# ECON 480-3
# LECTURE 16: CLUSTER COVARIANCE ESTIMATION

Ivan A. Canay
Northwestern University

Note: this class has gone through massive updates relative to the material in the lecture notes. Please use the information in these slides and ignore the material in the lecture notes (seriously, the notation is not even the same). I'll update notes later.

# Linear Model Setup

- Let $(Y, X, U)$ be a random vector where $Y$ and $U$ take values in $\mathbf{R}$ and $X$ takes values in $\mathbf{R}^{k+1}$.
- Let $\beta \in \mathbf{R}^{k+1}$ be such that
$$Y = X'\beta + U\,. \tag{1}$$
- **Today**: we observe a sample of size $n$ of $(Y, X)$ in a context where they are grouped into $q$ mutually independent known clusters, indexed by $j = 1, \ldots, q$.
- **Clustering**: can be due to sampling scheme or by the researcher knowing the correlation structure.
- Cannot exploit ordering of observations within a cluster (there may not be one).
- **Two tools**: (a) develop appropriate LLNs and CLTs for dependent processes, and (b) study the rate of convergence of the LS estimator under different dependence structures.

# Notation

- Use **two indices** as we did when we covered panel data.
- **First index**: $j = 1, \ldots, q$ denotes the **clusters** (observations that may be related to each other).

  **Example**: a cluster could be a family, a school, an industry, or a city.
- **Second index**: $i = 1, \ldots, n_j$ denotes units within a cluster.

  **Example**: family members, students, firms, or individuals.
- **Notation**: let

$$X_j = (X_{1,j}, \ldots, X_{n_j,j})'$$

  be a $n_j \times (k+1)$ matrix of stacked observations for cluster $j$, and define $Y_j$ and $U_j$ analogously. Then,

$$Y_j = X_j \beta + U_j \ , \quad j = 1, \ldots, q \ , \quad \text{where} \quad E[X_j' U_j] = 0 \ .$$

- We assume that $(Y_j, X_j)$ are **independent** across clusters but remain **agnostic** about the dependence **within** clusters.
  **Goal**: derive standard errors that account for the fact that $(X_{i,j}, U_{i,j})$ and $(X_{i',j}, U_{i',j})$ may be arbitrarily correlated within a cluster.

# Law of Large Numbers

- To introduce LLNs and CLTs we focus on the **sample mean** of $X_{i,j}$:

$$\bar{X}_n = \frac{1}{n}\sum_{j=1}^{q} X_j' \mathbf{1}_j = \frac{1}{n}\sum_{j=1}^{q}\sum_{i=1}^{n_j} X_{i,j} \ ,$$

where $\mathbf{1}_j$ is an $n_j$-dimensional **vector of ones**. Hansen and Lee (2019) prove the following result.

**Theorem (WLLN for clustered means)**

*Suppose that as* $n \to \infty$

$$\max_{j \leqslant q} \frac{n_j}{n} \to 0 \qquad (\clubsuit)$$

*and that*

$$\lim_{M\to\infty} \sup_{i,j} E\left[ \|X_{i,j}\| I\{\|X_{i,j}\| > M\} \right] = 0 \ . \qquad (2)$$

*Then, as* $n \to \infty$

$$\|\bar{X}_n - E[\bar{X}_n]\| \xrightarrow{P} 0 \ .$$

- The condition in (2) states that $X_{i,j}$ is **uniformly integrable** (similar to other standard conditions).

# Comments

- Assumption (♣) states that each cluster size $n_j$ is asymptotically **negligible**.
- Automatically holds when $n_j$ is fixed as $q \to \infty$ (original framework)
- Implies $q \to \infty$, so we do not explicitly list this.
- It allows for considerable **heterogeneity** in cluster sizes and it allows the cluster sizes to grow with sample size, so long as the growth is not proportional.
- **Example**: $n_j = n^a$ for $0 \leqslant a < 1$, which leads to

$$n = \sum_{j=1}^{q} n_j = \sum_{j=1}^{q} n^a = qn^a \quad \Rightarrow \quad q = n^{1-a} \, .$$

- Assumption (♣) is **necessary** for parameter estimation consistency while allowing **arbitrary** within-cluster dependence. Otherwise a single cluster could dominate the sample average.

