and factors relating both.

5.5.1.2 Uniform LLN and CLT

Let $\mathcal{F}$ denote a class of real-valued functions admitting a first moment w.r.t. the distribution P, let Pf denote the corresponding moment $\mathbb{E}\left[f(Y_{\mathbf{1}})\right]$, with $\mathbf{1}$ as the $k-$tuple $(1,\cdots,k)$. Assume that

Assumption 5.5.3: Measurability Assumption

$\exists$ a countable subclass $\mathcal{G} \subset \mathcal{F}$ s.t. elements of $\mathcal{F}$ are pointwise limits of sequences of elements of $\mathcal{G}$

Consider

$$\mathbb{P}_n f = \frac{(n-k)!}{n!}\sum_{\mathbf{i}\in\mathbb{I}_{n,k}} f(Y_{\mathbf{i}})$$
$$\mathbb{G}_n f = \sqrt{n}\left(\mathbb{P}_n f - Pf\right)$$

and the restrictions on $\mathcal{F}$: for any $\eta > 0$ and any seminorm $\|\cdot\|$ on a space containing $\mathcal{F}$, let

- $N\left(\eta,\mathcal{F},\|\cdot\|\right)$: the minimal number of $\|\cdot\|$-closed balls of radius η with centers in $\mathcal{F}$ needed to cover $\mathcal{F}$
- $N_{[]}\left(\eta,\mathcal{F},\|\cdot\|\right)$: the minimal number of $\eta-$brackets needed cover $\mathcal{F}$, where an $\eta-$bracket for $f \in \mathcal{F}$ is a pair of functions (l,u) s.t. $l \leq f \leq u$ and $\|u-l\| < \eta$

Davezies et al. (2021) considered the seminorms $\|f\|_{\mu,r} = \left(\int |f|^t \, \mathrm{d}\mu\right)^{1/r}$ for any $r \geq 1$ and probability measure (cdf) μ. An envelope of $\mathcal{F}$ is measurable function F satisfying $F(u) \geq \sup_{f\in\mathcal{F}} |f(u)|$, satisfying

Assumption 5.5.4: Assumptions of $\mathcal{F}$

A The class $\mathcal{F}$

(i) either admits an envelope F with $PF < \infty$ and $\forall \eta > 0$,

$$\sup_{Q\in\boldsymbol{Q}} N\left(\eta\|F\|_{Q,1},\mathcal{F},\|\cdot\|_{Q,1}\right) < \infty$$

(ii) or satisfies $N_{[]}\left(\eta, \mathcal{F}, \|\cdot\|_{L_1(P)}\right) < \infty$ for all $\eta > 0$

B and it

(i) **uniform entropy integral**: either admits an envelope F with $PF^2 < \infty$ and

$$\int_0^\infty \sup_{Q \in \mathcal{Q}} \sqrt{\log N\left(\eta \|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}\right)} \mathrm{d}\eta < \infty$$

(ii) **bracketing entropy integral**: or satisfies $\int_0^\infty \sqrt{\log N_{[]}\left(\eta, \mathcal{F}, \|\cdot\|_{L_2(P)}\right)} \mathrm{d}\eta < \infty$

Assumption 5.5.4 are same as the conditions imposed on i.i.d. data for uniform LLNs and CLTs. Under these assumptions, Davezies et al. (2021) established the uniform LLNs and CLTs as

Theorem 5.5.5: Uniform LLNs and CLTs

Under Assumption 5.5.1 and 5.5.3,

- if (**A**) of Assumption 5.5.4 holds, $\sup_{f \in \mathcal{F}} \left|\mathbb{P}_n f - Pf\right| \xrightarrow{\text{a.s.}} 0$ and in L^1
- if (**B**) of Assumption 5.5.4 holds, $\mathbb{G}_n$ converges weakly in $l^\infty(\mathcal{F})$ to a centered Gaussian process $\mathbb{G}$ on $\mathcal{F}$ as $n \to \infty$, the convariance kernel K of $\mathbb{G}$ satisfies

$$K(f_1, f_2) = \frac{1}{(k-1)!^2} \sum_{(\pi,\pi') \in \mathfrak{S}(\{\mathbf{1}\}) \times \mathfrak{S}(\{\mathbf{1}'\})} \mathrm{Cov}\left(f_1\left(\mathbf{Y}_{\pi(\mathbf{1})}\right), f_2\left(\mathbf{Y}_{\pi'(\mathbf{1}')}\right)\right)$$

Here, (**A**) of Assumption 5.5.4 is stronger than necessary to obtain the uniform LLNs. To establish the exact characterization, consider the norms:

$$\|f\|_{1,1} = \frac{1}{n} \sum_{i_1=1}^{n} \left| \frac{1}{n-1} \sum_{i_2 \neq i_1} f(Y_{i_1,i_2}) + f(Y_{i_2,i_1}) \right|$$

$$\|f\|_{1,2} = \frac{1}{n(n-1)} \sum_{1 \leq i_1 < i_2 \leq n} \left| \mathbb{E}\left[f(Y_{i_1,i_2}) + f(Y_{i_2,i_1})\right] \mid U_{\{i_1,i_2\}} \right|$$

and the exact characterization is established as

Proposition 5.5.6: Exact Characterization of Uniform LLNs

Under Assumption 5.5.1 and 5.5.3, and $\mathcal{F}$ admits an envelop F with $PF < \infty$, then

$$\sup_{f \in \mathcal{F}} \left|\mathbb{P}_n f - Pf\right| \xrightarrow{\text{a.s.}} 0$$

if and only if both $\log N\left(\epsilon, \mathcal{F}, \|\cdot\|_{1,2}\right)/n^2$ and $\log N\left(\epsilon, \mathcal{F}, \|\cdot\|_{1,1}\right)/n$ tend to 0 in outer probability.

and 2 aspects of dissociated, exchangeable arrays are emphasized:

- **i.i.d. variations**: through the random entropy term related to $\|\cdot\|_{1,2}$, which only involves $\left(U_{\{i_1,i_2\}}\right)_{\mathbf{i} \in \mathbb{I}_{n,2}}$
- **U-statistic**: through the random entropy term related to $\|\cdot\|_{1,1}$, up to negligible terms, $\|f\|_{1,1}$ only depends on $(U_{i_1})_{1 \leq i_1 \leq n}$

5.5.1.3 Convergence of the bootstrap process

Davezies et al. (2021), extending the pigeonhole bootstrap (McCullagh, 2000; Owen, 2007), established the following bootstrap process:

1 n units are sampled independently in $\{1, \cdots, n\}$ with replacement and equal probability, W_i denotes the number of times unit i is sampled.
2 the $k-$tuple $\mathbf{i} = (i_1, \cdots, i_k) \in \mathbb{I}_{n,k}$ is then selected $W_i = \prod_{j=1}^{k} W_{i_j}$ times in the bootstrap sample

then consider $\mathbb{P}_n^*$ and $\mathbb{G}_n^*$ defined on $\mathcal{F}$ by

$$\mathbb{P}_n^* f = \frac{(n-k)!}{n!} \sum_{\mathbf{i} \in \mathbb{I}_{n,k}} W_{\mathbf{i}} f(Y_{\mathbf{i}})$$
$$\mathbb{G}_n^* f = \sqrt{n} \left(\mathbb{P}_n^* f - \mathbb{P}_n f \right)$$

the validity of the bootstrap is then established as:

Theorem 5.5.7: Bootstrapping Validity

Under Assumption 5.5.1 and 5.5.3, if (**B-i**) of Assumption 5.5.4 also holds, the process $\mathbb{G}_n^*$ converges weakly in $l^\infty(\mathcal{F})$ to $\mathbb{G}$, conditional on $(Y_{\mathbf{i}})_{\mathbf{i} \in \mathbb{I}_k}$ and outer almost surely.

The proof of this theorem boils down to proving

$$\sup_{h \in BL_1} \left| \mathbb{E}\left(h\left(\mathbb{G}_n^*\right) \mid (Y_{\mathbf{i}})_{\mathbf{i} \in \mathbb{I}_k} \right) - \mathbb{E}\left(h(\mathbb{G})\right) \right| \xrightarrow{\text{a.s.}_{outer}} 0$$

where BL_1 is the set of bounded and Lipschitz functions from $l^\infty(\mathcal{F})$ to $[0,1]$. With the standard bootstrap for i.i.d. data

$$\mathbb{E}\left[\mathbb{P}_n^*(f) \mid (Y_{\mathbf{i}})_{\mathbf{i} \in \mathbb{I}_k} \right] = \frac{1}{n^k} \sum_{\mathbf{i} \in \mathbb{I}_{n,k}} f(Y_{\mathbf{i}}) \xrightarrow{n \to \infty} \mathbb{P}_n f$$

hence, Davezies et al. (2021) established the a.s. conditional convergence of $\sqrt{n}\left(\mathbb{P}_n^* f - \frac{1}{n^k} \sum_{\mathbf{i} \in \mathbb{I}_{n,k}} f(Y_{\mathbf{i}}) \right)$.

5.5.1.4 Nonlinear estimators

Davezies et al. (2021) considered 2 classes of estimators: **Z-estimators** and **smooth functionals of the empirical cdf**:

Z-estimators Let

- Θ denote a normed space, endowed with norm $\|\cdot\|_\Theta$
- $\left(\psi_{\theta,h}\right)_{(\theta,h) \in \Theta \times \mathcal{H}}$ denote a class of real, measurable functions
- $\Psi(\theta)(h) = P\psi_{\theta,h}$, $\Psi_n(\theta)(h) = \mathbb{P}_n \psi_{\theta,h}$, $\Psi_n^*(\theta)(h) = \mathbb{P}_n^* \psi_{\theta,h}$
- for any real function g on $\mathcal{H}$, $\|g\|_{\mathcal{H}} = \sup_{h \in \mathcal{H}} |g(h)|$

The parameter of interest θ_0, satisfying $\Psi(\theta_0) = 0$, is estimated by $\hat{\theta} = \arg\min_{\theta \in \Theta} \|\Psi_n(\theta)\|_{\mathcal{H}}$, define the bootstrap counterpart of $\hat{\theta}$ as

$$\hat{\theta}^* = \arg\min_{\theta \in \Theta} \|\Psi_n^*(\theta)\|_{\mathcal{H}}$$

then we have the convergence

Theorem 5.5.8: Convergence of Z-estimators Bootstrap

Under Assumption 5.5.1, if also

1 $\|\Psi(\theta_m)\|_{\mathcal{H}} \to 0 \Rightarrow \|\theta_m - \theta_0\|_{\Theta} \to 0, \forall (\theta_m)_{m\in\mathbb{N}} \in \Theta$

2 $\left\{\psi_{\theta,h} : (\theta, h) \in \Theta \times \mathcal{H}\right\}$ satisfies Assumption 5.5.3 and (A) of 5.5.4, with $PF < \infty$

3 $\exists \delta > 0$ s.t. $\left\{\psi_{\theta,h} : \|\theta - \theta_0\|_{\Theta} < \delta, h \in \mathcal{H}\right\}$ satisfies Assumption 5.5.3 and (B) of 5.5.4, with $PF_{\delta}^2 < \infty$

4 $\lim_{\theta\to\theta_0} \sup_{h\in\mathcal{H}} P\left(\psi_{\theta,h} - \psi_{\theta_0,h}\right)^2 = 0$

5 $\forall \eta > 0, \left\|\Psi_n\left(\hat{\theta}\right)\right\|_{\mathcal{H}} = o_p(n^{-1/2})$ and $P\left(\left\|\sqrt{n}\Psi_n^*\left(\hat{\theta}^*\right)\right\|_{\mathcal{H}} > \eta \left|(Y_{\mathbf{i}})_{\mathbf{i}\in\mathbb{I}_k}\right.\right) = o_p(1)$

