ECON 480-3
LECTURE 15: HAC COVARIANCE ESTIMATION

Ivan A. Canay
Northwestern University

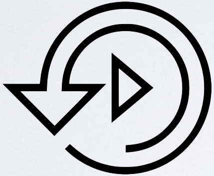
Past & Future

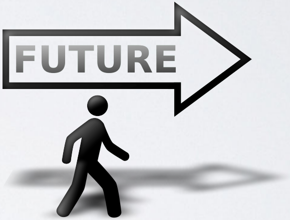
LAST CLASS

- HC Standard Errors
- Finite Sample Adjustments
- The Behrens-Fisher Problem

TODAY

- Stationarity
- Summability and mixing
- Naive Approaches
- Weighting and Truncation





Setup

- Let (Y, X, U) be st Y and U take values in $\mathbf{R}$ and X takes values in $\mathbf{R}^{k+1}$.
- The first component of X is a constant equal to one.
- Let $\beta \in \mathbf{R}^{k+1}$ be such that
$$Y = X'\beta + U \; .$$
Suppose that (1) $E[XU] = 0$, (2) that there is no perfect collinearity in X, that (3) $E[XX'] < \infty$, and that (4) $\text{Var}[XU] < \infty$.
- **Today**: we consider the case where the sample $(Y_1, X_1), \ldots, (Y_n, X_n)$ is not necessarily i.i.d. due to the presence of **dependence across observations**.
- **Autocorrelation**: the case where X_i and $X_{i'}$ may not be independent for $i \neq i'$.
- **Two tools**: (a) appropriate LLNs and CLTs for dependent processes, and (b) and description of the object we intend to estimate. For simplicity: assume $X_i = X_{1,i}$ is a scalar random variable and let the observations be naturally ordered (time series).

Limit theorems for dependent data

- Let's think about law of large numbers and central limit theorems to dependent data.
- **LLN**: when $\{X_i : 1 \leqslant i \leqslant n\}$ is i.i.d. with **mean** μ and **variance** σ_X^2, it follows that

$$\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}[X_i] = \frac{\sigma_X^2}{n} \to 0\,,$$

  and so convergence in probability follows by a simple application of Chebyshev's inequality.
- Without the independence, we need additional assumptions to control the **variance of the average**. We will start by assuming that the process we are dealing with are "stationary" as follows,

Definition

A process $\{X_i : 1 \leqslant i \leqslant n\}$ is **strictly stationarity** if for each j, the dist. of $\{X_i, \ldots, X_{i+j}\}$ is the same $\forall i$.

Definition

A process $\{X_i : 1 \leqslant i \leqslant n\}$ is **weakly stationary** if $E[X_i]$, $E[X_i^2]$, and, for each j, $\gamma_j \equiv \mathrm{Cov}[X_i, X_{i+j}]$, do not depend on i.

Stationarity: well defined mean

- **Stationarity**: the unique mean μ is well defined

- The variance of the sample average is

$$\text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{n}\text{Cov}[X_i, X_k] = \frac{1}{n^2}\left(n\gamma_0 + 2(n-1)\gamma_1 + 2(n-2)\gamma_2 + \cdots\right)$$

$$= \frac{1}{n}\left(\gamma_0 + 2\sum_{j=1}^{n}\gamma_j\left(1-\frac{j}{n}\right)\right) ,$$

where we have used the notation $\gamma_j = \text{Cov}[X_i, X_{i+j}]$, so that $\gamma_0 = \sigma_X^2$.

- For this variance to vanish, the last summation must not explode.

- A sufficient condition for this is **absolute summability**:

$$\sum_{j=-\infty}^{\infty} |\gamma_j| < \infty .$$

A law of large numbers follows one more time from an application of Chebyshev's inequality

Law of Large Numbers

Lemma

If $\{X_i : 1 \leqslant i \leqslant n\}$ is a (1) weakly stationary time series (with mean μ) with (2) absolutely summable auto-covariances, then a law of large numbers holds (in probability and L2).

Stationarity is not enough!: if $\zeta \sim N(0,1)$ and $X_i = \zeta\ \forall i$, then $\text{Cov}[X_i, X_{i'}] = 1\ \forall i, i'$.

