AEA Continuing Education Course

Time Series Econometrics

Lecture 4

Heteroskedasticity- and Autocorrelation-Robust Inference

or

Three Decades of HAC and HAR: What Have We Learned?

James H. Stock

Harvard University

January 6, 2015

Outline

HAC = Heteroskedasticity- and Autocorrelation-Consistent
HAR = Heteroskedasticity- and Autocorrelation-Robust

1) HAC/HAR Inference: Overview
2) Notational Preliminaries: Three Representations, Three Estimators
3) The PSD Problem and Equivalence of Sum-of-Covariance and Spectral Density Estimators
4) Three Approaches to the Bandwidth Problem
5) Application to Flat Kernel in the Frequency Domain
6) Monte Carlo Comparisons
7) Panel Data and Clustered Standard Errors
8) Summary

1) HAC/HAR Inference: Overview

The task: valid inference on β when X_t and u_t are possibly serially correlated:

$$Y_t = X_t'\beta + u_t,\ E(u_t|X_t) = 0,\ t = 1,\ldots, T$$

Asymptotic distribution of OLS estimator:

$$\sqrt{T}(\hat{\beta}-\beta) = \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\right)^{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_t u_t\right)$$

Assume throughout that WLLN and CLT hold:

$$\frac{1}{T}\sum_{t=1}^{T} X_t X_t' \xrightarrow{p} \Sigma_{XX} \text{ and } \frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_t u_t \xrightarrow{d} \mathrm{N}(0, \Omega),$$

so $$\sqrt{T}(\hat{\beta}-\beta) \xrightarrow{d} N\left(0, \Sigma_{XX}^{-1}\Omega\Sigma_{XX}^{-1}\right).$$

Σ_{XX} is easy to estimate, but what is Ω and how should it be estimated?

Ω: The Long-Run Variance of $X_t u_t$

Let $Z_t = X_t u_t$. Note that $EZ_t = 0$ (because $E(u_t|X_t) = 0$). Suppose Z_t is second order stationary. Then

$$\Omega_T = \operatorname{var}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t\right) = E\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t\right)^2$$

$$= \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T} E\left(Z_t Z_s{}'\right)$$

$$= \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T} \Gamma_{t-s} \quad (Z_t \text{ is second order stationary})$$

$$= \frac{1}{T}\sum_{j=-(T-1)}^{T-1} \left(T - |j|\right)\Gamma_{t-s} \quad \text{(adding along the diagonals)}$$

$$= \sum_{j=-(T-1)}^{T-1} \left(1 - \left|\frac{j}{T}\right|\right)\Gamma_j \to \sum_{j=-\infty}^{\infty} \Gamma_j$$

so

$$\Omega = \sum_{j=-\infty}^{\infty} \Gamma_j = 2\pi S_Z(0) \quad \text{(recall that } S_Z(\omega) = \frac{1}{2\pi}\sum_{j=-\infty}^{\infty} \Gamma_j e^{-i\omega j}\text{)}$$

Standard approach: Newey-West Standard Errors

- HAC/HAR SEs are generically needed in time series regression. The most common method (by far) for computing HAC/HAR SEs is to use the Newey-West (1987) estimator.
- Newey-West estimator: declining average of sample autocovariances

$$\hat{\Omega}^{NW} = \sum_{j=-m}^{m}\left(1-\left|\frac{j}{m}\right|\right)\hat{\Gamma}_j$$

where $\hat{\Gamma}_j = \frac{1}{T}\sum_{t=1}^{T}\hat{Z}_t\hat{Z}_{t-j}{}'$, where $\hat{Z}_t = X_t\hat{u}_t$.

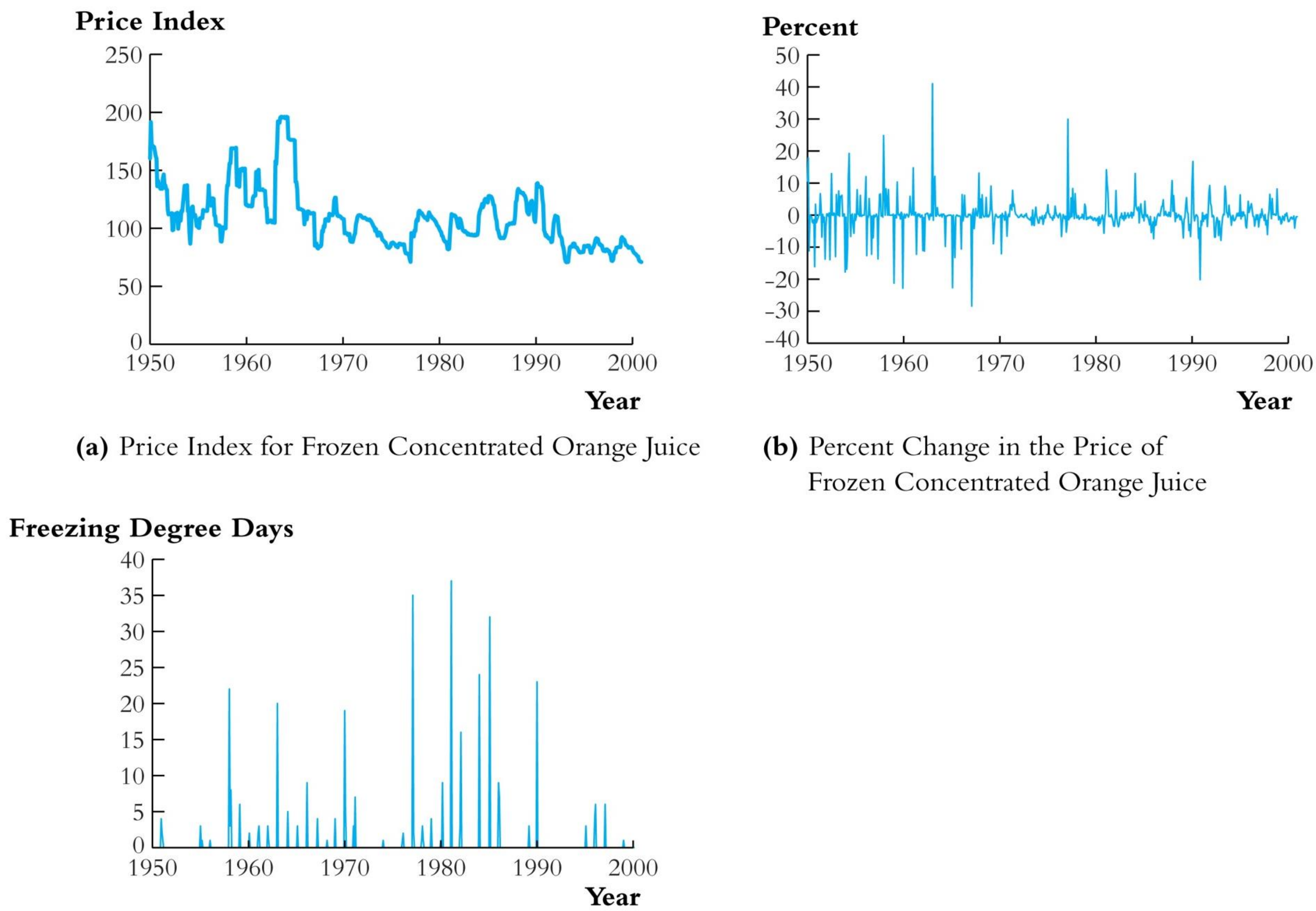
- Rule-of-thumb for m: $m = m_T = .75T^{1/3}$ (e.g. Stock and Watson, *Introduction to Econometrics*, 3rd edition, equation (15.17).
  - This rule-of-thumb dates to the 1990s. More recent research suggests it needs updating – and that, perhaps, the NW weights need to be replaced.