6 $\theta \mapsto \Psi(\theta)$ is Frechet-differentiable at θ_0, with continuously invertible derivative $\dot{\Psi}_{\theta_0}$

Then the convergence can be established as

- $\sqrt{n}\left(\hat{\theta} - \theta\right)$ converges in distribution to a centered Gaussian process $\mathbb{G}$
- conditional on $(Y_{\mathbf{i}})_{\mathbf{i}\in\mathbb{I}_k}$, $\sqrt{n}\left(\hat{\theta}^* - \hat{\theta}\right) \xrightarrow{d} \mathbb{G}$ almost surely

Smooth functionals of F_Y For the cdf of $Y_{\mathbf{i}}$, suppose that $\mathcal{Y} \subset \mathbb{R}^p$ for some $p \in \mathbb{N}^+$ and $\theta_0 = g(F_Y)$, where g is Hadamard differentiable[13]. Estimate θ_0 with $\hat{\theta} = g\left(\hat{F}_Y\right)$, where $\hat{F}_Y$ denotes the empirical cdf of $(Y_{\mathbf{i}})_{\mathbf{i}\in\mathbb{I}_{n,k}}$, and let $\hat{\theta}^*$ denote the bootstrap counterpart of $\hat{\theta}$. Davezies et al. (2021) established the convergence results as

Theorem 5.5.9: Convergence of Smooth Functionals of the Empirical CDF

Suppose that g is Hadamard differentiable at F_Y tangentially to a set $\mathbb{D}_0$, with derivative equal to g'_{F_Y}. Under Assumption 5.5.1,

- $\sqrt{n}\left(\hat{F}_Y - F_Y\right)$ converges weakly, as a process indexed by y, to a Gaussian process $\mathbb{G}$ with kernel K satisfying

$$K(y_1, y_2) = \frac{1}{(k-1)!^2} \sum_{(\pi,\pi')\in\mathfrak{S}(\{\mathbf{1}\})\times\mathfrak{S}(\{\mathbf{1}'\})} \operatorname{Cov}\left(\mathbf{1}_{\{Y_{\pi(\mathbf{1})}\le y_1\}}, \mathbf{1}_{\{Y_{\pi'(\mathbf{1}')}\le y_2\}}\right)$$

- If $\mathbb{G} \in \mathbb{D}_0$[a] with probability 1,

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \mathbb{V}\left(g'_{F_Y}(\mathbb{G})\right)\right)$$

conditional on $(Y_{\mathbf{i}})_{\mathbf{i}\in\mathbb{I}_k}$, $\sqrt{n}\left(\hat{\theta}^* - \hat{\theta}\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \mathbb{V}\left(g'_{F_Y}(\mathbb{G})\right)\right)$, almost surely.

[a] In practice, $\mathbb{D}_0$ often corresponds to the set of functions that are continuous everywhere or at a certain point y_0.

5.5.1.5 Extensions

Davezies et al. (2021) also considered several extensions of the main results:

- **Degenerate cases**: consider the simple $k = 2$ situations where $K(f, f) = 0, \forall f \in \mathcal{F}$. Generally, when $K(f, f) = 0$, the rate of convergence of $\mathbb{P}_n f - Pf$ is n^{-1} rather than $n^{-1/2}$, and the asymptotic distribution

[13] No linearity assumed under Hadamard differentiability.

is not necessarily normal. $\forall (i_1, i_2) \in \mathbb{I}_2$, let $Y_{i_1,i_2} = \tau\left(U_{i_1}, U_{i_2}, U_{\{i_1,i_2\}}\right)$ be the Aldous-Hoover-Kallenberg representation. WLoG, assume $U_\cdot$ to be uniform on $[0,1]$.

Under a more stringent version of (**B**-i) Assumption 5.5.4, that is, $\mathcal{F}$ admits an envelope F with $PF^2 < \infty$ and

$$\int_0^\infty \sup_{Q \in \mathcal{Q}} \log N\left(\eta \|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}\right) \mathrm{d}\eta < \infty$$

Davezies et al. (2021) showed that for $k = 2$, $K(f,f) = 0$ for all $f \in \mathcal{F}$, $\sqrt{n}\mathbf{G}_n$ converges **weakly** in $l^\infty(\mathcal{F})$ to $\mathbb{G}^d$. However, the proposed bootstrap process does **not** generally converge to $\mathbb{G}^d$.

- **An alternative bootstrap process**: Davezies et al. (2021) established that under Assumption 5.5.1 and 5.5.3, $Pf^2 < \infty$ for all $f \in \mathcal{F}$ and $\mathcal{F}$ admits an envelope F s.t. $PF^{1+\delta} < \infty$ for some $\delta > 0$, then if conditional on $(Y_\mathbf{i})_{\mathbf{i} \in \mathbb{I}_k}$, the process $\mathbb{G}_n^*$ outer almost surely converges weakly in $l^\infty(\mathcal{F})$ to a centered Gaussian process $\mathbb{G}$, the process $\mathbb{G}_n$ also converges weakly in $l^\infty(\mathcal{F})$ to $\mathbb{G}$. Combined with Theorem 5.5.7, we have

$$\mathbb{G}_n \to \mathbb{G} \Leftrightarrow \mathbb{G}_n^* \xrightarrow{\text{a.s.}_{outer}} \mathbb{G}$$

consider an alternative, the multiplier bootstrap process adapted to jointly exchangeable arrays. Let $(\xi_i)_{i=1}^n$ be a sequence of i.i.d. random variables, independent from the original data $(Y_\mathbf{i})_{\mathbf{i} \in \mathbb{I}_{n,2}}$, then the process

$$\mathbb{G}_n^{m*} : f \mapsto \frac{1}{\sqrt{n}} \sum_{i_1=1}^n \xi_{i_1} \left(\frac{1}{n-1} \sum_{1 \le i_2 \ne i_1 \le n} \left[f(Y_{i_1,i_2}) + f(Y_{i_2,i_1}) \right] - 2\mathbb{P}_n f \right)$$

also outer almost surely converges weakly in $l^\infty(\mathcal{F})$ to $\mathbb{G}$, just as the proposed process $\mathbb{G}_n^*$.

- **Separately exchangeable arrays** For the case of separately exchangeable arrays (the n units stem from k **different** populations, or multiway clustering as in Menzel (2021)), we must assume a stronger version of Assumption 5.5.1

Assumption 5.5.10: Stronger assumptions for separately exchangeable arrays

Consider randome variables $Y_\mathbf{i}$ where $\mathbf{i} = (i_1, \cdots, i_k) \in \mathbb{N}^{+k}$, implying that repetitions are allowed. Then assume that for any $(\pi_1, \cdots, \pi_k) \in \mathfrak{S}(\mathbb{N}^+)^k$,

$$(Y_\mathbf{i})_{\mathbf{i} \in \mathbb{N}^{+k}} \overset{d}{=} \left(Y_{\pi_1(i_1), \cdots, \pi_k(i_k)}\right)_{\mathbf{i} \in \mathbb{N}^{+k}}$$

and for any A, B, disjoint subsets of $\mathbb{N}^+$, $(Y_\mathbf{i})_{\mathbf{i} \in A^k}$ is independent of $(Y_\mathbf{i})_{\mathbf{i} \in B^k}$.

Under this stronger assumption, we have equality in distribution even for $\pi_1 = \cdots = \pi_k$. Here, denote

- $\mathbf{1} = (1, \cdots, 1)$
- $\mathbf{n} = (n_1, \cdots, n_k)$, where $n_j \ge 1$ denotes the number of units observed in population (or cluster) j, in general, $n_j \ne n_{j'}$ for $j \ne j'$
- sample at hand: $(Y_\mathbf{i})_{\mathbf{1} \le \mathbf{i} \le \mathbf{n}}$, where $\mathbf{i} \ge \mathbf{i}'$ means that $i_j \ge i_j'$, $\forall j = 1, \cdots, k$.
- $\underline{\mathrm{n}} = \min(n_1, \cdots, n_k)$

then the empirical measure and empirical process for separately exchagneable arrays are

$$\mathbb{P}_n f = \frac{1}{\prod_{j=1}^k n_j} \sum_{\mathbf{1} \le \mathbf{i} \le \mathbf{n}} f(Y_\mathbf{i}) \qquad \mathbb{G}_n f = \sqrt{\underline{\mathrm{n}}} \left(\mathbb{P}_n f - Pf\right)$$

Consider the **pigeonhole bootstrap** process (McCullagh, 2000), which is close the bootstrap in Theorem 5.5.7, except that weights are now independent from one coordinate to another:

1 For each $j \in \{1, \cdots, k\}$, n_j elements are sampled with replacement and equal probability in the set $\{1, \cdots, n_j\}$. For each i_j, let $W_{i_j}^j$ denote the number of times i_j is selected

2 k-tuple $\mathbf{i} = (i_1, \cdots, i_k)$ is then selected $W_{\mathbf{i}} = \prod_{j=1}^k W_{i_j}^j$ times in the bootstrap sample

the bootstrap process $\mathbb{G}_{\mathbf{n}}^*$ is then defined on $\mathcal{F}$ by

$$\mathbb{G}_{\mathbf{n}}^* f = \sqrt{\underline{n}} \left(\frac{1}{\prod_{j=1}^k n_j} \sum_{\mathbf{1} \leq \mathbf{i} \leq \mathbf{n}} (W_{\mathbf{i}} - 1) f(Y_{\mathbf{i}}) \right)$$

as with multisample U-statistics, assume an index $m \in \mathbb{N}^+$ and increasing functions $g_1, \cdots, g_k$ s.t. for all j, $n_j = g_j(m) \xrightarrow{m \to \infty} \infty$, and w.l.o.g., $\forall m \in \mathbb{N}^+, \exists j$ s.t. $g_j(m+1) > g_j(m)$, then

Theorem 5.5.11: Bootstrap Convergence: Separately Exchangeable Arrays

Under Assumption 5.5.3 and 5.5.10 and for every $j = 1, \cdots, k$, $\exists \lambda_j \geq 0$ s.t. $\underline{n}/n_j \to \lambda_j \geq 0$, then

1 If (A) of Assumption 5.5.4 holds, $\sup_{f \in \mathcal{F}} \left| \mathbb{P}_{\mathbf{n}} f - Pf \right| \xrightarrow{\text{a.s.}} 0$ and in L^1

2 If (B-i) of Assumption 5.5.4 holds, the process $\mathbb{G}_n$ coverges weakly in $l^\infty(\mathcal{F})$ to a centered Gaussian process $\mathbb{G}_\lambda$ on $\mathcal{F}$ as $n \to \infty$, and the covariance kernel K_λ of $\mathbb{G}_\lambda$ satisfies

$$K_\lambda(f_1, f_2) = \sum_{j=1}^k \lambda_j \text{Cov}\left(f_1(Y_{\mathbf{1}}), f_1\left(Y_{\mathbf{2}_j}\right) \right)$$

where $\mathbf{2}_j$ is the k-tuple with 2 in each entry but 1 in entry j.

3 If (B-i) of Assumption 5.5.4 holds, $\mathbb{G}_n^* \to \mathbb{G}_\lambda$ weakly, conditional on $(Y_{\mathbf{i}})_{\mathbf{i} \in \mathbb{N}^{+k}}$ and outer almost surely

Here, the case where $\lambda_j = 0$ for some j corresponds to **strongly unbalanced** designs with different rates of convergence to ∞ along the different dimensions of the array. In such case, only the dimensions with the slowest rate of convergence contribute to the asymptotic distribution.

5.5.2 CLT for the estimator with heterogeneous clusters

Yap (2023) provides general conditions such that the plug-in mean estimator is asymptotically normal, and the Cameron et al. (2011) variance estimator is consistent even when clusters are heterogeneous. The conditions mimic one-way clustering conditions, assuming that two observations are independent when they do not share any cluster, and **not** assuming separate exchangeability.