Mixing

Absolutely summability follows from mixing assumptions, i.e., assuming the sequence $\{X_i : 1 \leqslant i \leqslant n\}$ is α-mixing. Let α_n be a number such that

$$|P(A \cap B) - P(A)P(B)| \leqslant \alpha_n \ ,$$

for any $A \in \sigma(X_1, \ldots, X_j)$, $B \in \sigma(X_{j+n}, X_{j+n+1}, \ldots)$, where $\sigma(X)$ is the σ-field generated by X, and $j \geqslant 1$, $n \geqslant 1$.

If $\alpha_n \to 0$ as $n \to \infty$, the sequence is then said to be α-mixing, the idea being that X_j and X_{j+n} are then approximately independent for large n.

Limiting Variance

- From the new proof of LLN one can guess that the variance in a central limit theorem **should change**.
- Remember that we wish to **normalize the sum** in such a way that the limit variance would be 1.
- To this end, note that

$$\begin{aligned} \text{Var}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i\right] &= \gamma_0 + 2\sum_{j=1}^{n}\gamma_j\left(1-\frac{j}{n}\right) \\ &\to \gamma_0 + 2\sum_{j=1}^{\infty}\gamma_j = \Omega\,, \end{aligned}$$

  where Ω is called the **long-run variance**.
- There are many central limit theorems for serially correlated observations. Below we provide a commonly used version, see Billingsley (1995, Theorem 27.4).

THEOREM

Suppose that $\{X_i : 1 \leqslant i \leqslant n\}$ *is a* (1) *strictly stationary* (2) α_n*-mixing stochastic process with* (3) $E[|X|^{2+\delta}] < \infty$, $E[X] = 0$, *and*

$$\text{(4)} \sum_{n=1}^{\infty} \alpha_n^{\delta/(2+\delta)} < \infty \, .$$

Then Ω *in the previous slide is finite (i.e. summabilidy holds) and, provided* $\Omega > 0$,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \xrightarrow{d} N(0, \Omega) \, .$$

QUESTIONS?

Estimating long-run variances

- **Linear Model**: with i.i.d. data, one of the exclusion restrictions is $E[U_i|X_i] = 0$.

- When the data is **potentially dependent** (time series, panel data, clustered data), we have to describe the conditional mean relative to all variables that may be important.

- We say X_i is **weakly exogenous** if

$$E(U_i|X_i, X_{i-1}, \dots) = 0$$

where we assume the observations have a natural ordering (e.g., time series).

- **LS estimator** of β,

$$\sqrt{n}(\hat{\beta}_n - \beta) = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i U_i \, .$$

Under appropriate assumption on $\{X_i : 1 \leqslant i \leqslant n\}$ and $\{\eta_i \equiv X_i U_i : 1 \leqslant i \leqslant n\}$ we get

$$\frac{1}{n}\sum_{i=1}^{n} X_i X_i' \xrightarrow{P} \Sigma_X \equiv E[XX'] \quad \text{and} \quad \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \eta_i \xrightarrow{d} N(0, \Omega) \, .$$

Long-Run Variance of LS

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N\left(0, \Sigma_X^{-1}\Omega\Sigma_X^{-1}\right).$$

- The only thing that is different from the usual sandwich formula is the **meat**
- In this case $\Omega = \sum_{j=-\infty}^{\infty} \gamma_j$ where γ_j are now the autocovariances of η_i.
- This long-run variance is significantly harder to estimate than the usual variance-covariance matrices that arise under i.i.d. assumptions.
- **Today**: figure out how to estimate Ω by the so-called **HAC approach**
- **Simplification**: ignore the fact that in practice U_i will be replaced by a regression residual $\hat{U}_i$ (since such modification is easy to incorporate and follows similar steps to those in previous lectures).

Naive Approach

- Ω is the sum of **all** auto-covariances (an infinite number of them). However, we can only estimate $n-1$ of them with a sample of size n.
- **Idea 1**: What if we just use the ones we can estimate? This leads to:

$$\tilde{\Omega} \equiv \sum_{j=-(n-1)}^{n-1} \hat{\gamma}_j\,, \quad \hat{\gamma}_j = \frac{1}{n}\sum_{i=1}^{n-j} \eta_i \eta_{i+j}\,.$$

Simple Truncation

- **Idea 2**: what if we do not use all the covariances?

- This gives us a **truncated estimator**,

$$\bar{\Omega} \equiv \sum_{j=-m_n}^{m_n} \hat{\gamma}_j = \hat{\gamma}_0 + 2\sum_{j=1}^{m_n} \hat{\gamma}_j \,.$$

  where $m_n < n$, $m_n \to \infty$, and $m_n/n \to 0$ as $n \to \infty$.