# QUESTIONS?

# Rates of Convergence

- Under i.i.d. sampling the rate of convergence of the sample mean is $n^{-1/2}$
- Clustering may or may not affect this rate of convergence (often it does).
- **Hansen(2007)**: if dependence within clusters is **weak**, the rate of convergence is $n^{-1/2}$ but if the dependence is **strong**, the rate of convergence is determined by the number of clusters: $q^{-1/2}$.
- **Donut-hole problem**: What if the dependence is in between weak and strong?
- **Hansen and Lee (2019)**: the rate of convergence can be in between or even slower than these rates.
- **Analysis**: the convergence rate can be calculated as the standard deviation of the sample mean:

$$\text{Var}[\bar{X}_n] = \frac{1}{n^2} \sum_{j=1}^{q} \text{Var}[X_j' \mathbf{1}_j] \quad \Rightarrow \quad \text{sd}(\bar{X}_n) = \frac{1}{n} \left( \sum_{j=1}^{q} \text{Var}[X_j' \mathbf{1}_j] \right)^{1/2} .$$

# Rates of Convergence: Examples

$$\mathrm{sd}(\bar{X}_n) = \frac{1}{n}\Big(\sum_{j=1}^{q} \mathsf{Var}[X_j'\mathbf{1}_j]\Big)^{1/2}.$$

### Example (No dependence gives $\sqrt{n}$ rate)

Suppose $n_j = n^a$ (so $q = n^{1-a}$) and let $X_{i,j} \in \mathbf{R}$, $\mathsf{Var}[X_{i,j}] = 1$, and $\mathsf{Cov}[X_{i,j}, X_{i',j}] = 0$. Then:

$$\mathsf{Var}[X_j'\mathbf{1}_j] = \sum_{i=1}^{n_j} \mathsf{Var}[X_{i,j}] = n_j \quad \Rightarrow \quad \mathrm{sd}(\bar{X}_n) = \frac{1}{n}(qn^a)^{1/2} = n^{-1/2}.$$

### Example (High dependence gives $\sqrt{q}$ rate)

Suppose $n_j = n^a$ (so $q = n^{1-a}$) and let $X_{i,j} \in \mathbf{R}$, $\mathsf{Var}[X_{i,j}] = 1$, and $\mathsf{Cov}[X_{i,j}, X_{i',j}] = 1$. Then:

$$\mathsf{Var}[X_j'\mathbf{1}_j] = \sum_{i=1}^{n_j}\sum_{i'=1}^{n_j} \mathsf{Cov}[X_{i,j}, X_{i',j}] \propto n_j^2 = n^{2a} \quad \Rightarrow \quad \mathrm{sd}(\bar{X}_n) = n^{-(1-a)/2} = \frac{1}{n}(qn^{2a})^{1/2} = q^{-1/2}.$$

# Rates of Convergence: Examples

$$\text{sd}(\bar{X}_n) = \frac{1}{n}\Big(\sum_{j=1}^{q} \text{Var}[X_j'\mathbf{1}_j]\Big)^{1/2}.$$

## Example (In-between rate with strong dependence)

Suppose $n_j = n^a$ (so $q = n^{1-a}$) and let $X_{i,j} \in \mathbf{R}$, $\text{Var}[X_{i,j}] = 1$, and $\text{Cov}[X_{i,j}, X_{i',j}] = 1/|i - i'|$. Then:

$$\text{Var}[X_j'\mathbf{1}_j] = \sum_{i=1}^{n_j}\sum_{i'=1}^{n_j} \text{Cov}[X_{i,j}, X_{i',j}] = \sum_{i=1}^{n_j}\sum_{i'=1}^{n_j} \frac{1}{|i-i'|} =$$

# Rates of Convergence: Examples

$$\text{sd}(\bar{X}_n) = \frac{1}{n}\left(\sum_{j=1}^{q} \text{Var}[X_j' \mathbf{1}_j]\right)^{1/2}.$$

## Example (Slower than $\sqrt{n}$ and $\sqrt{q}$ rate with heterogeneity)

**Two types of clusters**: $q_1 = n/2$ with $n_j = 1$ (many small) and $q_2 = n^{1-a}/2$ with $n_j = n^a$ (few large).
**Note**: $q = q_1 + q_2 = O(n)$. Within each cluster the observations are identical and have unit variance.