Consider for vectors $\{\mathbf{W}_i\}_{i=1}^n$, where $\mathbf{W}_i := (W_{i1}, \cdots, W_{iK})' \in \mathbb{R}^K$ and i is the unit of observation, for the population of size n. The goal is to establish a central list theorem (CLT) for a weighted sum of the random vector, $\sum_i \omega_i \mathbf{W}_i$ where ω_i are non-stochastic scalar weights as $n \to \infty$.

Notation Consider 2 clustering dimension G and H, then

- $g(i), h(i)$: the cluster where observations i belongs on the G and H dimensions, respectively
- **partition**: For $C \in \{G, H\}$, let $\mathcal{N}_c^C$ denote the set of observations in cluster c on dimension C
- $N_c^C = \left| \mathcal{N}_c^C \right|$: the cluster size for $C \in \{G, H\}$, and $N_{gh} := \left| \mathcal{N}_g^G \cap \mathcal{N}_h^H \right|$

Assumptions Several assumptions are imposed to establish the main result

Assumption 5.5.12: Assumptions of Yap (2023)

1 **dependence structure**: $\mathbf{W}_i \perp \mathbf{W}_j$ if $g(i) \neq g(j)$ and $h(i) \neq h(j)$
2 **additional assumptions**: For $C \in \{G, H\}$ and $k \in \{1, 2, \cdots, K\}$, $\exists K_0 < \infty$ s.t.
   - $\forall i, \mathbb{E}\left[W_{ik}^4\right] \leq K_0$: bounded fourth moment, stronger than the one-way clustering condition.
   - $\frac{1}{\lambda_n} \max_c \left(\sum_{i \in \mathcal{N}_c^C} |\omega_i|\right)^2 \to 0$: the cluster with the largest weight to have relatively **small** contribution to the total variance, s.t. removing one cluster does not change the variance substantively, allowing the ratio of the size of any 2 clusters to diverge.
   - $\frac{1}{\lambda_n} \sum_c \sum_{i,j \in \mathcal{N}_c^C} A_{ij} \left|\omega_i \omega_j\right| \leq K_0$: ruling out the purely interactive model in Menzel (2021)

The key feature of Assumption 5.5.12 (1) is that it is agnostic about the dependent structure when W_i and W_j share at least one cluster. The DGP can be arbitrarily heterogeneous across different clusters. Consider a $0-1$ indicator:

$$A_{ij} \coloneqq \mathbf{1}\left[\mathbf{W}_i \not\perp \mathbf{W}_j\right]$$

then we have $A_{ij} = A_{ji}$ and $A_{ii} = 1$. Yap (2023) established the following results

Theorem 5.5.13: Asymptotic Normality of Yap (2023)

Under Assumption 5.5.12,

$$Q_n^{-1/2} \sum_{i=1}^{n} \omega_i \left(\mathbf{W}_i - \mathbb{E}\left[\mathbf{W}_i\right]\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \mathbf{I}_K\right)$$

further

- If $\mathbb{E}\left[\mathbf{W}_i\right] = 0, \forall i$, then $Q_n^{-1} \hat{Q}_n \xrightarrow{\mathrm{p}} \mathbf{I}_K$, where $\hat{Q}_n \coloneqq \sum_i \sum_{j \in \mathcal{N}_i} \omega_i \omega_j \mathbf{W}_i \mathbf{W}_j'$
- If $\mathbb{E}\left[\mathbf{W}_i\right] = \mu, \forall i$[a] and $\frac{1}{\lambda_n} \sum_c \sum_{i,j \in \mathcal{N}_c^C} \left|\omega_i \omega_j\right| \leq K_0$ for some $K_0 < \infty$, then

$$\overline{\mathbf{W}} \xrightarrow{\mathrm{p}} \mu \qquad\qquad Q_n^{-1} \hat{Q}_n \xrightarrow{\mathrm{p}} \mathbf{I}_K$$

for $\hat{\mathbf{W}} = \left(\sum_i \omega_i \mathbf{W}_i\right) / \left(\sum_j \omega_j\right)$ and $\hat{Q}_n \coloneqq \sum_i \sum_{j \in \mathcal{N}_i} \omega_i \omega_j \left(\mathbf{W}_i - \overline{\mathbf{W}}\right)\left(\mathbf{W}_j - \overline{\mathbf{W}}\right)'$

one-way clustering is a special case of these results when one dimension is weakly nested within the other.

[a] Here, the variance estimator need not be consistent, or even conservative. See Remark 2 of Yap (2023) for details.

5.5.3 One-way robust clustering with large clusters

Sasaki and Wang (2022) established robust clustering for the cases wher the distribution of cluster sizes follows a power law with exponent less than 2, the conventional clustering method fails[14]. Conssider again the linear model

$$Y_{gi} = \mathbf{X}_{gi}' \boldsymbol{\theta} + U_{gi} \qquad\qquad \mathbb{E}\left[\mathbf{U}_g \mid \mathbf{X}_g\right] = 0$$

[14] That is, $\sup_g N_g / N \nrightarrow 0$. An example is that California consists of 10% of total sample among the 51 US states.

with a clustered sample $\left\{\left\{\left(Y_{gi}, \mathbf{X}'_{gi}\right)'\right\}_{i=1}^{N_g}\right\}_{g=1}^{G}$. The OLS estimator gives

$$\hat{\boldsymbol{\theta}} = \left(\sum_{g=1}^{G}\sum_{i=1}^{N_g}\mathbf{X}_g i\mathbf{X}'_{gi}\right)^{-1}\left(\sum_{g=1}^{G}\sum_{i=1}^{N_g}\mathbf{X}_{gi}Y_{gi}\right) = \left(\sum_{g=1}^{G}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{g=1}^{G}\mathbf{X}'_g\mathbf{Y}_g\right)$$

and the commonly cluster-robust variance estimators take the form of

$$\hat{V}^{\text{CR}}_{\hat{\boldsymbol{\theta}}} = a_n\left(\sum_{g=1}^{G}\mathbf{X}'_g\mathbf{X}_g\right)^{-1}\left(\sum_{g=1}^{G}\hat{\mathbf{S}}_g\hat{\mathbf{S}}'_g\right)\left(\sum_{g=1}^{G}\mathbf{X}'_g\mathbf{X}_g\right)^{-1}$$

where $a_n \to 1$ is a suitable finite-sample adjustment, and $\hat{\mathbf{S}}_g = \sum_i^{N_g} = 1\mathbf{X}_{gi}\hat{U}_{gi}$ with $\hat{U}_{gi} = Y_{gi} - \mathbf{X}'_{gi}\hat{\boldsymbol{\theta}}$. Conventionally, the adjustment is

$$a_n = \left(\frac{\sum_{g=1}^{G}N_g - 1}{\sum_{g=1}^{G}N_g - p}\right)\left(\frac{G}{G-1}\right)$$

with p denoting the dimension of $\mathbf{X}_{gi}$. Asymptotically,

$$\sqrt{G}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) = \underbrace{\left(\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{N_g}\mathbf{X}_{gi}\mathbf{X}'_{gi}\right)^{-1}}_{\xrightarrow{\text{p}}\mathbf{Q}}\underbrace{\left(\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\sum_{i=1}^{N_g}\mathbf{X}_{gi}U_{gi}\right)}_{\xrightarrow{\text{d}}\mathcal{N}(0,\mathbf{V})} \xrightarrow{\text{d}} \mathcal{N}\left(\mathbf{0}, \mathbf{Q}^{-1}\mathbf{V}\mathbf{Q}^{-1}\right)$$

where $\mathbf{Q} = \mathbb{E}\left[\sum_{i=1}^{N_g}\mathbf{X}_{gi}\mathbf{X}'_{gi}\right]$, $\mathbf{V} = \text{Var}\left[\sum_{i=1}^{N_g}\mathbf{X}_{gi}U_{gi}\right]$ and $\sum_{i=1}^{N_g}\mathbf{X}_{gi}U_{gi}$ have finite second moments.

Main results for a given $k \in \{1, \cdots, p\}$, let $\sum_g$ and Z_{gi} be the k-th coordinate of $\sum_{i=1}^{N_g}\mathbf{X}_{gi}U_{gi}$ and the $k-$th coordinate of $\mathbf{X}_{gi}U_{gi}$ respectively. Consider the following property, **regularly varying (RV)**, of a distribution function F,

$$\frac{1 - F(xt)}{1 - F(t)} \xrightarrow{t\to\infty} x^{-\alpha}, \ \forall x > 0$$

for some constant $\alpha > 0$, which is referred to as the **tail exponent**, measuring the tail heaviness of F. Then, let

- F denote the marginal distribution of Z_{gi}
- C^n denote, for each $n \geq 2$, the copula s.t.

$$\mathbb{P}\left(Z_{g1} \leq z_1, \cdots, Z_{gn} \leq z_n\right) = C^n\left(F(z_1), \cdots, F(z_n)\right)$$

Sasaki and Wang (2022) make the following assumptions

Assumption 5.5.14: Assumptions on the Distribution of $\{Z_{gi}\}_{g,i}$

- $\{Z_{gi}\}_{g,i}$ is identically distributed with CDF F, which is **RV** at infinity with $\alpha > 1$[a]
- The copula density $c(u_1, \cdots, u_n) = \partial^n C(u_1, \cdots, u_n)/\partial u_1 \cdots \partial u_n$ exists and is **uniformly bounded**[b]
- N_g is independent of $(Z_{g1}, Z_{g2}, \cdots)$ and its distribution H is **RV** at infinity with $\beta > 1$[c]

[a] Z_{gi} should have a finite mean, and a regularly varying tail (satisfied by many common heavy-tailed distributions including Pareto, Student-t, Cauchy, F). The tail condition cahracterize the moment conditions as

$$\mathbb{E}\left[|Z_{gi}|^r\right] = \infty \forall r < \alpha \qquad \mathbb{E}\left[|Z_{gi}|^r\right] < \infty \forall r > \alpha$$

[b] Allowing dependence among $Z_{g1}, \cdots, Z_{gN_g}$ within each cluster g

[c] The distribution of the cluster size N_g is also regularly varying. One way to think about this is considering N_g as the integer part of some continuous random variable with a regularly varying tail.

Under Assumption 5.5.14, Sasaki and Wang (2022) establish that

Theorem 5.5.15: When Conventional Robust Clustering Fails

If $\beta < \alpha$, then $\forall z > 0$, as $t \to \infty$

$$\frac{\mathbb{P}\left(\sum_g > zt\right)}{\mathbb{P}\left(\sum_g > t\right)} = z^{-\beta}$$

here, the tail of the distribution of N_g dominates that of Z_{gi}, the tail heaviness of the summation $\sum_g = \sum_{i=1}^{N_g} Z_{gi}$ is dictated by that of N_g. Therefore, even if Z_{gi} has a finite $r-$th moment for $r < \alpha$, the $r-$th moment of $\sum_g$ might still be infinite if $r > \beta$. Hence, the conventional robust clustering may fail to exist even if Z_{gi} has a bounded second moment. Sasaki and Wang (2022) illustrated this problem with state clustering of PSID and some recent studies published on Econometrica. The way they choose to show that $\beta < 2$ is by using the Hill plot (Drees et al., 2000).