Four examples…


Florida's Orange Groves

Source: "USDA Assesses Freeze Damage of Florida Oranges," Feb. 1, 2011 at
http://blogs.usda.gov/2011/02/01/usda-assesses-freeze-damage-of-florida-oranges/

FIGURE 15.1 Orange Juice Prices and Florida Weather, 1950–2000



(a) Price Index for Frozen Concentrated Orange Juice

(b) Percent Change in the Price of Frozen Concentrated Orange Juice

(c) Monthly Freezing Degree Days in Orlando, Florida

Example 1: OJ prices and Freezing degree-days:

$$\Delta \ln P_t = \alpha + \beta(\mathrm{L})FDD_t + u_t$$

Example 2: GDP growth and monetary policy shock:

$$\Delta \ln GDP_t = \alpha + \beta(\mathrm{L})\varepsilon_t^m + u_t$$

Example 3: Multiperiod asset returns:

$$\Delta \ln(P_{t+k}/P_t) = \alpha + \beta X_t + u_t^{t+l}, \text{ e.g. } X_t = \text{dividend yield}_t$$

Example 4: (GMM) Hybrid New Keynesian Phillips Curve:

$$\pi_t = \lambda x_t + \gamma_f E_t \pi_{t+1} + \gamma_b \pi_{t-1} + \eta_t$$

where x_t = marginal cost/output gap/unemployment gap and π_t = inflation. Suppose $\gamma_b + \gamma_f = 1$ (empirically supported); then

$$\Delta\pi_t = \lambda x_t + \gamma_f(E_t\pi_{t+1} - \pi_{t-1}) + \eta_t$$

Instruments: $\{\pi_{t-1}, x_{t-1}, \pi_{t-2}, x_{t-2}, \ldots\}$

- η_t could be serially correlated by omission of supply shocks

Digression: Why not just use GLS?

The path to GLS: suppose u_t follows an AR(1)

$$Y_t = X_t'\beta + u_t,$$

$$u_t = \rho u_{t-1} + \varepsilon_t,\ \varepsilon_t \text{ serially uncorrelated}$$

This suggests Cochrane-Orcutt quasi-differencing:

$$(1\text{-}\rho\mathrm{L})Y_t = ((1\text{-}\rho\mathrm{L})X_t)' + \varepsilon_t \text{ or } \tilde{y}_t = \tilde{x}_t'\beta + \varepsilon_t$$

(Feasible GLS uses an estimate of ρ – not the issue here)

Validity of the quasi-differencing regression requires $E(\varepsilon_t|\tilde{x}_t) = 0$:

$$E(\varepsilon_t|\tilde{x}_t) = E(u_t - \rho u_{t-1}|x_t - \rho x_{t-1}) = 0$$

For general ρ, this requires all the cross-terms to be zero:

(i) $E(u_t|x_t) = E(u_{t-1}|x_{t-1}) = 0$

(ii) $E(u_t|x_{t-1}) = 0$

(iii) $E(u_{t-1}|x_t) = 0$ – this condition fails in examples 1-4

2) Notational Preliminaries: Three Representations, Three Estimators

The challenge: estimate $\Omega = \sum_{j=-\infty}^{\infty} \Gamma_j$

- This is hard: the sum has ∞'s!

- Draw on the literature on estimation of the spectral density to estimate Ω

- Three estimators of the spectral density:

(1) Sum-of-covariances: $\hat{\Omega}^{sc} = \sum_{j=-(T-1)}^{T-1} k_T(j)\hat{\Gamma}_j$

(2) Weighted periodogram: $\hat{\Omega}^{wp} = 2\pi \sum_{l=-(T-1)}^{T-1} K_T(l) I_{\hat{Z}\hat{Z}}(2\pi l / T)$

(3) VARHAC: $\hat{\Omega}^{VARHAC} = \hat{A}(1)^{-1}\hat{\Sigma}_{\hat{u}\hat{u}}\hat{A}(1)^{-1}$

We follow the literature and focus on (1) and (2)

(1) Sum-of-covariances estimator of Ω

$$\Omega = \sum_{j=-\infty}^{\infty} \Gamma_j$$

Because Z_t is stationary and Ω exists, Γ_j dies off. This suggests and estimator of Ω based a weighted average of the first few sample estimators of Γ:

$$\hat{\Omega}^{sc} = \sum_{j=-(T-1)}^{T-1} k_T(j)\hat{\Gamma}_j$$

where $\hat{\Gamma}_j = \frac{1}{T}\sum_{t=1}^{T} Z_t Z_{t-j}'$ (throughout, use the convention $Z_t = 0$, $t<1$ or $t > T$)

$k_T(.)$ is the weighting function or "kernel":

- Example: $k_T(j) = 1 - |j/m_T|$ = "triangular weight function" = "Bartlett kernel" = "Newey-West weights" with truncation parameter m_T
- We return to kernel and truncation parameter choice problem below

(2) Smoothed periodogram estimator of Ω

The periodogram as an inconsistent estimator of the spectral density:

- Fourier transform of Z_t at frequency ω: $d_Z(\omega) = \dfrac{1}{\sqrt{2\pi T}}\sum_{t=1}^{T} Z_t e^{-i\omega t}$
- The periodogram is $I_{ZZ}(\omega) = d_Z(\omega)\overline{d_Z(\omega)}'$

Asymptotically, $I_{ZZ}(\omega)$ is distributed as $S_Z(0)\times(\chi_2^2/2)$ (scalar case)

- Mean:

$$E\, I_{ZZ}(\omega) = E(d_Z(\omega)\overline{d_Z(\omega)}')$$

$$= \frac{1}{2\pi} E\left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t e^{i\omega t}\right|^2$$

$$= \frac{1}{2\pi}\sum_{j=-\infty}^{\infty} \Gamma_j e^{-i\omega j} = S_Z(\omega)$$

- Distribution (Brillinger (1981), Priestley (1981), Brockwell and Davis (1991)):

$$d_Z(\omega) = \frac{1}{\sqrt{2\pi T}}\sum_{t=1}^{T} Z_t e^{i\omega t}$$

$$= \frac{1}{\sqrt{2\pi}}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t \cos\omega t + i\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t \sin\omega t\right)$$

$= z_1 + iz_2$, say, where z_1 and z_2 are i.i.d. mean zero normal

So

$$I_{ZZ}(\omega) = d_Z(\omega)\overline{d_Z(\omega)}' = z_1^2 + z_2^2 \xrightarrow{d} S_Z(\omega)\times(\chi_2^2/2)$$

- For ω evaluated at $\omega_j = 2\pi j/T$, $j = 0, 1,\ldots, T$, $d_Z(\omega_j)$ and $d_Z(\omega_k)$ are asymptotically independent (orthogonality of sins and cosines).
- The weighted periodogram estimator averages the periodogram near zero:

$$\hat{\Omega}^{wp} = 2\pi \sum_{l=-(T-1)}^{T-1} K_T(l) I_{ZZ}(2\pi l / T)$$

(3) VAR-HAC estimator of Ω

Approximate the dynamics of Z_t by a vector autoregression: A(L)$Z_t = u_t$

so Z_t has the vector MA representation, Z_t = A(L)$^{-1}u_t$

Thus

$$S_Z(\omega) = \frac{1}{2\pi} A\left(e^{i\omega}\right)^{-1} \Sigma_{uu} \overline{A\left(e^{i\omega}\right)}^{-1\prime}$$

so

$$S_Z(0) = \frac{1}{2\pi} A\left(1\right)^{-1} \Sigma_{uu} A\left(1\right)^{-1\prime}$$

This suggests the VAR-HAC estimator (Priestley (1981), Berk (1974); den Haan and Levin (1997),

$$\hat{\Omega}^{VARHAC} = \hat{A}(1)^{-1} \hat{\Sigma}_{\hat{u}\hat{u}} \hat{A}(1)^{-1}$$

where $\hat{A}(1)$ and $\hat{\Sigma}_{\hat{u}\hat{u}}$ are obtained from a VAR estimated using $\hat{Z}_t$.

3) The PSD Problem and Equivalence of Sum-of-Covariance and Spectral Density Estimators

Not all estimators of Ω are positive semi-definite – including some natural ones. Consider the m-period return problem – so under the null $\beta = 0$, u_t is a MA(m-1). This suggests using a specific sum of covariances estimator:

$$\tilde{\Omega} = \sum_{j=-(m-1)}^{m-1} \hat{\Gamma}_j .$$

But $\tilde{\Omega}$ isn't psd with probability one! Consider $m = 2$ and the scalar case:

$$\tilde{\Omega} = \sum_{j=-1}^{1} \hat{\gamma}_j = \hat{\gamma}_0 \left(1 + 2\frac{\hat{\gamma}_1}{\hat{\gamma}_0} \right) < 0 \text{ if } \frac{\hat{\gamma}_1}{\hat{\gamma}_0} = \text{first sample autocorrelation} < -0.5$$

Solutions to the PSD problem

- Restrict kernel/weight function so that estimator is PSD with probability one (standard method)
- Hybrid, e.g. use $\tilde{\Omega}$ but switch to PSD method if $\tilde{\Omega}$ isn't psd – won't pursue (not used in empirical work)

Choice of kernel so that $\hat{\Omega}^{sc}$ is psd w.p.1

Step 1:

Note that $\hat{\Omega}^{wp}$ is psd w.p.1 if the frequency-domain weight function is non-negative. Recall that $\hat{\Omega}^{wp}$ is psd if $\lambda'\hat{\Omega}^{wp}\lambda \geq 0$ for all λ. Now

$$\begin{aligned}\lambda'\hat{\Omega}^{wp}\lambda &= 2\pi \sum_{l=-(T-1)}^{T-1} K_T(l)\left(\lambda' I_{ZZ}(2\pi l/T)\lambda\right) \\ &= 2\pi \sum_{l=-(T-1)}^{T-1} K_T(l)\left(\lambda' d_Z(\omega_l)\overline{d_Z(\omega_l)}'\lambda\right) \\ &= 2\pi \sum_{l=-(T-1)}^{T-1} K_T(l)\left|\lambda' d_Z(\omega_l)\right|^2 \geq 0\end{aligned}$$

with probability 1 if $K_T(l) \geq 0$ for all l.

- $K_T(l) \geq 0$, all l, is necessary and sufficient for $\hat{\Omega}^{wp}$ to be psd

Step 2: $\hat{\Omega}^{wp}$ and $\hat{\Omega}^{sc}$ are equivalent!

$$\hat{\Omega}^{wp} = 2\pi \sum_{l=-(T-1)}^{T-1} K_T(l) I_{ZZ}(2\pi l / T)$$

$$= 2\pi \sum_{l=-(T-1)}^{T-1} K_T(l) \left(\frac{1}{\sqrt{2\pi T}} \sum_{t=1}^{T} Z_t e^{i2\pi lt/T} \right) \left(\frac{1}{\sqrt{2\pi T}} \sum_{s=1}^{T} Z_s e^{-i2\pi ls/T} \right)$$

$$= \sum_{l=-(T-1)}^{T-1} K_T(l) \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} Z_t Z_s' e^{-i2\pi l(s-t)/T}$$

$$= \sum_{l=-(T-1)}^{T-1} K_T(l) \sum_{j=-(T-1)}^{T-1} \frac{1}{T} \sum_{t=1}^{T} Z_t Z_{t-j}' e^{-i2\pi lj/T}$$

$$= \sum_{j=-(T-1)}^{T-1} \frac{1}{T} \sum_{t=1}^{T} Z_t Z_{t-j}' \sum_{l=-(T-1)}^{T-1} K_T(l) e^{-i(2\pi j/T)l}$$

$$= \sum_{j=-(T-1)}^{T-1} \hat{\Gamma}_j k_T(j) = \hat{\Omega}^{sc}, \text{ where } k_T(j) = \sum_{l=-(T-1)}^{T-1} K_T(l) e^{-i(2\pi j/T)l}$$

Result: $\hat{\Omega}^{sc}$ is psd w.p.1 if and only if k_T is the (inverse) Fourier transform of a nonnegative frequency domain weight function K_T. Also, k_T is real if K_T is symmetric (then $k_T(j) = K_T(0) + 2\sum_{l=1}^{T-1} K_T(l)\cos\left[(2\pi j / T)l\right]$).