# Comments

- The final example illustrates the importance of considering **heterogeneous** cluster sizes.

- The reason why the convergence rate is slower than both $n^{-1/2}$ and $q^{-1/2}$ is because the number of clusters is determined by the large number of small clusters ($q_1$), but the convergence rate is determined by the (relatively) small number of large clusters ($q_2 n^{2a}$).

- **Lessons from the Examples**: the convergence rate $\mathrm{sd}(\bar{X}_n)$ can be:
  1. Equal the square root of the sample size $\sqrt{n}$
  2. Equal to the square root of the number of clusters $\sqrt{q}$
  3. In-between these two
  4. Slower than both.

- When $\bar{X}_n$ is a vector: its elements may converge at **different rates**.

## Lesson

Under cluster dependence the convergence rate is context-dependent and variable-dependent, and it is therefore important to allow for general rates of convergence without imposing arbitrary rate ex-ante.

# QUESTIONS?

# Central Limit Theorem

- Under i.i.d. sampling the standard deviation of the sample mean is of order $O(n^{-1/2})$, so $\sqrt{n}$ seems to be the **natural scaling** to obtain the central limit theorem (CLT).

- **However**: clustering can alter the rate of convergence, so it is essential to **standardize** the sample mean by the actual variance rather than an assumed rate.

- The variance-covariance matrix of $\sqrt{n}\bar{X}_n$ is

$$\Omega_n = E\left[n(\bar{X}_n - E[\bar{X}_n])(\bar{X}_n - E[\bar{X}_n])'\right] = \frac{1}{n}\sum_{j=1}^{q} E\left[(X_j'\mathbf{1}_j - E[X_j'\mathbf{1}_j])(X_j'\mathbf{1}_j - E[X_j'\mathbf{1}_j])'\right].$$

- **Right scaling** is then $\Omega_n^{-1/2}\sqrt{n} \Rightarrow$ look at limit distribution of $\Omega_n^{-1/2}\sqrt{n}(\bar{X}_n - E[\bar{X}_n])$

- We denote by $\lambda_n = \lambda_{\min}(\Omega_n)$ the **minimum eigenvalue** of $\Omega_n$ in the statement of the next theorem. We also use $\mathbb{I}_{k+1}$ to denote the identity matrix of dimension $k+1$.

# Central Limit Theorem

**Theorem (Central Limit Theorem)**

*Suppose that for some* $2 \leqslant r < \infty$

$$\lim_{M\to\infty} \sup_{i,j} E\left[\|X_{i,j}\|^r I\{\|X_{i,j}\| > M\}\right] = 0 , \tag{3}$$

*and*

$$\frac{\left(\sum_{j=1}^{q} n_j^r\right)^{2/r}}{n} \leqslant C < \infty \tag{4}$$

*Assume further that as* $n \to \infty$

$$\max_{j\leqslant q} \frac{n_j^2}{n} \to 0 \tag{5}$$

*and*

$$\lambda_n \geqslant \lambda > 0 . \tag{6}$$

*Then, as* $n \to \infty$

$$\Omega_n^{-1/2}\sqrt{n}(\bar{X}_n - E[\bar{X}_n]) \xrightarrow{d} N\left(0, \mathbb{I}_{k+1}\right) . \tag{7}$$

# Comments

- **Assumption** (3) states that $\|X_{i,j}\|^r$ is **uniformly integrable**. When $r = 2$ this is similar to the Lindeberg condition for the CLT under independent heterogeneous sampling.

- **Assumption** (4) involves a **trade-off** between the cluster sizes and the number of moments $r$. It is least restrictive for large $r$, and more restrictive for small $r$.