Proposed alternative estimators Sasaki and Wang (2022) proposed an simple fix as

$$\hat{\boldsymbol{\theta}}^{\text{WCR}} = \left(\sum_{g=1}^{G} N_g^{-1} \sum_{i=1}^{N_g} \mathbf{X}_g i \mathbf{X}'_{gi}\right)^{-1} \left(\sum_{g=1}^{G} N_g^{-1} \sum_{i=1}^{N_g} \mathbf{X}_{gi} Y_{gi}\right) = \left(\sum_{g=1}^{G} N_g^{-1} \mathbf{X}'\mathbf{X}\right)^{-1} \left(\sum_{g=1}^{G} N_g^{-1} \mathbf{X}'_g \mathbf{Y}_g\right)$$

$$\hat{V}^{\text{WCR}}_{\hat{\boldsymbol{\theta}}} = a_n \left(\sum_{g=1}^{G} N_g^{-1} \mathbf{X}'_g \mathbf{X}_g\right)^{-1} \left(\sum_{g=1}^{G} N_g^{-2} \hat{\mathbf{S}}_g \hat{\mathbf{S}}'_g\right) \left(\sum_{g=1}^{G} N_g^{-1} \mathbf{X}'_g \mathbf{X}_g\right)^{-1}$$

where $a_n \to 1$. Again, for a given $k \in \{1, \cdots, p\}$, let $\sum_g$ and Z_{gi} be the k-th coordinate of $\sum_{i=1}^{N_g} \mathbf{X}_{gi} U_{gi}$ and the $k-$th coordinate of $\mathbf{X}_{gi} U_{gi}$ respectively, and then similar to Thm.5.5.15,

Theorem 5.5.16: Sasaki and Wang (2022)'s Modification

Under Assumption 5.5.14, $\forall z > 0$, as $t \to \infty$, $\frac{\mathbb{P}(\tilde{\sum}_g > zt)}{\mathbb{P}(\tilde{\sum}_g > t)} = z^{-\beta}$ where $\tilde{\sum}_g = \sum_g / N_g = \sum_{i=1}^{N_g} Z_{gi}$

Thm.5.5.16 established that $\tilde{\sum}_g$ has the same tailthe original score Z_{gi}. It requires the second moments of

the score $\mathbf{X}_{gi}U_{gi}$ is **bounded** regardless of whether the second moment of N_g if finite or not.

Theorem 5.5.17: CLT of Sasaki and Wang (2022)'s Clustering

Under Assumption 5.5.14, and in addition, assume that $(N_g, \mathbf{X}_g, \mathbf{X}_g)$ is i.i.d. across g[a], and $\mathbb{E}\left[N_g^{-1}\mathbf{X}_g'\mathbf{X}_g\right]$ is non-singular[b]. With $\alpha > 2$, then

$$\sqrt{G}\left(\hat{\boldsymbol{\theta}}^{\text{WCR}} - \boldsymbol{\theta}\right) \xrightarrow{\text{p}} \mathcal{N}\left(\mathbf{0}, \mathbf{V}^{\text{WCR}}\right)$$

as $G \to \infty$, where

$$\mathbf{V}^{\text{WCR}} = \left(\mathbb{E}\left[N_g^{-1}\mathbf{X}_g'\mathbf{X}_g\right]\right)^{-1}\left(\mathbb{E}\left[N_g^{-2}\mathbf{X}_g'\mathbf{X}_g\mathbf{U}_g^2\right]\right)\left(\mathbb{E}\left[N_g^{-1}\mathbf{X}_g'\mathbf{X}_g\right]\right)^{-1}$$

and $G\hat{\mathbf{V}}^{\text{WCR}} \xrightarrow{\text{p}} \mathbf{V}_{\hat{\boldsymbol{\theta}}}^{\text{WCR}}$ as $G \to \infty$.

[a] Requiring the i.i.d. sampling **across** clusters, while allowing for arbitrary dependence within each cluster.
[b] Ruling out multi-collinearity.

Thm.5.5.13 gives that the proposed clustering approach based on $\hat{\boldsymbol{\theta}}^{\text{WCR}}$ works as far as $\alpha > 2$ is true, regardless of whether $\beta < 2$ is true or not.

5.5.4 Two-Way Robust Clustering: Justification

Chiang and Sasaki (2023) developed a CLT for means of two-way clustered triangular arrays under mild conditions, which provides a theoretical justification for the asymptotic gaussianity of the statistics commonly occurs under two-way clustering.

Consider

$$D_{it} = f_{NT}\left(\alpha_i, \gamma_t, \epsilon_{it}\right)$$

normalize it, w.l.o.g., to $\mathbb{E}\left[D_{it}\right] = 0$, define

$$\hat{\theta}_{NT} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}$$

the goal is to show the **asymptotic Gaussianity** of $\hat{\theta}_{NT}$. The Hoeffding-type decomposition gives

$$\hat{\theta}_{NT} = \underbrace{\sum_{i=1}^{N} a_i + \sum_{t=1}^{T} b_t}_{:=L_{NT}} + \underbrace{\sum_{i=1}^{N}\sum_{t=1}^{T} w_{it}}_{:=W_{NT}} + \underbrace{\sum_{i=1}^{N}\sum_{t=1}^{T} r_{it}}_{:=R_{NT}}$$

where

$$a_i = \frac{1}{N}\mathbb{E}\left[D_{it} \mid \alpha_i\right] \qquad b_t = \frac{1}{T}\mathbb{E}\left[D_{it} \mid \gamma_t\right]$$

$$w_{it} = \frac{1}{NT}\left(\mathbb{E}\left[D_{it} \mid \alpha_i, \gamma_t\right] - \mathbb{E}\left[D_{it} \mid \alpha_i\right] - \mathbb{E}\left[D_{it} \mid \gamma_t\right]\right) \qquad r_{it} = \frac{1}{NT}\left(D_{it} - \mathbb{E}\left[D_{it} \mid \alpha_i, \gamma_t\right]\right)$$

straightforwardly,

$$\begin{aligned}\mathbb{E}[a_i] = \mathbb{E}[b_t] = \mathbb{E}[w_{it}] = \mathbb{E}[r_{it}] = 0\\ \mathbb{E}[w_{it} \mid \alpha_i] = \mathbb{E}\left[w_{it} \mid \gamma_t\right] = 0\\ \mathbb{E}[a_i w_{it}] = \mathbb{E}[a_i r_{it}] = \mathbb{E}\left[\gamma_t w_{it}\right] = \mathbb{E}\left[\gamma_t r_{it}\right] = 0\end{aligned}$$

here, $W_{NT} = \sum_{i=1}^{N}\sum_{t=1}^{T} w_{it}$ is the potentially non-guassian part, as discussed by Menzel (2021). Chiang and Sasaki (2023) argue that if the DGP is treated as a triangular array, W_{NT} is often asymptotically gaussian. For a simple, generic DGP

$$D_{it} = \alpha_{i0} + \gamma_{t0} + \sum_{j=1}^{J} \lambda_j \alpha_{ij}\gamma_{tj} + \epsilon_{ij}$$

where $\{\alpha_{ij}\}_{j=0}^{J}$ are $i-$specific latent factors, $\{\gamma_{tj}\}_{j=0}^{J}$ are $t-$specific latent factors, ϵ_{it} is an idiosyncratic component. Suppose that the factor loading λ_j are non-zero, then Chiang and Sasaki (2023) provide a summary on when to use the TWCR standard error for $\hat{\theta} = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}$: Except for the extreme

Table 5.1: Chiang and Sasaki (2023)'s Summary on TWCR SE Validity

Small J	α_{i0}	γ_{t0}	ϵ_{it}	TWCR SE valid?
Yes	Degenerate	Degenerate	Degenerate	**No**
Other cases				**Yes**

case where the number of factors J is **small**, latent factors on both dimension i and t are degenerate, the idiosyncratic component ϵ_{it} is also degenerate.

5.6 Clustering in Experiments

5.6.1 When to cluster

Abadie et al. (2023) highlight 3 common misconceptions on clustering adjustments:

- clustering when there is a nonzero correlation between residuals for units within the same cluster
- when clustering makes a difference, one should cluster
- either fully adjust for clustering, or not at all

and they propose a new design-based clustering framework to address the overuse and overconvservatism of the commonly used clustering.

They propose a framework with **3 sources** of sampling variations:

1 variation across samples in which units are observed in each cluster
2 variation in which clusters are observe
3 variation in the treatment assignment across units

Standard framework for clustering focuses on the first 2 sources of uncertainty. How much the 3 sources of variations matter depends on **the sampling process**, **the assignment process**, and **the heterogeneity** in the treatment effects across clusters.

5.6.1.1 Sampling process

Consider a **sequence of populations** indexed by k, and

- the k-th population has n_k units, indexed by $i = 1, \cdots, n_k$
- the population is partitioned into m_k clusters, and $m_{k,i} \in \{1, \cdots, m_k\}$ denote the cluster to which unit i of population k belongs
- number of units in cluster m of population k is $n_{k,m} \geq 1$
- 2 potential outcomes: treated $y_{k,i}(1)$, control $y_{k,i}(0)$

Thus, the population is characterized by triples $\left(m_{k,i}, y_{k,i}(0), y_{k,i}(1)\right)$, for units $1, \cdots, n_k$ and clusters $1, \cdots, m_k$. Then,

- **population ATE**

$$\tau_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \left(y_{k,i}(1) - y_{k,0}(0)\right)$$

- **population ATE by clutser**

$$\tau_{k,m} = \frac{1}{n_{k,m}} \sum_{i=1}^{n_k} \mathbf{1}\left\{m_{k,i} = m\right\} \left(y_{k,i}(1) - y_{k,i}(0)\right)$$

naturally. we have

$$\tau_k = \sum_{m=1}^{m_k} \frac{n_{k,m}}{n_k} \tau_{k,m}$$

For unit i in population k, there are 2 components of the stochastic nature:

- **Sampling process**: a random variable $R_{k,i} = 1$ if unit i belongs to the sample, 0 if not. The sampling process is independent of the potential outcomes and the assignments, and consists of 2 stages
  - **i** clusters are sampled with cluster sampling probability $q_k \in (0, 1]$
    - $q_k = 1$: sample all clusters, i.e. **random sampling**
    - $q_k < 1$: **clustered sampling**
    - $q_k \to 0$: only a small fraction of the clusters sampled
  - **ii** units are sampled from the subpopulation consisting of all the sampled clusters, with probability $p_k \in (0, 1]$
    - $p_k = 1$: sample all units in the population
    - $p_k \to 0$: only a small sample of units sampled
- **Assignment process** the stochastic treatment indicator $W_{k,i} \in \{0, 1\}$ is determined by a 2-stage process as well:
  - **i** for cluster m in population k, **randomly** draw an assignment probability $A_{k,m} \in [0, 1]$ from a distribution with μ_k (bounded away from 0 and 1 uniformly in k), variance σ_k^2, **independently** for each cluster
    - $\sigma_k^2 = 0$: $A_{k,m}$ is the same across all clusters, i.e., **random assignment**
    - $\sigma_k^2 > 0$: assignment probabilities depend on clustering, and
      - **clustered assignment**, $\sigma_k^2 = \mu_k(1 - \mu_k)$: **no within-cluster** variation in $W_{k,i}$[15]
      - **partially clustered assignment**, $0 < \sigma_k^2 < \mu_k(1 - \mu_k)$: assignment depends on cluster but not all units in the same cluster necessarily have the same value of $W_{k,i}$
  - **ii** each unit in cluster m is assigned to the treatment independently, with cluster-specific probability $A_{k,m}$

[15] This is the upper bound of σ_k^2, attained when $A_{k,m}$ can only take the values 0 or 1

5.6.1.2 LS estimator and variance

Let

$$N_{k,1} = \sum_{i=1}^{n_k} R_{k,i} W_{k,i} \qquad N_{k,0} = \sum_{i=1}^{n_k} R_{k,i}(1 - W_{k,i})$$

be the number of **treated** and **untreated** units in the sample (both randomly variables), total sample size is then $N_k = N_{k,1} + N_{k,0}$, then for the regression,

$$Y_{k,i} = \alpha + \tau_k W_{k,i} + \epsilon_{k,i}$$

the OLS estimator of β is equal to the difference in means[16]:

$$\hat{\tau}_k = \frac{1}{N_{k,1} \vee 1} \sum_{i=1}^{n_k} R_{k,i} W_{k,i} Y_{k,i} - \frac{1}{N_{k,0} \vee 1} \sum_{i=1}^{n_k} R_{k,i}(1 - W_{k,i}) Y_{k,i}$$