- **Finite sample bias**: truncation introduces finite sample bias. As m_n increases, the bias due to truncation should be smaller and smaller. But we don't want to increase m_n too fast for the reason stated above (we don't want to sum up noises).

- **Negative Estimator**: in small samples this estimator may be negative, $\bar{\Omega} < 0$ (or in vector case, $\bar{\Omega}$ not positive definite).

  Example: take $m_n = 1$, so that $\bar{\Omega} = \hat{\gamma}_0 + 2\hat{\gamma}_1$. In small samples, we may find $\hat{\gamma}_1 < -\frac{1}{2}\hat{\gamma}_0$, then $\bar{\Omega}$ will be negative.

Weighting and truncation: the HAC estimator

- **Newey and West (1987)**: create a **weighted sum** of sample auto-covariances with weights guaranteeing positive-definiteness:

$$\hat{\Omega}_n \equiv \sum_{j=-(n-1)}^{n-1} k\left(\frac{j}{m_n}\right)\hat{\gamma}_j \,.$$

  We need conditions on m_n and $k(\cdot)$ to give us consistency and positive-definiteness.

- **First**: $m_n \to \infty$ as $n \to \infty$ but not too fast. Today we assume $m_n^3/n \to 0$, but the result can be proved under $m_n^2/n \to 0$.

- **Second**: $k(\cdot)$ needs to be such that it guarantees positive-definiteness by down-weighting high lag covariances, but we also need $k(j/m_n) \to 1$ as $n \to \infty$ for consistency.

- As with non-parametric density estimation, there exist a variety of kernels that satisfy all the properties needed for consistency and positive-definiteness.

Popular Kernels

Barlett Kernel (Newey and West, 1987)

$$k(x) = \begin{cases} 1 - |x| & \text{if } |x| \leqslant 1 \\ 0 & \text{otherwise} \end{cases} .$$

Parzen kernel (Gallant, 1987)

$$k(x) = \begin{cases} 1 - 6x^2 + 6|x|^3 & \text{if } |x| \leqslant 1/2 \\ 2(1-|x|)^3 & \text{if } 1/2 \leqslant |x| \leqslant 1 \\ 0 & \text{otherwise} \end{cases} .$$

Quadratic spectral kernel (Andrews, 1991)

$$k(x) = \frac{25}{12\pi^2 x^2}\left(\frac{\sin(6\pi x/5)}{6\pi x/5} - \cos(\sin(6\pi x/5))\right) .$$

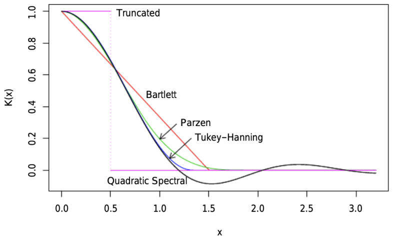


Figure 1: Kernel functions for kernel-based HAC estimation

- All symmetric at 0. The first two have bounded support $[-1, 1]$ and the QS has unbounded support.
- **First two**: the weight assigned to $\hat{\gamma}_j$ decreases with $|j|$ and becomes zero for $|j| \geqslant m_n$. Hence, m_n in these functions is also known as a truncation lag parameter.
- **QS**: the weight decreases to zero at $|j| = 1.2m_n$ but then exhibits damped sine waves afterwards.

Consistency of HAC estimator

- For the first two kernels, we can write

$$\hat{\Omega}_n \equiv \sum_{j=-m_n}^{m_n} k\left(\frac{j}{m_n}\right)\hat{\gamma}_j \, .$$

  Truncation at m_n is explicit. In the results we focus on this representation to simplify the arguments.

Theorem

Assume that $\{\eta_i : 1 \leqslant i \leqslant n\}$ *is a weakly stationary sequence with mean zero and autocovariances* $\gamma_j = \text{Cov}[\eta_i, \eta_{i+j}]$ *that satisfy absolute summability. Assume that*

1. $m_n \to \infty$ *as* $n \to \infty$ *and* $m_n^3/n \to 0$.
2. $k(x) : \mathbf{R} \to [-1, 1]$, $k(0) = 1$, $k(x)$ *is continuous at 0, and* $k(-x) = k(x)$.
3. *For all* j *the sequence* $\xi_{i,j} = \eta_i \eta_{i+j} - \gamma_j$ *is stationary and*

$$\sup_j \sum_{k=1}^{\infty} |\text{Cov}(\xi_{i,j}, \xi_{i+k,j})| < C$$

   for some constant C *(limited dependence).*

Then, $\hat{\Omega}_n \xrightarrow{P} \Omega$.