  As $r \to \infty$ it approaches $\max_{j \leqslant q} \frac{n_j^2}{n} = O(1)$ which is implied by (5).

- **Assumption** (5) allows for growing and heterogeneous cluster sizes.
  - It allows clusters to grow uniformly at the rate $n_j = n^a$ for any $0 \leqslant a \leqslant (r-2)/2(r-1)$.
  - **Note**: this requires the cluster sizes to be bounded if $r = 2$.
  - It also allows for only a small number of clusters to grow.
  - **Example**: $n_j = \bar{n}$ (bounded) for $q - k$ clusters and $n_j = q^{a/2}$ for $k$ clusters, with $k$ fixed. In this case the assumption holds for any $a < 1$ and $r = 2$.

- **Assumption** (6) specifies that $\text{Var}[\sqrt{n} c' \bar{X}_n]$ does not vanish for any conformable vector $c \neq 0$.

# Cluster Covariance Estimation

- We now consider estimation of $\Omega_n$
- Under the assumption $E[X_j' \mathbf{1}_j] = 0$ for all $j = 1, \ldots, q$ (which will hold later) we obtain

$$\Omega_n = \frac{1}{n} \sum_{j=1}^{q} E\left[ (X_j' \mathbf{1}_j - E[X_j' \mathbf{1}_j])(X_j' \mathbf{1}_j - E[X_j' \mathbf{1}_j])' \right] = \frac{1}{n} \sum_{j=1}^{q} E\left[ X_j' \mathbf{1}_j \mathbf{1}_j' X_j \right].$$

- The **natural estimator** is:

$$\widehat{\Omega}_n = \frac{1}{n} \sum_{j=1}^{q} X_j' \mathbf{1}_j \mathbf{1}_j' X_j = \frac{1}{n} \sum_{j=1}^{q} \left( \sum_{i=1}^{n_j} X_{i,j} \right) \left( \sum_{i=1}^{n_j} X_{i,j} \right)'.$$

- This estimator is **robust** to dependence within clusters. It allows for **arbitrary within-cluster correlation** patterns and it also allows for $E\left[ X_j' \mathbf{1}_j \mathbf{1}_j' X_j \right]$ to vary across $j$ (**heterogeneity**).

# Consistency of the CCE

**Theorem (Consistency of CCE)**

*Under the same assumptions Theorem CLT and assuming* $E[X_j'\mathbf{1}_j] = 0$, *we obtain as* $n \to \infty$ *that*

$$||\widehat{\Omega}_n - \Omega_n|| \xrightarrow{P} 0$$

*and*

$$\widehat{\Omega}_n^{-1/2}\sqrt{n}\bar{X}_n \xrightarrow{d} N\left(0, \mathbb{I}_{k+1}\right).$$

- This shows that the CCE estimator is consistent and that replacing the covariance matrix in the CLT with the estimated covariance matrix does not affect the asymptotic distribution.
- **Implication**: Cluster-robust t-statistics are asymptotically standard normal.
- **Rate adaptive**: we do not need to know the actual rate of convergence of $\bar{X}_n$ as the CCE estimator captures this rate of convergence.
- For a proof of all theorems, see Hansen and Lee (2019, JoE).

# QUESTIONS?

# Applying Results to Linear Regression

$$Y_j = X_j\beta + U_j\,, \quad j = 1,\dots,q\,, \quad \text{where} \quad E[X_j' U_j] = 0\,.$$

- The LS estimator of $\beta$ is given by

$$\sqrt{n}(\hat{\beta}_n - \beta) = \left(\frac{1}{n}\sum_{j=1}^{q} X_j' X_j\right)^{-1} \frac{1}{\sqrt{n}}\sum_{j=1}^{q} X_j' U_j\,.$$

- **Notation**: Let

$$\Sigma_n = \frac{1}{n}\sum_{j=1}^{q} E[X_j' X_j] \quad \text{and} \quad \Omega_n = \frac{1}{n}\sum_{j=1}^{q} E[X_j' U_j U_j' X_j]\,.$$

## Consistency

If the rate conditions in Theorem WLLN hold, $\Sigma_n$ has full rank, $\lambda_{\min}(\Sigma_n) \geqslant C > 0$, and the uniform integrability condition in (2) holds for $X_{i,j}X_{i,j}'$ and $X_{i,j}'U_{i,j}$, then $\hat{\beta}_n$ is **consistent** for $\beta$.