Abadie et al. (2023) assume that

- $m_k q_k \to \infty$: the expected number of sampled clusters goes to infinity
- $\liminf_{k\to\infty} p_k \min_m n_{k,m} > 0$: average number of observations sampled per cluster (conditional on sampled) does not go to 0
- $\limsup_{k\to\infty} \frac{\max_m n_{k,m}}{\min_n n_{k,m}} < \infty$: the imbalance between the number of units across clusters is bounded

Large k distribution of $\hat{\tau}_k$ Let

$$\alpha_k = \frac{1}{n_k} \sum_{i=1}^{n_k} y_{k,i}(0) \qquad u_{k,i}(1) = y_{k,i} - (\alpha_k + \tau_k) \qquad u_{k,i}(0) = y_{k,i}(0) - \alpha_k$$

then

$$\frac{\sqrt{N_k}\,(\hat{\tau}_k - \tau_k)}{\sqrt{v_k}} \xrightarrow{d} \mathcal{N}(0,1)$$

where

$$\begin{aligned} v_k =& \frac{1}{n_k} \sum_{i=1}^{n_k} \left(\frac{u_{k,i}^2(1)}{\mu_k} + \frac{u_{k,i}^2(0)}{1-\mu_k} \right) \\ & - p_k \frac{1}{n_k} \sum_{i=1}^{n_k} \left(u_{k,i}(1) - u_{k,i}(0) \right)^2 - p_k \sigma_k^2 \frac{1}{n_k} \sum_{i=1}^{n_k} \left(\frac{u_{k,i}(1)}{\mu_k} + \frac{u_{k,i}(0)}{1-\mu_k} \right)^2 \\ & + p_k \left(1 - q_k\right) \frac{1}{n_k} \sum_{m=1}^{m_k} \left(\sum_{i=1}^{n_k} \mathbf{1}\left\{ m_{k,i} = m \right\} \left(u_{k,i}(1) - u_{k,i}(0) \right) \right)^2 \\ & + p_k \sigma_k^2 \frac{1}{n_k} \sum_{m=1}^{m_k} \left(\sum_{i=1}^{n_k} \mathbf{1}\left\{ m_{k,i} = m \right\} \left(\frac{u_{k,i}(1)}{\mu_k} + \frac{u_{k,i}(0)}{1-\mu_k} \right) \right)^2 \end{aligned}$$

Under the following cases:

[16] $\leftarrow a \vee b = \max\{a, b\}$

- **random sampling** ($q_k = 1$) and **random assignment** ($\sigma_k^2 = 0$): then simply v_k as

$$v_k = \underbrace{\frac{1}{n_k}\sum_{i=1}^{n_k}\left(\frac{u_{k,i}^2(1)}{\mu_k} + \frac{u_{k,i}^2(0)}{1-\mu_k}\right)}_{\text{robust variance estimator}} - \underbrace{p_k \frac{1}{n_k}\sum_{i=1}^{n_k}\left(u_{k,i}(1) - u_{k,i}(0)\right)^2}_{\text{finite-sample correction}}$$

  here, the finite-sample correction vanishes if there is either **no heterogeneity** in the treatment effects ($u_{k,i}(1) - u_{k,i}(0) = y_{k,i}(1) - y_{k,i}(0) - \tau_k = 0$), or the sample is **small** ($p_k \simeq 0$)
- **clustered sampling** ($q_k < 1$), the component

$$p_k\left(1 - q_k\right)\frac{1}{n_k}\sum_{m=1}^{m_k}\left(\sum_{i=1}^{n_k}\mathbf{1}\left\{m_{k,i} = m\right\}\left(u_{k,i}(1) - u_{k,i}(0)\right)\right)^2 = p_k(1-q_k)\frac{1}{n_k}\sum_{m=1}^{m_k} n_{k,m}^2\left(\tau_{k,m} - \tau_k\right)^2$$

  this term vanishes when there is no heterogeneity in the average treatment effect across clusters. Information on whether to include this term to adjust for clustered sampling ($q_k < 1$) must come from **outside the sample**.
- **clustered assignment** ($\sigma_k^2 > 0$), then 2 more terms added

$$-p_k\sigma_k^2\frac{1}{n_k}\sum_{i=1}^{n_k}\left(\frac{u_{k,i}(1)}{\mu_k} + \frac{u_{k,i}(0)}{1-\mu_k}\right)^2 + p_k\sigma_k^2\frac{1}{n_k}\sum_{m=1}^{m_k}\left(\sum_{i=1}^{n_k}\mathbf{1}\left\{m_{k,i} = m\right\}\left(\frac{u_{k,i}(1)}{\mu_k} + \frac{u_{k,i}(0)}{1-\mu_k}\right)\right)^2$$

  the sign of it depends on the amount of variation in potential outcomes that can be explained by the clusters. In contrast to the lack of sample information about the need to adjust for clustered sampling, the sample is potentially informative about the need to account for clustered assignment.

The 5 terms of v_k might be of **different** orders:

- $\frac{1}{n_k}\sum_{i=1}^{n_k}\left(\frac{u_{k,i}^2(1)}{\mu_k} + \frac{u_{k,i}^2(0)}{1-\mu_k}\right)$: average of bounded terms, of **order** $O(1)$
- $p_k\frac{1}{n_k}\sum_{i=1}^{n_k}\left(u_{k,i}(1) - u_{k,i}(0)\right)^2 + p_k\sigma_k^2\frac{1}{n_k}\sum_{i=1}^{n_k}\left(\frac{u_{k,i}(1)}{\mu_k} + \frac{u_{k,i}(0)}{1-\mu_k}\right)^2$: **at most** of the same order as the first term. If $p_k \simeq 0$ (the sample is small relative to the population of sampled clusters), dominated by the first term
- the order of the last 2 terms depends on asymptotic of cluster sizes

$$p_k\left(1 - q_k\right)\frac{1}{n_k}\sum_{m=1}^{m_k}\left(\sum_{i=1}^{n_k}\mathbf{1}\left\{m_{k,i} = m\right\}\left(u_{k,i}(1) - u_{k,i}(0)\right)\right)^2$$
$$+ p_k\sigma_k^2\frac{1}{n_k}\sum_{m=1}^{m_k}\left(\sum_{i=1}^{n_k}\mathbf{1}\left\{m_{k,i} = m\right\}\left(\frac{u_{k,i}(1)}{\mu_k} + \frac{u_{k,i}(0)}{1-\mu_k}\right)\right)^2$$

  – if cluster sizes are bounded as k increases: they are also order $O(1)$
  – if cluster sizes increase with k: they can be of higher order and dominate the variance

  it also depends on p_k, clustering in sampling, clustering in assignment, heterogeneity in potential outcomes.

Robust and cluster robust variance estimators Let

$$\hat{U}_{k,i} = Y_{k,i} - \hat{\alpha}_k - \hat{\tau}_k W_{k,i}$$

be the residuals from the regression of $Y_{k,i}$ on a constant and $W_{k,i}$, then the 2 common estimators of the variance of $\sqrt{N_k}\left(\hat{\tau}_k - \tau_k\right)$ are

- **robust variance estimator**:

$$\hat{V}_k^{\text{robust}} = \frac{1}{\overline{W}_k^2\left(1-\overline{W}_k\right)^2}\left\{\frac{1}{N_k}\sum_{i=1}^{n_k} R_{k,i}\hat{U}_{k,i}^2\left(W_{k,i}-\overline{W}_k\right)^2\right\}$$

where $\overline{W}_k = \frac{1}{N_k \vee 1}\sum_{i=1}^{n_k} R_{k,i}W_{k,i}$, let $v_k^{\text{robust}} = \frac{1}{n_k}\sum_{i=1}^{n_k}\left(\frac{u_{k,i}^2(1)}{\mu_k} + \frac{u_{k,i}^2(0)}{1-\mu_k}\right)$, then under some regularity conditions,

$$\frac{\hat{V}_k^{\text{robust}}}{v_k} = \frac{v_k^{\text{robust}}}{v_k} + o_p(1)$$

notice that $v_k^{\text{robust}} - v_k$ can be positive or negative in general, so the robust variance estimator can be invalid in large samples.

- **cluster variance estimator**:

$$\hat{V}_k^{\text{cluster}} = \frac{1}{\overline{W}_k^2(1-\overline{W}_k)^2}\times\left\{\frac{1}{N_k}\sum_{m=1}^{m_k}\left(\sum_{i=1}^{n_k}\mathbf{1}\left\{m_{k,i}=m\right\}R_{k,i}\hat{U}_{k,i}\left(W_{k,i}-\overline{W}_k\right)\right)^2\right\}$$

define

$$\begin{aligned}
v_k^{\text{cluster}} =& \frac{1}{n_k}\sum_{i=1}^{n_k}\left(\frac{u_{k,i}^2(1)}{\mu_k}+\frac{u_{k,i}^2(0)}{1-\mu_k}\right)\\
&- p_k\frac{1}{n_k}\sum_{i=1}^{n_k}\left(u_{k,i}(1)-u_{k,i}(0)\right)^2 - p_k\sigma_k^2\frac{1}{n_k}\sum_{i=1}^{n_k}\left(\frac{u_{k,i}(1)}{\mu_k}+\frac{u_{k,i}(0)}{1-\mu_k}\right)^2\\
&+ p_k\frac{1}{n_k}\sum_{m=1}^{m_k}\left(\sum_{i=1}^{n_k}\mathbf{1}\left\{m_{k,i}=m\right\}\left(u_{k,i}(1)-u_{k,i}(0)\right)\right)^2\\
&+ p_k\sigma_k^2\frac{1}{n_k}\sum_{m=1}^{m_k}\left(\sum_{i=1}^{n_k}\mathbf{1}\left\{m_{k,i}=m\right\}\left(\frac{u_{k,i}(1)}{\mu_k}+\frac{u_{k,i}(0)}{1-\mu_k}\right)\right)^2
\end{aligned}$$

then $\hat{V}_k^{\text{cluster}}$ is close to v_k^{cluster} s.t.

$$\frac{\hat{V}_k^{\text{cluster}}}{v_k} = \frac{v_k^{\text{cluster}}}{v_k} + o_p(1)$$

and $v_k^{\text{cluster}} - v_k$ is always **nonnegative**. For large k, the cluster variance can be conservative.