Kernel and bandwidth choice

The class of estimators here is very large. What is a recommendation for empirical work?

Two distinct questions:

(i) What kernel to use?

(ii) Given the kernel, what bandwidth to use?

It turns out that problem (ii) is more important in practice than problem (i).

Some final preliminaries

- Closer look at four kernels:
  - Newey-West (triangular in time domain)
  - Flat in time domain
  - Flat in frequency domain
  - Epinechnikov (Quadratic Spectral) – certain optimality properties
- Link between time domain and frequency domain kernels

Flat kernel in frequency domain

In general:

$$\hat{\Omega}^{wp} = 2\pi \sum_{l=-(T-1)}^{T-1} K_T(l) I_{ZZ}(2\pi l / T)$$

Flat kernel:

$$K_T(l) = \begin{cases} \dfrac{1}{2B_T+1} & \text{if } |l| \le B_T \\ 0 & \text{if } |l| > B_T \end{cases}$$

Then $\hat{\Omega}^{wp}$ becomes

$$\hat{\hat{\Omega}} = \frac{2\pi}{2B_T+1} \sum_{l=-B_T}^{B_T} I_{ZZ}\left(\frac{2\pi l}{T}\right)$$

The time-domain kernel corresponding to the flat frequency-domain kernel is

$$k_T(j) = \sum_{l=-(T-1)}^{T-1} K_T(l) e^{-i(2\pi j/T)l}$$

$$= \frac{1}{2B_T+1}\sum_{l=-B_T}^{B_T} e^{-i(2\pi j/T)l}$$

$$= \ldots \to_{T\to\infty} \frac{\sin(2\pi j/m_T)}{2\pi j/m_T}$$, where $m_T = T/B_T$

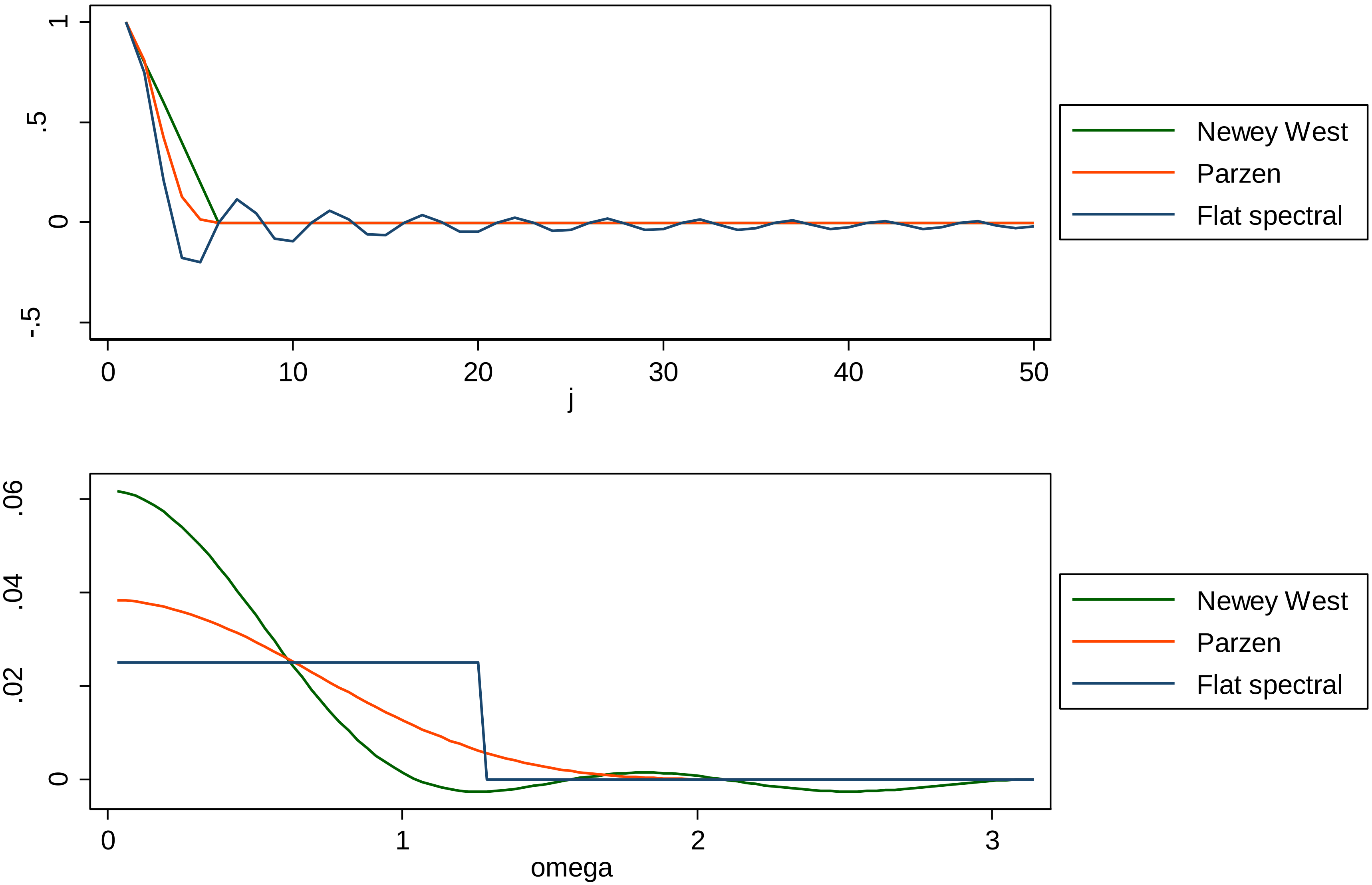
Important points:

- $m_T B_T = T$: using few periodogram ordinates corresponds to using many covariances
- Flat in frequency domain (which is psd) produces some negative weights in the sum-of-covariance kernel

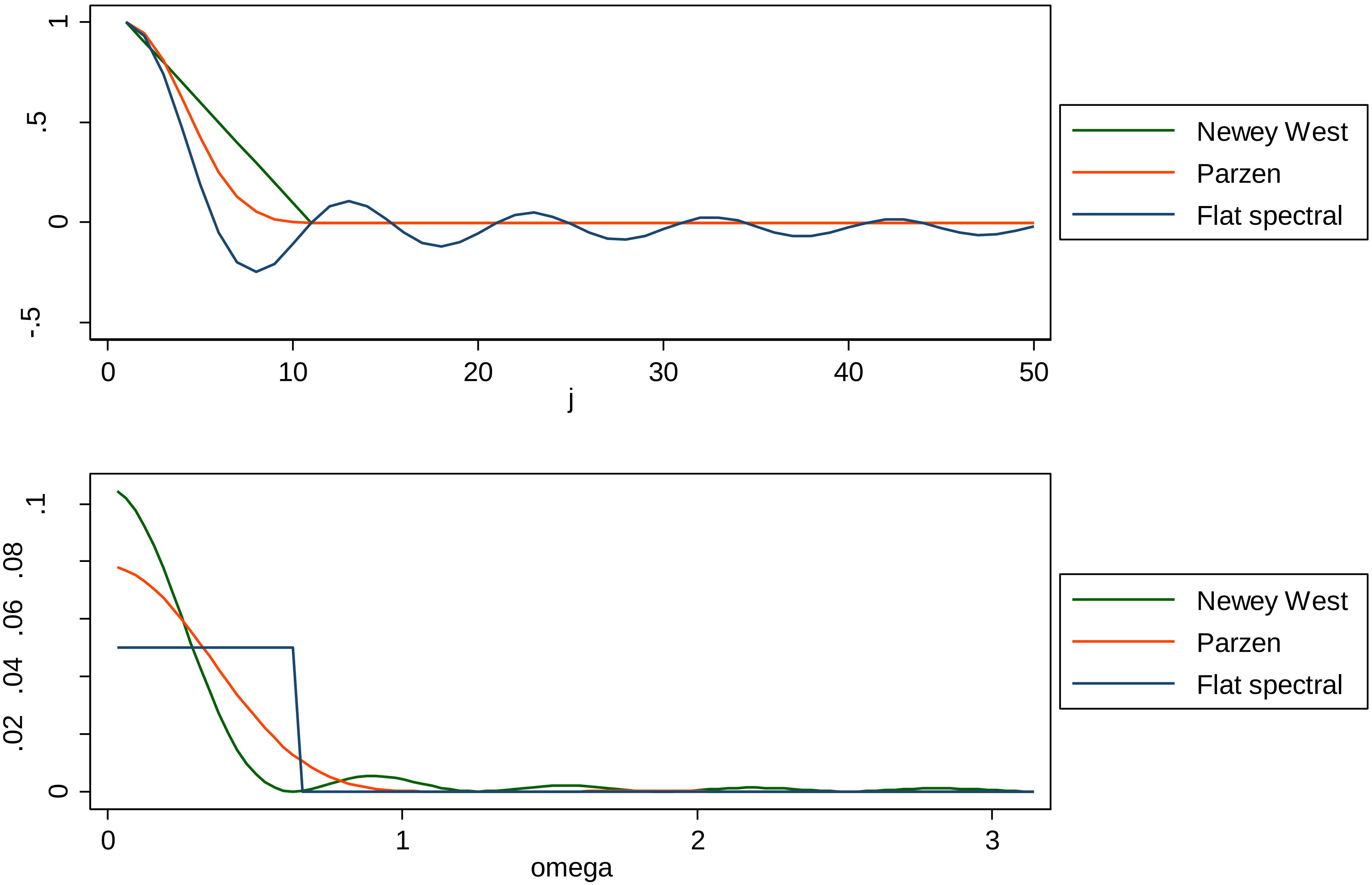
Three PSD kernels in pictures

Kernel	$k(x)$, $x = \lvert j \rvert / m$	$K(u)$, $u = \lvert l \rvert / B$
Newey-West	$1-\lvert x \rvert$ if $\lvert x \rvert \le 1$	
Parzen	$1 - 6x^2 + 6\lvert x \rvert^3$ if $\lvert x \rvert < .5$ $2(1-\lvert x \rvert)^3$ if $.5 \le \lvert x \rvert \le 1$	
Flat spectral		1 if $\lvert u \rvert \le 1$

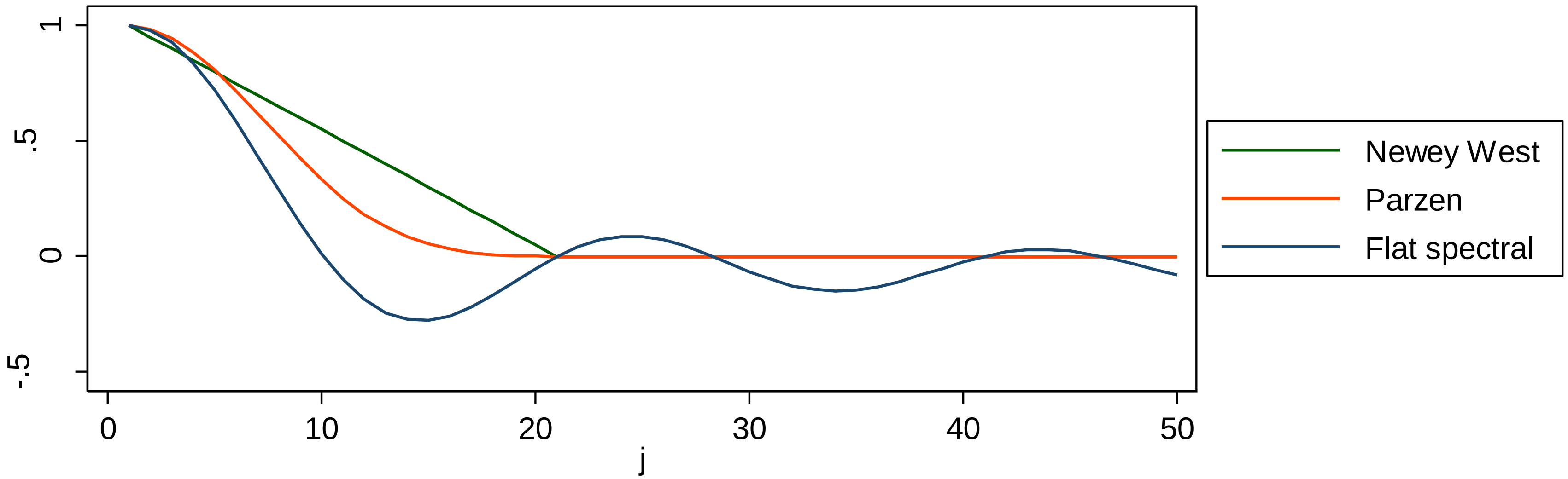
Three PSD Kernels: m = 5, B = 40, T = 200