# Applying Results to Linear Regression

## Asymptotic Normality

To properly **normalize** $\sqrt{n}(\hat{\beta}_n - \beta)$ we define

$$\mathbb{V}_n = \Sigma_n^{-1}\Omega_n\Sigma_n^{-1}$$

as the rate of convergence may not be $\sqrt{n}$.

Next, we assume the rate conditions in Theorem CLT hold for some $r$, $\Sigma_n$ has full rank, $\lambda_{\min}(\Sigma_n) \geqslant C > 0$, $\lambda_{\min}(\Omega_n) \geqslant C > 0$, and the uniform integrability condition in (3) holds for $X_{i,j}X_{i,j}'$ and $X_{i,j}'U_{i,j}$. It follows that as $n \to \infty$:

$$\mathbb{V}_n^{-1/2}\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \mathbb{I}_{k+1}) \ .$$

- All we now need is an **estimator** $\widehat{\mathbb{V}}_n$ with the property that

$$\widehat{\mathbb{V}}_n^{-1/2}\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \mathbb{I}_{k+1}) \ .$$

# The CCE

**Definition (Cluster Covariance Estimator: CCE)**

$$\widehat{\mathbb{V}}_n = \left(\frac{1}{n}\sum_{j=1}^{q} X_j' X_j\right)^{-1} \frac{1}{n}\sum_{j=1}^{q} X_j'\hat{U}_j\hat{U}_j' X_j \left(\frac{1}{n}\sum_{j=1}^{q} X_j' X_j\right)^{-1},$$

where $\hat{U}_j = Y_j - X_j\hat{\beta}_n$ are the LS residuals.

- Under the same conditions listed before,

$$(1)\ \|\widehat{\mathbb{V}}_n - \mathbb{V}_n\| \xrightarrow{P} 0 \quad \text{and} \quad (2)\ \widehat{\mathbb{V}}_n^{-1/2}\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \mathbb{I}_{k+1})\ .$$

- In the special case with $n_j = 1$ for all $j = 1, \ldots, q$, this estimator becomes the HC estimator.
- `Stata` uses a multiplicative adjustment to reduce the bias,

$$\widehat{\mathbb{V}}_{\text{stata}} = \frac{n-1}{n-k-1}\frac{q}{q-1}\hat{\mathbb{V}}_n\ .$$

- This estimator allows for arbitrary within-cluster correlation patterns and heteroskedasticity across clusters. Unlike HAC estimators, it does not require the selection of a kernel or bandwidth parameter.

# Inference

- For $s \in \{0, 1, \ldots, k\}$ let $\beta_s$ be the $s$th element of $\beta$ and let $\hat{V}_{n,s}$ be the $(s+1)$th diagonal element of $\widehat{\mathbb{V}}_n$.

- Consider testing
$$H_0 : \beta_s = c \text{ versus } H_1 : \beta_s \neq c$$
at level $\alpha$. Using the results we just derived, it follows that under the null hypothesis
$$t_{\text{stat}} = \frac{\sqrt{n}(\hat{\beta}_{n,s} - c)}{\sqrt{\hat{V}_{n,s}}} = \frac{\hat{\beta}_{n,s} - c}{\sqrt{\frac{1}{n}\hat{V}_{n,s}}} \xrightarrow{d} N(0,1) \text{ as } n \to \infty .$$
The test that rejects $H_0$ when $|t_{\text{stat}}| > z_{1-\frac{\alpha}{2}}$ is consistent in levels.

- **Important!**: this result holds **regardless** of the rate of convergence of the LS estimator!
  - $q \to \infty$ and $n_j$ fixed (traditional panel data asymptotics).
  - $q \to \infty$ and $n_j \to \infty$.
  - Either case with strong within-cluster dependence.
  - Either case with weak within-cluster dependence.
  - Heterogeneous clusters with different sizes and degrees of dependence.
  - Covariates with different degrees of dependence.