QUESTIONS?

Sketch of Proof

$$\hat{\Omega}_n \equiv \sum_{j=-m_n}^{m_n} k\left(\frac{j}{m_n}\right)\hat{\gamma}_j \quad \text{and} \quad \Omega \equiv \sum_{j=-\infty}^{\infty} \gamma_j \,.$$

Sketch of Proof

$$\hat{\Omega}_n - \Omega = -\sum_{|j|>m_n} \gamma_j + \sum_{j=-m_n}^{m_n} \left(k\left(\frac{j}{m_n}\right) - 1 \right) \gamma_j + \sum_{j=-m_n}^{m_n} k\left(\frac{j}{m_n}\right) (\hat{\gamma}_j - \gamma_j) \,.$$

Let $f_n(j) \equiv \left| k\left(\frac{j}{m_n}\right) - 1 \right| |\gamma_j|$ and note $f_n(j) \leqslant g(j) \equiv 2|\gamma_j|$.

Extra Slide

Sketch of Proof

$$\hat{\Omega}_n - \Omega = -\sum_{|j|>m_n} \gamma_j + \sum_{j=-m_n}^{m_n} \left(k\left(\frac{j}{m_n}\right) - 1\right)\gamma_j + \sum_{j=-m_n}^{m_n} k\left(\frac{j}{m_n}\right)(\hat{\gamma}_j - \gamma_j)\,.$$

Let $\gamma_j^* \equiv E[\hat{\gamma}_j] = \frac{n-j}{n}\gamma_j$ since $\hat{\gamma}_j = \frac{1}{n}\sum_{i=1}^{n-j} \eta_i \eta_{i+j}$

Sketch of Proof

$$\text{WTS}: \sum_{j=-m_n}^{m_n} |\hat{\gamma}_j - \gamma_j^*| \xrightarrow{P} 0 \quad \text{Recall} \quad \sup_j \sum_{k=1}^{\infty} |\text{Cov}(\xi_{i,j}, \xi_{i+k,j})| \leqslant C.$$

Step 1: Let $\xi_{i,j} \equiv \eta_i \eta_{i+j} - \gamma_j$ and show that $E[(\hat{\gamma}_j - \gamma_j^*)^2] \leqslant C/n$.

Sketch of Proof

$$\text{WTS}: \sum_{j=-m_n}^{m_n} |\hat{\gamma}_j - \gamma_j^*| \xrightarrow{P} 0 \quad \text{that is} \quad P\left\{ \sum_{j=-m_n}^{m_n} |\hat{\gamma}_j - \gamma_j^*| > \epsilon \right\} \to 0\,.$$

Note: The event $A = \{\sum_{j=-m_n}^{m_n} |\hat{\gamma}_j - \gamma_j^*| > \epsilon\}$ implies $B = \{|\hat{\gamma}_j - \gamma_j^*| > \frac{\epsilon}{2m_n+1}$ for at last some $j\}$

Concluding Remarks

- We proved **consistency** but did not proved positive definiteness of our HAC estimator.
- **Required**: to characterize positive definiteness using the Fourier transformation of $\hat{\Omega}$.
- **Bandwidth choice.** After the original paper by Newey-West (1987), a series of papers addressed the issue of bandwidth choice (notably, Andrews (1991)).
- **General idea**: bias-variance trade-off in the choice of bandwidth m_n. A bigger m_n reduces the cut-off bias, however, it increases the number of estimated covariances used (and hence the variance of the estimate).
- Andrews (1991): choose m_n by minimizing the mean squared error (MSE) of the HAC estimator,

$$MSE(\hat{\Omega}_n) = \text{bias}(\hat{\Omega}_n)^2 + \text{Var}(\hat{\Omega}_n) \ .$$

- He showed that the optimal bandwidth is $m_n = C^* n^{1/r}$, where $r = 3$ for the Barlett kernel and $r = 5$ for other kernels. He also derived the optimal constant C^*, which depends on the kernel used among other things.
- In finite samples, inference on $\hat{\beta}_n$ based on HAC standard errors may perform **quite poorly**.

THE END!