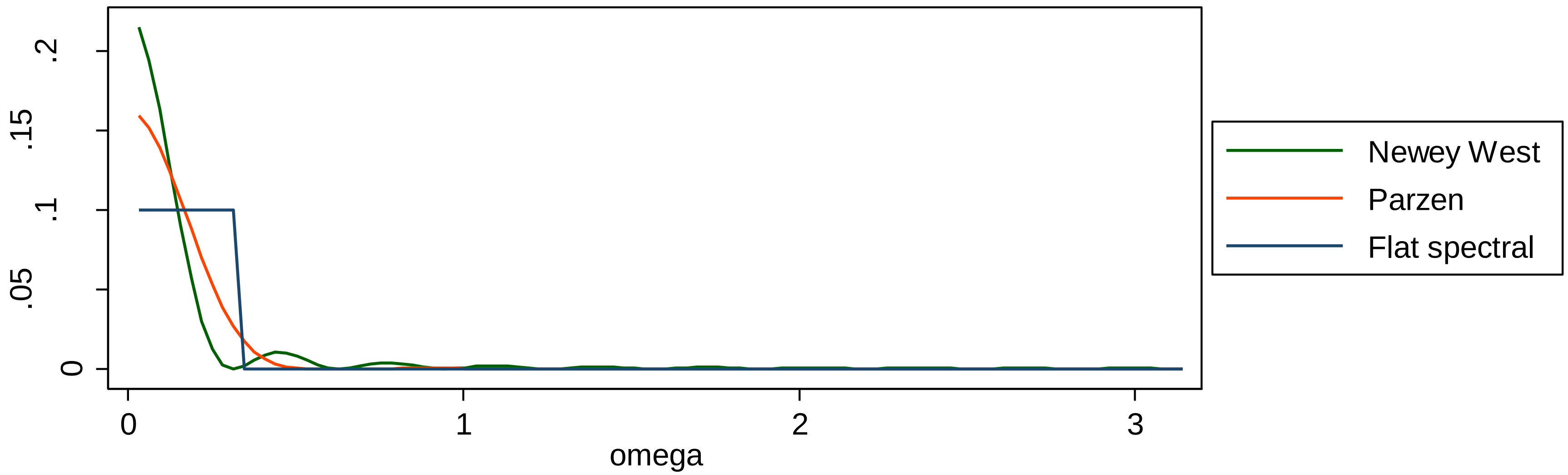


Three PSD Kernels: m = 10, B = 20, T = 200

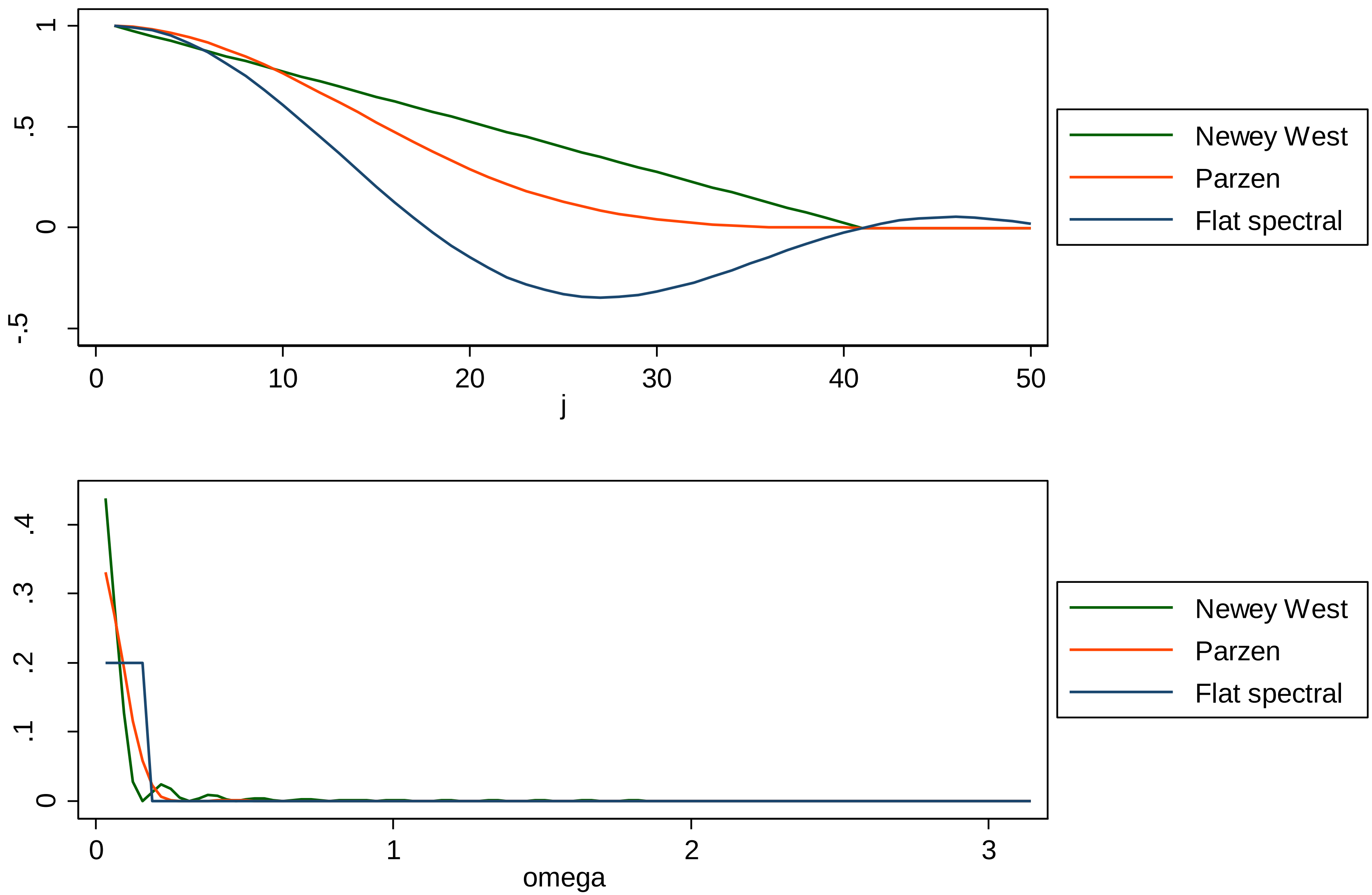


Three PSD Kernels: m = 20, B = 10, T = 200





Three PSD Kernels: m = 40, B = 5, T = 200



4) Three Approaches to the Bandwidth Problem

As in all nonparametric problems, there is a fundamental tradeoff between bias and variance when choosing smoothing parameters.