# QUESTIONS?

# Small $q$ ad-hoc adjustments

- Cluster-robust inference asymptotics are based on **many clusters**: $q \to \infty$.
- Often, however, there are only a **few clusters** (few regions, few schools, few states, etc).
- **Finite-sample adjustments** extend some of the ideas we discussed with HC standard errors.
- Bell and McCaffrey (2002): propose a bias-reduction modification **analogous to that of HC2**

$$\widehat{\mathbb{V}}_{\text{bm}} = \left( \frac{1}{n} \sum_{j=1}^{q} X_j' X_j \right)^{-1} \frac{1}{n} \sum_{j=1}^{q} X_j' \tilde{U}_j \tilde{U}_j' X_j \left( \frac{1}{n} \sum_{j=1}^{q} X_j' X_j \right)^{-1} ,$$

where

$$\tilde{U}_j = (\mathbb{I}_{n_j} - \mathbb{P}_{jj})^{-1/2} \hat{U}_j ,$$

$\mathbb{I}_{n_j}$ is the $n_j \times n_j$ identity matrix, $P_{jj}$ is the $n_j \times n_j$ matrix defined as

$$\mathbb{P}_{jj} = X_j (\mathbb{X}' \mathbb{X})^{-1} X_j' ,$$

and $\mathbb{X}$ is the $n \times (k+1)$ matrix constructed by stacking $X_1$ through $X_q$.

- The also propose a $t$ critical value with dof adjustment (same intuition as in Behrens-Fisher)

# Simulations

Table 1 reports simulations results for five designs in Imbens and Kolesar. The model is

$$Y_i = \beta_0 + \beta_1 X_i + U_i, \text{ with } X_i = V_{C_i} + W_i \text{ and } U_i = \nu_{C_i} + \eta_i \,.$$

$C_i$ denotes the cluster of $i$, all variables are $N(0,1)$, and $q = 10$. Report CS coverage.

| | dof | I | II | III | IV | V |
|---|---|---|---|---|---|---|
| $\hat{V}_n$ | $\infty$ | 84.7 | 73.9 | 79.6 | 85.7 | 81.7 |
| | $q-1$ | 89.5 | 86.9 | 85.2 | 90.2 | 86.4 |
| $\hat{V}_{\text{stata}}$ | $\infty$ | 86.7 | 78.8 | 81.9 | 87.6 | 83.6 |
| | $q-1$ | 91.1 | 90.3 | 87.2 | 91.8 | 88.1 |
| $\hat{V}_{\text{bm}}$ | $\infty$ | 89.2 | 84.7 | 87.2 | 89.1 | 87.7 |
| | $q-1$ | 93.0 | 93.3 | 91.3 | 92.8 | 91.4 |
| | $\kappa_{\text{bm}}$ | 94.4 | 95.3 | 94.4 | 94.2 | 96.6 |

TABLE: Design in Imbens and Kolesar/CGM: $1-\alpha = 95\%$.

# Final Comments

- When $q$ is **small**, $\widehat{\mathbb{V}}_{\text{bm}}$ (more so $\widehat{\mathbb{V}}_n$) typically leads to CSs that under-cover.

- BM (2002): more traction from the dof adjustment to the $t$-distribution that is used to compute the critical value of the test.

- Adjustment performs well sometimes, but is ad-hoc (no formal results).

- The literature on inference with **few clusters** (i.e, $q$ fixed) has made significant progress recently and the main alternatives to using CCE are:
  - **The Wild Bootstrap**: see Cameron, Gelbach, and Miller (2008) and Canay, Santos, and Shaikh (2021).
  - **Exact t-approach**: see Ibragimov and Mueller (2010)
  - **Approximate Randomization Tests**: see Canay, Romano, and Shaikh (2017).

- CCE is still a very good option when $q$ is large, but remember that its performance does not improve if $n_j$ gets large while $q$ remains small.

# THE END!