Econometrics **November 20, 2023**

Topic 5: Two-Way Cluster-Robust (TWCR) 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

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

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 $s^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ 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 $\boldsymbol{\Omega}_g = \mathbb{V}\left[\mathbf{u}_g \mid \mathbf{X}_g\right] = \mathbb{E}\left[\mathbf{u}_g\mathbf{u}_g' \mid \mathbf{X}_g\right]$, and there is a consistent estimate $\hat{\boldsymbol{\Omega}}_g$ of $\boldsymbol{\Omega}_g$, we can estimate $\mathbb{E}\left[\mathbf{X}_g'\mathbf{u}_g\mathbf{u}_g'\mathbf{X}_g\right] = \mathbb{E}\left[\mathbf{X}_g'\boldsymbol{\Omega}_g\mathbf{X}_g\right]$ 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

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

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 $i \neq j$

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

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,

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

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

[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

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

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

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

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 $\mathrm{Var}(\alpha_i) > 0$ or $\gamma_t > 0$, the sample distribution

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

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

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

  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

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

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

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

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

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

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

then the bootstrap distribution is

– in the non-degenerate case,

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

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

– in the degenerate case,

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

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 $\mathbb{E}[\epsilon_{it}] = 0$, $\mathrm{Var}(\alpha_i) = \sigma^2_\alpha$, $\mathrm{Var}(\gamma_t) = \sigma^2_\gamma$, $\mathrm{Var}(\epsilon)_{it} = \sigma^2_\epsilon$. **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 2^{nd} 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 $Y_{i_1,\cdots,i_D} := h\left(X_{i_1}, \cdots, X_{i_D}\right)$ 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

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

with

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

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, $\mathbb{E}\left[e_{it} \mid a_i, g_t, v_{it}\right] = 0$, $\mathbb{E}[v_{it} \mid a_i] = \mathbb{E}\left[v_{it} \mid g_t\right] = 0$
- e_{it}, (a_i, g_t) and v_{it} are **uncorrelated**

then, rewrite the sample mean as

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

and the unconditional variances of the projections with

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

let $w_{it} \coloneqq v_{it} + e_{it}$, 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

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

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,

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

under Assumption 5.4.2, we have compact integral operators

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

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: $\sigma_{ak} := \mathrm{Cov}\left(a_i, \phi_k(\alpha_i)\right)$
  - $\left(\psi_1(\gamma_t), \cdots, \psi_K(\gamma_t)\right)$ can be correlated with g_t: $\sigma_{gk} := \mathrm{Cov}\left(g_t, \psi_k(\gamma_t)\right)$

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 $v(\alpha, \gamma) := \mathbb{E}\left[Y_{it} \mid \alpha_i = \alpha, \gamma_t = \gamma\right] - \mathbb{E}\left[Y_{it} \mid \alpha_i = \alpha\right] - \mathbb{E}\left[Y_{it} \mid \gamma_t = \gamma\right] + \mathbb{E}\left[Y_{it}\right]$ 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 $\bar{c}_k$: $c_k \leq \bar{c}_k, \forall k = 1, 2, \cdots$, 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

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

the sample analogs are:

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

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:

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

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 $\mathbf{c} \coloneqq \left(\bar{(c)}_k\right)_{k\geq 0}$ 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

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

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

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

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

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

where

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

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]

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

A Theoretical

Chiang and Sasaki (2023)

[9] Following Eq. (5.15),

References

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

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.

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.

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.

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.