Comparison Comparing v_k^{robust} and v_k, we have

$$\begin{aligned}
v_k^{\text{robust}} - v_k =& p_k\frac{1}{n_k}\sum_{i=1}^{n_k}\left(u_{k,i}(1)-u_{k,i}(0)\right)^2 - p_k\left(1-q_k\right)\frac{1}{n_k}\sum_{m=1}^{m_k}\left(\sum_{i=1}^{n_k}\mathbf{1}\left\{m_{k,i}=m\right\}\left(u_{k,i}(1)-u_{k,i}(0)\right)\right)^2\\
&+ p_k\sigma_k^2\frac{1}{n_k}\sum_{i=1}^{n_k}\left(\frac{u_{k,i}(1)}{\mu_k}+\frac{u_{k,i}(0)}{1-\mu_k}\right)^2 - p_k\sigma_k^2\frac{1}{n_k}\sum_{m=1}^{m_k}\left(\sum_{i=1}^{n_k}\mathbf{1}\left\{m_{k,i}=m\right\}\left(\frac{u_{k,i}(1)}{\mu_k}+\frac{u_{k,i}(0)}{1-\mu_k}\right)\right)^2
\end{aligned}$$

which consists of 2 terms: the first term

$$p_k\frac{1}{n_k}\left[\sum_{i=1}^{n_k}\left(u_{k,i}(1)-u_{k,i}(0)\right)^2 - \left(1-q_k\right)\sum_{m=1}^{m_k}n_{k,m}^2\left(\tau_{k,m}-\tau_k\right)^2\right]$$

- $= 0$ with **homogeneous treatment effects**: $u_{k,i}(1) - u_{k,i}(0) = 0$, for $i = 1, \cdots, n_k$ and $\tau_{k,m} - \tau_k = 0$ for all $m = 1, \cdots, m_k$
- > 0 with **all clusters sampled**, $q_k = 1$, and **heterogeneous treatment effects**
- **possibly** < 0: $q_k < 1$ (only a fraction of clusters are sampled), and $n_{k,m}^2$ large enough.

and the second term

$$\begin{aligned}
&p_k\sigma_k^2\frac{1}{n_k}\sum_{i=1}^{n_k}\left(\frac{u_{k,i}(1)}{\mu_k}+\frac{u_{k,i}(0)}{1-\mu_k}\right)^2 - p_k\sigma_k^2\frac{1}{n_k}\sum_{m=1}^{m_k}\left(\sum_{i=1}^{n_k}\mathbf{1}\left\{m_{k,i}=m\right\}\left(\frac{u_{k,i}(1)}{\mu_k}+\frac{u_{k,i}(0)}{1-\mu_k}\right)\right)^2 \\
=&p_k\sigma_k^2\sum_{m=1}^{m_k}\frac{n_{k,m}}{n_k}\left[\frac{1}{n_{k,m}}\sum_{i=1}^{n_k}\mathbf{1}\left\{m_{k,i}=m\right\}\left(\frac{u_{k,i}(1)}{\mu_k}+\frac{u_{k,i}(0)}{1-\mu_k}\right)^2\right. \\
&\left. - n_{k,m}\left(\frac{1}{n_{k,m}}\sum_{i=1}^{n_k}\mathbf{1}\left\{m_{k,i}=m\right\}\left(\frac{u_{k,i}(1)}{\mu_k}+\frac{u_{k,i}(0)}{1-\mu_k}\right)\right)^2\right]
\end{aligned}$$

- $= 0$ if there is no clustered assignment: $\sigma_k^2 = 0$
- close to 0 if the heterogeneity in potential outcomes is small: $u_{k,i}(1), u_{k,i}(0)$ close to 0
- > 0 if there is heterogeneity in potential outcomes, but average potential outcomes are **nearly constant** across clusters
- < 0 if the clusters explain enough heterogeneity in potential outcomes, could be very large if $n_{k,m}$ is large

Compare v_k^{cluster} and v_k, we have

$$\begin{aligned}
v_k^{\text{cluster}} - v_k &= p_k q_k \frac{1}{n_k}\sum_{m=1}^{m_k}\left(\sum_{i=1}^{n_k}\mathbf{1}\left\{m_{k,i}=m\right\}\left(u_{k,i}(1)-u_{k,i}(0)\right)\right)^2 \\
&= \left(\frac{p_k n_k}{m_k}\right) q_k \left\{\frac{1}{m_k}\sum_{m=1}^{m_k}\left(\frac{n_{k,m}m_k}{n_k}\right)^2\left(\tau_{k,m}-\tau_k\right)^2\right\}
\end{aligned}$$

hence,

- $v_k^{\text{cluster}} - v_k \geq 0$: cluster standard errors are (very) conservative in general
- $v^{\text{cluster}} - v_k \simeq 0$ when q_k is small (the expected fraction of clusters in the sample), or when the average treatment effect is nearly **constant** between clusters.

5.6.1.3 New variance estimators

Abadie et al. (2023) proposed 2 estimators of the variance of $\hat{\tau}_k$, one analytic based on a correction to $\hat{V}_k^{\text{cluster}}$, one based on resampling.

Random sampling: $q_k = 1$ When all clusters are observed ($q_k = 1$), but allowing for general p_k, let

$$U_{k,i} = W_{k,i}u_{k,i}(1) + (1 - W_{k,i})u_{k,i}(0)$$

first, approximate the normalized error of $\hat{\tau}_k$ by a normalized sample average over clusters

$$\frac{\sqrt{N_k}\left(\hat{\tau}_k - \tau_k\right)}{\sqrt{v_k}} = \frac{1}{\sqrt{n_k p_k v_k}\mu_k\left(1-\mu_k\right)}\sum_{m=1}^{m_k} C_{k,m} + o_p(1)$$

where $C_{k,m} = \sum_{i=1}^{n_k} \mathbf{1}\{m_{k,i} = m\} R_{k,i} (W_{k,i} - \mu_k) U_{k,i}$ are independent across clusters. Then For the term $C_{k,m}$, its expectation $\mathbb{E}[C_{k,m}]$ is **not** 0 in general for each cluster separately, but equals to 0 when sum over **all** clusters:

$$\mathbb{E}[C_{k,m}] = n_{k,m} p_k \mu_k (1 - \mu_k)(\tau_{k,m} - \tau_k) \qquad \sum_{m=1}^{m_k} \mathbb{E}[C_{k,m}] = p_k \mu_k (1 - \mu_k) \sum_{m=1}^{m_k} n_{k,m} (\tau_{k,m} - \tau_k) = 0$$

Now, since we have

$$\frac{\hat{V}_k^{\text{cluster}}}{v_k} = \frac{1}{n_k p_k v_k} \left(\frac{1}{\mu_k(1-\mu_k)} \right)^2 \sum_{m=1}^{m_k} C_{k,m}^2 + o_p(1)$$

and $\text{var}(C_{k,m}) \leq \mathbb{E}\left[C_{k,m}^2\right]$, the expectation of $\frac{1}{n_k p_k v_k} \left(\frac{1}{\mu_k(1-\mu_k)} \right)^2 \sum_{m=1}^{m_k} C_{k,m}^2$ is $\hat{V}_k^{\text{cluster}}$ is bigger than the variance of $\frac{\sqrt{N_k}(\hat{\tau}_k - \tau_k)}{\sqrt{v_k}}$, leading to $\hat{V}_k^{\text{cluster}}$ being conservative.

Plug in $\sum_{m=1}^{m_k} \mathbb{E}[C_{k,m}] = 0$ back into normalized $\hat{\tau}_k$, get

$$\begin{aligned}
\frac{\sqrt{N_k}(\hat{\tau}_k - \tau_k)}{\sqrt{v_k}} &= \frac{1}{\sqrt{n_k p_k v_k} \mu_k (1 - \mu_k)} \sum_{m=1}^{m_k} C_{k,m} + o_p(1) \\
&= \frac{1}{\sqrt{n_k p_k v_k} \mu_k (1 - \mu_k)} \sum_{m=1}^{m_k} \left(C_{k,m} - \mathbb{E}[C_{k,m}] \right) + o_p(1) \\
&= \frac{1}{\sqrt{n_k p_k v_k} \mu_k (1 - \mu_k)} \sum_{m=1}^{m_k} \left(C_{k,m,1} + C_{k,m,2} \right) + o_p(1)
\end{aligned}$$

where

$$\begin{aligned}
C_{k,m,1} &= \sum_{i=1}^{n_k} \mathbf{1}\{m_{k,i} = m\} (R_{k,i} - p_k)(\tau_{k,m} - \tau_k) \mu_k (1 - \mu_k) \\
C_{k,m,2} &= \sum_{i=1}^{n_k} \mathbf{1}\{m_{k,i} = m\} R_{k,i} \left[(W_{k,i} - \mu_k) U_{k,i} - (\tau_{k,m} - \tau_k) \mu_k (1 - \mu_k) \right]
\end{aligned}$$

$C_{k,m,1}$ and $C_{k,m,2}$ are **mean 0** and **uncorrelated**, and **uncorrelated across clusters**, and

- variance of $\frac{\sum_{m=1}^{m_k} C_{k,m,1}}{\sqrt{n_k p_k} \mu_k (1-\mu_k)}$ is
$$(1 - p_k) \sum_{m=1}^{m_k} \frac{n_{k,m}}{n_k} (\tau_{k,m} - \tau_k)^2$$
- a direct estimator of the variance of $\sum_{m=1}^{m_k} C_{k,m,2}$ is
$$\sum_{m=1}^{m_k} \left(\sum_{i=1}^{n_k} \mathbf{1}\{m_{k,i} = m\} R_{k,i} \left(\left(W_{k,i} - \overline{W}_k \right) \hat{U}_{k,i} - \left(\hat{\tau}_{k,m} - \hat{\tau}_k \right) \overline{W}_k \left(1 - \overline{W}_k \right) \right) \right)^2$$
where $\hat{\tau}_{k,m}$ is the estimated treatment effect (sample average) in cluster m. This estimator could be biased from the correlation between the estimation errors of its components (due to $\overline{W}_k$), to address this, Abadie et al. (2023) proposed a sample-splitting estimation procedure:
  1. split the sample randomly into 2 subsamples, indexed by $Z_{k,i} \in \{0, 1\}$, where $Z_{k,i} = 1$ for the second subsample. Let $\overline{Z}_k$ be the mean of $Z_{k,i}$

2 use the first subsample $Z_{k,i} = 0$, get estimates $\hat{\tau}^*_{k,m}$, $\hat{\alpha}^*_k$ and $\hat{\tau}^*_k$
3 for the second subsample $Z_{k,i} = 1$, calculate the residuals $\hat{U}^*_{k,i} = Y_{k,i} - \hat{\alpha}^*_k - \hat{\tau}^*_k W_{k,i}$
4 estimate the normalized variance (for $q_k = 1$) as

$$\begin{aligned}\hat{V}_k^{\text{CCV}}(1) = &\frac{1}{N_k \overline{W}_k^2 \left(1-\overline{W}_k\right)^2} \times \\ &\sum_{m=1}^{m_k} \left[\frac{1}{\overline{Z}_k^2} \left(\sum_{i=1}^{n_k} \mathbf{1}\left\{m_{k,i} = m\right\} R_{k,i} Z_{k,i} \times \left(\left(W_{k,i} - \overline{W}_k\right) \hat{U}^*_{k,i} - \left(\hat{\tau}^*_{k,m} - \hat{\tau}^*_k\right) \overline{W}_k \left(1-\overline{W}_k\right) \right) \right)^2 \right. \\ &\left. - \frac{1-\overline{Z}_k}{\overline{Z}_k^2} \sum_{i=1}^{n_k} \mathbf{1}\left\{m_{k,i} = m\right\} R_{k,i} Z_{k,i} \left(\left(W_{k,i} - \overline{W}_k\right) \hat{U}^*_{k,i} - \left(\hat{\tau}^*_{k,m} - \hat{\tau}^*_k\right) \overline{W}_k \left(1-\overline{W}_k\right) \right)^2 \right] \\ &+ \left(1-p_k\right) \sum_{m=1}^{m_k} \frac{\overline{N}_{k,m}}{N_k} \left(\hat{\tau}_{k,m} - \hat{\tau}_k\right)^2\end{aligned}$$

where $\overline{N}_{k,m}$ is the size of the sample in cluster m. For clusters with no variation in the treatment variable, replace $\hat{\tau}_{k,m}$ with $\hat{\tau}_k$, for clusters with no variation in the treatment variable for a particular subsample, replace $\hat{\tau}^*_{k,m}$ with $\hat{\tau}^*_k$.

for more precise $\hat{V}_k^{\text{CCV}}(1)$, do the sample splitting multiple times and take the average of all variance estimators.

Not all clusters are sampled: $q_k < 1$ First, notice that the variance for the general q_k case is a convex combination of the true variance at $q_k = 1$ and the cluster variance:

$$\begin{aligned} v_k\left(q_k\right) - v_k^{\text{cluster}} &= q_k \times \left(v_k(1) - v_k^{\text{cluster}}\right) \\ \Rightarrow v_k\left(q_k\right) &= q_k \times v_k(1) + \left(1-q_k\right) \times v_k^{\text{cluster}} \end{aligned}$$

then naturally, the variance estimator is

$$\hat{V}_k^{\text{CCV}} = \hat{q}_k \times \hat{V}_k^{\text{CCV}}(1) + \left(1-\hat{q}_k\right) \times \hat{V}_k^{\text{cluster}}$$

where computing $\hat{q}_k$ requires knowledge of m_k, the total number of clusters.