- In frequency domain:

$$\hat{\Omega}^{wp} = 2\pi \sum_{l=-B}^{B} K_T(l) I_{ZZ}(2\pi l / T)$$

Larger B decreases variance, but increases bias

- In time domain:

$$\hat{\Omega}^{sc} = \sum_{j=-m}^{m} k_T(j)\hat{\Gamma}_j$$

Larger m increases variance, but decreases bias

- Recall $m_T B_T = T$

How should this bias-variance tradeoff be resolved?

First generation answer:

Obtain as good an estimate of Ω as possible (Andrews [1991])

- "Good" means:
  - psd with probability 1
  - consistent (HAC)
  - minimize mean squared error:

$$\text{MSE}(\hat{\Omega}) = E(\hat{\Omega} - \Omega)^2 = \text{bias}(\hat{\Omega})^2 + \text{var}(\hat{\Omega})$$

  - This yields a bandwidth m_T that increases with, but more slowly than, T
- Practical issue:
  - if true spectral density is flat in neighborhood of zero, you should include many periodogram ordinates (large B); equivalently, if true Γ_j's are small for $j \neq 0$ then you should include few $\hat{\Gamma}_j$'s
  - But, you don't know the true spectral density!!
  - So, in practice you can estimate and plug in, or use a rule-of-thumb.
  - The $m = .75T^{1/3}$ rule of thumb assumes X_t and u_t are AR(1) with coefficient 0.5
- Then use asymptotic chi-squared critical values to evaluate test statistics.

Big problem with the first generation answer

- The resulting estimators do a very bad job of controlling size when the errors are in fact serially correlated, even with a modest amount of serial correlation
    - den Haan and Levin (1997) provided early complete Monte Carlo assessment
    - We will look at MC results later

- Why? The key insight is that the min MSE problem isn't actually what we are interested in – we are actually interested in size control or equivalently coverage rates of confidence intervals.
    - For coverage rates of confidence intervals, what matters is not bias^2, but bias (Velasco & Robinson [2001]; Kiefer & Vogelsang [2002]; Sun, Phillips, and Jin (2008))

- Practical implication: use fewer periodogram ordinates (smaller *B*) i.e. more autocovariances (larger *m*).

<u>Approach #2: Retain consistency, but minimize size distortion</u>

Sketch of asymptotic expansion of size distortion

for details see Velasco and Robinson (2001), Sun, Phillips, and Jin (2008)

Consider the case of a single X and the null hypothesis $\beta = \beta_0$. Then $u_t = Y_t - X_t\beta_0$, and $Z_t = X_t u_t$, so the Wald test statistic is,

$$W_T = \frac{\left(T^{-1/2}\sum_1^T Z_t\right)^2}{\hat{\Omega}}$$

The probability of rejection under the null thus is,

$$\Pr[W_T < c] = \Pr\left[\frac{\left(T^{-1/2}\sum_1^T Z_t\right)^2}{\hat{\Omega}} < c\right]$$

where c is the asymptotic critical value (3.84 for a 5% test). The size distortion is obtained by expanding this probability…

First, note that $T^{-1/2}\sum_1^T Z_t$ and $\hat{\Omega}$ are asymptotically independent. Now

$$\Pr[W_T < c] = \Pr\left[\frac{\left(T^{-1/2}\sum_1^T Z_t\right)^2}{\hat{\Omega}} < c\right] = \Pr\left[\frac{\left(T^{-1/2}\sum_1^T Z_t\right)^2}{\Omega} < c\frac{\hat{\Omega}}{\Omega}\right]$$

$$= E\left\{\Pr\left[\frac{\left(T^{-1/2}\sum_1^T Z_t\right)^2}{\Omega} < c\frac{\hat{\Omega}}{\Omega}\right]\Bigg|\hat{\Omega}\right\}$$

$$\approx E\left[F\left(c\frac{\hat{\Omega}}{\Omega}\right)\right]$$, where F = chi-squared c.d.f

$$= E\left[F(c) + cF'(c)\left(\frac{\hat{\Omega}-\Omega}{\Omega}\right) + \frac{1}{2}cF''(c)\left(\frac{\hat{\Omega}-\Omega}{\Omega}\right)^2 + \ldots\right]$$

so the size distortion approximation is,

$$\Pr[W_T < c] - F(c) \approx cF'(c)\frac{bias(\hat{\Omega})}{\Omega} + \frac{1}{2}cF''(c)\frac{MSE(\hat{\Omega})}{\Omega^2}$$

or

$$\Pr[W_T < c] - F(c) \approx cF'(c)\frac{bias(\hat{\Omega})}{\Omega} + \frac{1}{2}cF''(c)\frac{\text{var}(\hat{\Omega})}{\Omega^2} + \text{smaller terms}$$

Thus minimizing the size distortion entails minimizing a linear combination of bias and variance – *not* bias^2 and variance

Approach #3: "Fixed b" asymptotics

- Drop consistency – but use correct critical values that account for additional variance (HAR)
  - This decision has a cost – consistency provides first-order asymptotic efficiency of tests – but this isn't worth much if you don't have size control

- Fixed *b* corresponds in our notation to fixed *B* (or, equivalently, to $m \propto T$)
  - The fixed-*b* calculations typically use a FCLT approach, see Kiefer-Vogelsang (2002), Müller (2007), Sun (2013).
  - We will sidestep the FCLT results by using classical results from the spectral density estimation literature for the flat kernel in the frequency domain.