Bootstrap estimator Abadie et al. (2023) proposed a bootstrap procedure where units (clusters) have different assignments and assignment probabilities from the original sample:

- **Input**:
  - Sample $\left(Y_{k,i}, W_{k,i}, m_{k,i}\right)$
  - Fraction sampled clusters q_k
  - Number of bootstrap replications B
- **Stage 1**:

a create pseudo population by replicating each cluster $1/q_k$ times
b for each cluster in the pseudo population, calculate the *assignment probability* $\overline{W}_{k,m}$
c create a bootstrap sample of clusters by *randomly drawing clusters* from the pseudo population from **a**, where cluster m is sampled with probability q_k
d for each sampled cluster, draw an *assignment probability* $A_{k,m}$ from the empirical distribution of the $\overline{W}_{k,m}$ from **b**

- **Stage 2**:
  - **a** randomly draw $\lfloor N_{k,m} A_{k,m} \rfloor$[17] units with replacement from the set of treated units in cluster m
  - **b** randomly draw $\lfloor N_{k,m}(1 - A_{k,m}) \rfloor$ units with replacement from the set of control units in cluster m
- **Calculation**:
  - **1** for the units constructed in Stage 2, collect the values for $(Y_{k,i}, W_{k,i}, m_{k,i})$ and calculate the least-squares or FE estimator
  - **2** calculate the standard deviation of the estimator over the B bootstrap samples

5.6.1.4 Fixed-effects estimator

The fixed-effects estimator is based on a regression of the outcome on the treatment indicator and indicators for each of the clusters in the sample:

$$\hat{\tau}_k^{\text{fixed}} = \frac{\sum_{m=1}^{m_k} \sum_{i=1}^{n_k} \mathbf{1}\{m_{k,i} = m\} R_{k,i} Y_{k,i} \left(W_{k,i} - \overline{W}_{k,m}\right)}{\sum_{m=1}^{m_k} \sum_{i=1}^{n_k} \mathbf{1}\{m_{k,i} = m\} R_{k,i} W_{k,i} \left(W_{k,i} - \overline{W}_{k,m}\right)}$$

assume:

- $m_k q_k / \left(\frac{p_k n_k}{m_k}\right) \to 0$: the expected number of sampled clusters is small relative to the expected number of sampled observations per sampled cluster
  - this implies $p_k n_k / m_k \to \infty$, $n_k p_k q_k \to \infty$, and $p_k \min_m n_{k,m} \to \infty$
- $A_{k,m}$, the supports of the cluster probabilities, are bounded away from 0 and 1

together, $\hat{\tau}_k^{\text{fixed}}$ is well-defined with probability approaching 1.

Let $\alpha_{k,m} = \frac{1}{n_{k,m}} \sum_{i=1}^{n_k} \mathbf{1}\{m_{k,i} = m\} y_{k,i}(0)$. For an observation i, with $m_{k,i} = m$, define the within-cluster residuals $e_{k,i}(0) = y_{k,i}(0) - \alpha_{k,m}$ and $e_{k,i}(1) = y_{k,i} - \tau_{k,m} - \alpha_{k,m}$, let

$$\tilde{v}_k = \frac{f_k}{\left(\mu_k (1 - \mu_k) - \sigma_k^2\right)^2}$$

where

$$\begin{aligned} f_k =& \mathbb{E}\left[A_{k,m}\left(1 - A_{k,m}\right)^2\right] \frac{1}{n_k} \sum_{i=1}^{n_k} e_{k,i}^2(1) + \mathbb{E}\left[A_{k,m}^2\left(1 - A_{k,m}\right)\right] \frac{1}{n_k} \sum_{i=1}^{n_k} e_{k,i}^2(0) \\ & - p_k \mathbb{E}\left[A_{k,m}^2\left(1 - A_{k,m}\right)^2\right] \frac{1}{n_k} \sum_{i=1}^{n_k} \left(e_{k,i}(1) - e_{k,i}(0)\right)^2 \\ & + \left(\mathbb{E}[A_{k,m}(1 - A_{k,m})] - (5 + p_k)\mathbb{E}[A_{k,m}^2(1 - A_{k,m})^2] + 2q_k \left(\mathbb{E}[A_{k,m}(1 - A_{k,m})]\right)^2\right) \sum_{m=1}^{m_k} \frac{n_{k,m}}{n_k} (\tau_{k,m} - \tau_k)^2 \\ & + \left(p_k \mathbb{E}[A_{k,m}^2(1 - A_{k,m})^2] - p_k q_k (\mathbb{E}[A_{k,m}(1 - A_{k,m})])^2\right) \sum_{m=1}^{m_k} \frac{n_{k,m}^2}{n_k} (\tau_{k,m} - \tau_k)^2 \end{aligned}$$

the large-k distribution of the fixed-effects estimator is then

$$\frac{\sqrt{N_k}\left(\hat{\tau}_k^{\text{fixed}} - \tau_k\right)}{\sqrt{\tilde{v}_k}} \xrightarrow{d} \mathcal{N}(0, 1)$$

[17] $\lfloor x \rfloor$ is the floor function takes the largest integer smaller than or equal to x

Robust estimator Let $\tilde{U}_{k,i} = \tilde{Y}_{k,i} - \hat{\tau}_k^{\text{fixed}} \tilde{W}_{k,i}$, where $\tilde{Y}_{k,i} = Y_{k,i} - \overline{Y}_{k,m_{k,i}}$, $\tilde{W}_{k,i} = W_{k,i} - \overline{W}_{k,m_{k,i}}$, the robust estimator of the variance of $\sqrt{N_k}\left(\hat{\tau}_k^{\text{fixed}} - \tau_k\right)$ is

$$\tilde{V}_k^{\text{robust}} = \frac{\frac{1}{N_k}\sum_{i=1}^{n_k} R_{k,i}\tilde{W}_{k,i}^2\tilde{U}_{k,i}^2}{\left(\frac{1}{N_k}\sum_{i=1}^{n_k} R_{k,i}\tilde{W}_{k,i}^2\right)^2}$$

consider

$$\tilde{v}_k^{\text{robust}} = \frac{f_k^{\text{robust}}}{\left(\mu_k\left(1-\mu_k\right) - \sigma_k^2\right)^2}$$

where

$$\begin{aligned} f_k^{\text{robust}} =& \mathbb{E}\left[A_{k,m}\left(1-A_{k,m}\right)^2\right]\frac{1}{n_k}\sum_{i=1}^{n_k} e_{k,i}^2(1) + \mathbb{E}\left[A_{k,m}^2\left(1-A_{k,m}\right)\right]\frac{1}{n_k}\sum_{i=1}^{n_k} e_{k,i}^2(0) \\ & \mathbb{E}\left[A_{k,m}\left(1-A_{k,m}\right)\left(1-3A_{k,m}\left(1-A_{k,m}\right)\right)\right]\sum_{m=1}^{m_k}\frac{n_{k,m}}{n_k}(\tau_{k,m}-\tau_k)^2 \end{aligned}$$

and

$$\tilde{V}_k^{\text{robust}} = \tilde{v}_k^{\text{robust}} + o_p(1)$$

Cluster estimator similarly,

$$\tilde{V}_k^{\text{cluster}} = \frac{\frac{1}{N_k}\sum_{m=1}^{m_k}\left(\sum_{i=1}^{n_k}\mathbf{1}\left\{m_{k,i}=m\right\}R_{k,i}\tilde{W}_{k,i}\tilde{U}_{k,i}\right)^2}{\left(\frac{1}{N_k}\sum_{i=1}^{n_k} R_{k,i}\tilde{W}_{k,i}^2\right)^2}$$

consider

$$\tilde{v}_k^{\text{cluster}} = \frac{f_k^{\text{cluster}}}{\left(\mu_k\left(1-\mu_k\right) - \sigma_k^2\right)^2}$$

where

$$\begin{aligned} f_k =& \mathbb{E}\left[A_{k,m}\left(1-A_{k,m}\right)^2\right]\frac{1}{n_k}\sum_{i=1}^{n_k} e_{k,i}^2(1) + \mathbb{E}\left[A_{k,m}^2\left(1-A_{k,m}\right)\right]\frac{1}{n_k}\sum_{i=1}^{n_k} e_{k,i}^2(0) \\ & - p_k\mathbb{E}\left[A_{k,m}^2\left(1-A_{k,m}\right)^2\right]\frac{1}{n_k}\sum_{i=1}^{n_k}\left(e_{k,i}(1)-e_{k,i}(0)\right)^2 \\ & + \left(\mathbb{E}[A_{k,m}(1-A_{k,m})] - (5+p_k)\mathbb{E}[A_{k,m}^2(1-A_{k,m})^2]\right)\sum_{m=1}^{m_k}\frac{n_{k,m}}{n_k}(\tau_{k,m}-\tau_k)^2 \\ & + p_k\mathbb{E}[A_{k,m}^2(1-A_{k,m})^2]\sum_{m=1}^{m_k}\frac{n_{k,m}^2}{n_k}(\tau_{k,m}-\tau_k)^2 \end{aligned}$$

and then again,

$$\frac{\tilde{V}_k^{\text{cluster}}}{\tilde{v}_k} = \frac{\tilde{v}_k^{\text{cluster}}}{\tilde{v}_k} + o_p(1)$$

Adjusted estimator Similar to the LS case, the robust estimator **underestimate** the true variance, the cluster estimator is generally **conservative**. Abadie et al. (2023) proposed a convex combination of $\tilde{V}_k^{\text{cluster}}$ and $\tilde{V}_k^{\text{robust}}$:

$$\tilde{V}_k^{\text{CCV}} = \hat{\lambda}_k \tilde{V}_k^{\text{cluster}} + \left(1 - \hat{\lambda}_k\right) \tilde{V}_k^{\text{robust}}$$

with a estimated correction weight

$$\hat{\lambda}_k = 1 - \hat{q}_k \frac{\left(\frac{1}{M_k} \sum_{m=1}^{m_k} Q_{k,m} \overline{W}_{k,m} \left(1 - \overline{W}_{k,m}\right)\right)^2}{\frac{1}{M_k} \sum_{m=1}^{m_k} Q_{k,m} \overline{W}_{k,m} \left(1 - \overline{W}_{k,m}\right)}$$

where

- $Q_{k,m}$ is an indicator = 1 if cluster m of population k is sampled
- $M_k = \sum_{m=1}^{m_k} Q_{k,m}$ is the total number of sampled clusters

5.6.2 What Level to Cluster at

In regression analysis, clustered standard errors are reliable when the regression's dependent and independent variables are uncorrelated across clusters (Cameron and Miller, 2015). De Chaisemartin and Ramirez-Cuellar (2024) examine paired RCTs and small-strata experiments, where the treatments of the two units in the same pair are negatively correlated.

5.6.2.1 Setup

Consider a population of $2P$ units, not assume that the units are i.i.d. sample drawn from a superpopulation. Instead, population is fixed, its characteristics are not random[18]. The $2P$ units are matched into P pairs, by grouping together units with the closest value of some baseline variables predicting the outcome, then

- the pairs are indexed by $p \in \{1, \cdots, P\}$
- the two units in pair p are indexed by $g \in \{1, 2\}$
- unit g in pair p has n_{gp} observations: the number of observations in pair p $n_p = n_{1p} + n_{2p}$
- population: $n = \sum_{p=1}^{P} n_p$
- $n_{gp} > 1$ for some units if RCT is clustered

Treatment is assigned as follows: $\forall p \in \{1, \cdots, P\}$ and $g \in \{1, 2\}$, let W_{gp} be an indicator = 1 if unit g in pair p is treated, and = 0 otherwise. Assume

Assumption 5.6.1: Paired assignment

- $\forall p$, $W_{1p} + W_{2p} = 1$: within each pair, one of the two units is treated
- $\mathbb{P}\left(W_{gp} = 1\right) = \frac{1}{2}, \forall g, p$: two units have the same probability of being treated
- $\left(W_{1p}, W_{2p}\right)_{p=1}^{P}$ is jointly independent across p: treatments are independent across pairs

[18] This structure reflects the RCTs where the sample is a convenience sample, consisting of volunteers to receive the treatment, or of units located in areas where conducting the research was easier.

Let $y_{igp}(1)$ and $y_{igp}(0)$ represent the potential outcomes of observation i in unit g and pair p with/without the treatment. Assume that potential outcomes are fixed, the observed outcome is

$$Y_{igp} = y_{igp}(1)W_{gp} + y_{igp}(0)(1 - W_{gp})$$

then the ATE is

$$\tau = \frac{1}{n}\sum_{p=1}^{P}\sum_{g=1}^{2}\sum_{i=1}^{n_{gp}}\left[y_{igp}(1) - y_{igp}(0)\right]$$

Consider 2 estimators of τ, for $i = 1, \cdots, n_{gp}$; $g = 1, 2$; $p = 1, \cdots, P$.:

- OLS estimator of the observed outcome Y_{igp} on a constant and W_{gp}:

$$Y_{igp} = \hat{\alpha} + \hat{\tau}W_{gp} + \epsilon_{igp} \tag{5.25}$$

- pair-fixed-effects estimator of Y_{igp} on W_{gp} and a set of **pair** fixed effects:

$$Y_{igp} = \hat{\tau}_{fe}W_{gp} + \sum_{p=1}^{P}\hat{\gamma}_p\delta_{igp} + u_{igp} \tag{5.26}$$

5.6.2.2 Unit-/Pair-Clustered Variance Estimators

Given the number of treated and untreated observations in pair p as

$$T_p = n_{1p}W_{1p} + n_{2p}W_{2p} \qquad = C_p = n_{1p}\left(1 - W_{1p}\right) + n_{2p}\left(1 - W_{2p}\right)$$

and the total number of treated and untreated observations $T = \sum_{p=1}^{P} T_p$, $C = \sum_{p=1}^{P} C_p$. Let the sum of residuals ϵ_{igp} for the treated and untreated observations in pair p be

$$SET_p = \sum_{g=1}^{2}\sum_{i=1}^{n_{gp}} W_{gp}\epsilon_{igp} \qquad SEU_p = \sum_{g=1}^{2}\sum_{i=1}^{n_{gp}}\left(1 - W_{gp}\right)\epsilon_{igp}$$

then the clustered variance estimators are defined as

- **paired**-clustered variance estimators (PCVEs):

$$\hat{\mathbb{V}}_{\text{pair}}\left(\hat{\tau}\right) = \sum_{p=1}^{P}\left(\frac{SET_p}{T} - \frac{SEU_p}{C}\right)^2$$

$$\hat{\mathbb{V}}_{\text{pair}}\left(\hat{\tau}_{fe}\right) = \sum_{p=1}^{P}\omega_p^2\left(\hat{\tau}_p - \hat{\tau}_{fe}\right)^2$$

- **unit**-clustered variance estimators (UCVEs):

$$\hat{\mathbb{V}}_{\text{unit}}\left(\hat{\tau}\right) = \sum_{p=1}^{P}\left[\left(\frac{SET_p}{T}\right)^2 + \left(\frac{SEU_p}{C}\right)^2\right]$$

$$\hat{\mathbb{V}}_{\text{unit}}\left(\hat{\tau}_{fe}\right) = \sum_{p=1}^{P}\omega_p^2\left(\hat{\tau}_p - \hat{\tau}_{fe}\right)^2\left[\left(\frac{n_{1p}}{n_p}\right)^2 + \left(\frac{n_{2p}}{n_p}\right)^2\right]$$

Assumption 5.6.2: Number of Observations

There is a strictly positive integer N such that for all p, $n_{1p} = n_{2p} = N$.

Assumption 5.6.2 requires that all units have the same number of observations, let

$$\hat{\tau}_p = \sum_g \left[W_{gp} \frac{1}{n_{gp}} \sum_i Y_{igp} - (1 - W_{gp}) \frac{1}{n_{gp}} \sum_i Y_{igp} \right]$$

denote the difference between the average outcome of treated and untreated observations in pair p, then under Assumption 5.6.2,

$$\hat{\tau} = \hat{\tau}_{fe} = \sum_{p=1}^{P} \frac{\hat{\tau}_p}{P}$$

both estimators are unbiased for the ATE, and

$$\mathbb{V}(\hat{\tau}) = \mathbb{V}\left(\hat{\tau}_{fe}\right) = \frac{1}{P^2} \sum_{p=1}^{P} \mathbb{V}\left(\hat{\tau}_p\right)$$

Let

- $\tau_p \equiv \frac{1}{n_p} \sum_{g=1}^{2} \sum_{i=1}^{n_{gp}} \left[y_{igp}(1) - y_{igp}(0) \right]$ be the ATE in pair p
- and $\forall d \in \{0, 1\}$, let
  - $\overline{y}_{gp}(d) \equiv \frac{1}{n_{gp}} \sum_i y_{igp}(d)$ denotes average outcome with treatment d in pair p's unit g
  - $\overline{y}_p(d) \equiv \frac{1}{2} \sum_g \overline{y}_{gp}(d)$ denotes average outcome with treatment d in pair p
  - $\overline{y}(d) \equiv \sum_p \overline{y}_p(d)/P$ denotes average outcome with treatment d in the entire population

Then, we have the following lemma:

Lemma 5.6.3: Comparing Two types of Clustered Variance Estimators

1 Under Assumption 5.6.1 and 5.6.2, $\hat{\mathbb{V}}_{pair}(\hat{\tau}) = \hat{\mathbb{V}}_{pair}(\hat{\tau})_{fe}$, and

$$\mathbb{E}\left[\frac{P}{P-1} \hat{\mathbb{V}}_{pair}(\hat{\tau}) \right] = \mathbb{V}(\hat{\tau}) + \frac{1}{P(P-1)} \sum_{p=1}^{P} (\tau_p - \tau)^2 \geq \mathbb{V}(\hat{\tau})$$

2 Under Assumption 5.6.2, $\hat{\mathbb{V}}_{pair}(\hat{\tau}) = 2\hat{\mathbb{V}}_{unit}\left(\hat{\tau}_{fe}\right)$

3 Under Assumption 5.6.1 and 5.6.2,

$$\begin{aligned} \mathbb{E}\left[\frac{P}{P-1} \left(\hat{\mathbb{V}}_{unit}(\hat{\tau}) - \hat{\mathbb{V}}_{pair}(\hat{\tau}) \right) \right] = & \frac{2}{P} \left(\frac{1}{P-1} \sum_p \left(\overline{y}_p(0) - \overline{y}(0) \right) \left(\overline{y}_p(1) - \overline{y}(1) \right) \right. \\ & \left. - \frac{1}{P} \sum_p \sum_g \frac{1}{2} \left(\overline{y}_{gp}(0) - \overline{y}_p(0) \right) \left(\overline{y}_{gp}(1) - \overline{y}_p(1) \right) \right) \end{aligned}$$

Lemma 5.6.3 states that

1 PCVEs without and with pair fixed effects are equal, and after a DOF correction, their expectation is at least as large as the variance of $\hat{\tau}$:

  - if the treatment effect is heterogeneous across pairs: $\sum_{p=1}^{P} \left(\tau_p - \tau\right)^2 > 0$ (strict inequality), the PCVEs are upward-biased estimators for the variance of $\hat{\tau}$
2. UCVEs are only a half of PCVEs
3. without pair fixed effects, the expectation of the difference between UCVEs and PCVEs is **proportional** to the difference between the between-pair and with-in pair covariance of the two potential outcomes[19]

Intuitively, the UCVEs are biased bacause clustering at the unit level does **not** account for the perfect negative correlation of the treatments of the two units in the same pair. If

- **pair fixed effects are *not* included** in the regression: UCVEs generally **overeestimate** the variance of $\hat{\tau}$
- **pair fixed effects are included** in the regression: UCVEs can **underestimate** the variance of $\hat{\tau}$[20]

[19] In most applications, both terms should be positive since they should be positively correlated.

[20] With pair fixed effects in the regression, the sample residuals u_{igp} are by construction uncorrelated with the pair fixed effects, which implies that for every p,

$$\sum_{i,g} u_{igp} = 0$$

Splitting the summation by $g = 1$ and $g = 2$, under Assumption 5.6.2, the average residuals of observations in unit g of pair p, the previous display implies that $\overline{u}_{1,p} = -\overline{u}_{2,p}$, hence $\left(\overline{u}_{1,p}\right)^2 = \left(\overline{u}_{2,p}\right)^2$, that is, the squares of the average residuals are equal in the treated and control units of each pair. Then with pair fixed effects, the UCVE is proportional to

$$\frac{1}{(2P)^2} \sum_{p=1}^{P} \sum_{g=1}^{2} \left(\overline{u}_{g,p}\right)^2$$

the sum of their average squared residuals across all units, divided by the number of units squared.

References

Alberto Abadie, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. When should you adjust standard errors for clustering? *The Quarterly Journal of Economics*, 138(1):1–35, 2023.

Manuel Arellano. Computing robust standard errors for within-groups estimators. *Oxford bulletin of Economics and Statistics*, 49(4):431–434, 1987.

A Colin Cameron and Douglas L Miller. A practitioner's guide to cluster-robust inference. *Journal of human resources*, 50(2):317–372, 2015.

A Colin Cameron, Jonah B Gelbach, and Douglas L Miller. Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, 29(2):238–249, 2011.

Harold D Chiang and Yuya Sasaki. On using the two-way cluster-robust standard errors. *arXiv preprint arXiv:2301.13775*, 2023.

Laurent Davezies, Xavier D'Haultfœuille, and Yannick Guyonvarch. Empirical process results for exchangeable arrays. *The Annals of Statistics*, 49(2):845–862, 2021.

Clément De Chaisemartin and Jaime Ramirez-Cuellar. At what level should one cluster standard errors in paired and small-strata experiments? *American Economic Journal: Applied Economics*, 16(1):193–212, 2024.

Holger Drees, Sidney Resnick, and Laurens de Haan. How to make a hill plot. *The Annals of Statistics*, 28(1): 254–274, 2000.

Christian B Hansen. Asymptotic properties of a robust variance matrix estimator for panel data when t is large. *Journal of Econometrics*, 141(2):597–620, 2007.

Teunis Kloek. Ols estimation in a model where a microvariable is explained by aggregates and contemporaneous disturbances are equicorrelated. *Econometrica: Journal of the Econometric Society*, pages 205–207, 1981.

Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

Peter McCullagh. Resampling and exchangeable arrays. *Bernoulli*, pages 285–301, 2000.

Konrad Menzel. Bootstrap with cluster-dependence in two or more dimensions. *Econometrica*, 89(5):2143–2188, 2021.

Brent R Moulton. An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *The review of Economics and Statistics*, pages 334–338, 1990.

Art B Owen. The pigeonhole bootstrap. *The Annals of Applied Statistics*, pages 386–411, 2007.

Yuya Sasaki and Yulong Wang. Non-robustness of the cluster-robust inference: with a proposal of a new robust method. *arXiv preprint arXiv:2210.16991*, 2022.

Andrew J Scott and D Holt. The effect of two-stage sampling on ordinary least squares methods. *Journal of the American statistical Association*, 77(380):848–854, 1982.

Samuel B Thompson. Simple formulas for standard errors that cluster by both firm and time. *Journal of financial Economics*, 99(1):1–10, 2011.

Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: journal of the Econometric Society*, pages 817–838, 1980.

Luther Yap. General conditions for valid inference in multi-way clustering. *arXiv preprint arXiv:2301.03805*, 2023.