5) Application to Flat Kernel in the Frequency Domain

Consider scalar X_t and flat-kernel in frequency domain:

$$\hat{\hat{\Omega}} = \frac{2\pi}{2B_T}\sum_{l=-B}^{B} I_{\hat{Z}\hat{Z}}\left(\frac{2\pi l}{T}\right) = \frac{2\pi}{B_T}\sum_{l=1}^{B} I_{\hat{Z}\hat{Z}}\left(\frac{2\pi l}{T}\right)$$

- This adjusts the kernel to drop $\omega = 0$ since $I_{\hat{Z}\hat{Z}}(0) = 0$ (OLS residuals are orthogonal to X)
- The second equality holds because
  (i) in scalar case, $I_{ZZ}(\omega) = I_{ZZ}(-\omega)$, and
  (ii) $I_{\hat{Z}\hat{Z}}(0) = 0$ because $d_{\hat{Z}}(0) = 0$ ($\hat{u}_t$ are OLS residuals)
- This kernel plays a special historical role in frequency domain estimation.

We now provide explicit results for the three approaches:

i. Fixed B (this kernel delivers asymptotic t_{2B} inference!)
ii. Min MSE
iii. Min size distortion

i. Fixed *b*

- For this kernel, you don't need to use FCLT approach – the result for its fixed-B distribution is very old and is a cornerstone of classical theory of frequency domain estimation (e.g. Brillinger (1981)). For X_t, u_t stationary, with suitable moment conditions,

  (a) $\hat{\hat{\Omega}} \xrightarrow{d} \Omega \times (\chi^2_{2B}/2B)$, that is,

  $$\hat{\hat{\Omega}} \sim \Omega \times (\chi^2_{2B}/2B)$$

  (b) Moreover $\hat{\hat{\Omega}}$ is asymptotically independent of $T^{-1/2}\sum_1^T Z_t \sim \mathrm{N}(0,\Omega)$

- It follows that, for B fixed, the t statistic has an asymptotic t_{2B} distribution:

  $$t = \frac{T^{-1/2}\sum_1^T Z_t}{\hat{\Omega}^{1/2}} \xrightarrow{d} t_{2B}$$

- This result makes the size/power tradeoff clear – using t_{2B} distribution has power loss relative to asymptotically efficient normal inference – but the power loss is slight for $B \geq 10$ (say).

Sketch of (a) and (b):

Consider scalar case, and recall that $I_{\hat{Z}\hat{Z}}(0) = 0$ (OLS residuals), so

(a) Distribution of $\hat{\hat{\Omega}}$ with B fixed:

$$\hat{\hat{\Omega}} = \frac{2\pi}{B}\sum_{l=1}^{B} I_{\hat{Z}\hat{Z}}\left(\frac{2\pi l}{T}\right)$$

$$\sim \frac{2\pi}{B}\sum_{l=1}^{B} S_{ZZ}\left(\frac{2\pi l}{T}\right)\xi_l \text{, where } \xi_l \sim \chi_2^2 / 2$$

$$= \frac{2\pi}{B}\sum_{l=1}^{B}\left[S_{ZZ}(0) + \frac{1}{2}\left(\frac{2\pi l}{T}\right)^2 S_{ZZ}''(0) + \ldots\right]\xi_l$$

$$\approx \frac{2\pi}{B}\sum_{l=1}^{B} S_{ZZ}(0)\xi_l$$

$$= 2\pi S_{zz}(0) \times (\chi_{2B}^2 / 2B)$$

$$= \Omega \times (\chi_{2B}^2 / 2B)$$

(b) $\hat{\hat{\Omega}}$ is independent of $T^{-1/2}\sum_1^T Z_t$. This follows from the result above that $d_Z(\omega_l)$ and $d_Z(\omega_k)$ are asymptotically independent, applied here to $d_Z(0)$ (the numerator) and d_Z at other ω_l's (the denominator)

ii. and iii. – Preliminaries for the asymptotic expansions

Bias

$$E\left(\hat{\hat{\Omega}}-\Omega\right) = E\left[\frac{2\pi}{B}\sum_{l=1}^{B} I_{\hat{Z}\hat{Z}}\left(\frac{2\pi l}{T}\right) - S_{ZZ}(0)\right]$$

$$\approx \frac{2\pi}{B}\sum_{l=1}^{B}\left[S_{ZZ}\left(\frac{2\pi l}{T}\right) - S_{ZZ}(0)\right]$$

$$= \frac{2\pi}{B}\sum_{l=1}^{B}\left\{\left[S_{ZZ}(0) + \frac{2\pi l}{T}S_{ZZ}'(0) + \frac{1}{2}\left(\frac{2\pi l}{T}\right)^2 S_{ZZ}''(0) + \ldots\right] - S_{ZZ}(0)\right\}$$

$$= \frac{2\pi}{B}\sum_{l=1}^{B}\left\{\left[S_{ZZ}(0) + \frac{2\pi l}{T}S_{ZZ}'(0) + \frac{1}{2}\left(\frac{2\pi l}{T}\right)^2 S_{ZZ}''(0) + \ldots\right] - S_{ZZ}(0)\right\}$$

Because $S_{ZZ}(\omega) = S_{ZZ}(-\omega)$, $S_{ZZ}'(0) = 0$, and after dividing by Ω,

$$E\left(\hat{\hat{\Omega}}-\Omega\right)\Big/\Omega = \left[\frac{2\pi}{B}\sum_{l=1}^{B}\frac{1}{2}\left(\frac{2\pi l}{T}\right)^2\right]S_{ZZ}''(0)\Big/2\pi S_{ZZ}(0) = \frac{1}{2d}\left(\frac{B}{T}\right)^2$$

where $d = \dfrac{3S_{ZZ}(0)}{4\pi^2 S_{ZZ}''(0)}$.

Variance

$$\frac{\operatorname{var}(\hat{\hat{\Omega}})}{\Omega^2} = \operatorname{var}\left[\frac{2\pi}{B}\sum_{l=1}^{B} I_{\hat{Z}\hat{Z}}\left(\frac{2\pi l}{T}\right)\right]\Big/\Omega^2$$

$$\approx \frac{4\pi^2}{B^2}\sum_{l=1}^{B}\operatorname{var}\left[I_{ZZ}\left(\frac{2\pi l}{T}\right)\right]\Big/\left(2\pi S_{ZZ}(0)\right)^2$$

$$= \frac{4\pi^2}{B^2}\sum_{l=1}^{B} S_{ZZ}\left(\frac{2\pi l}{T}\right)^2 \Big/ 4\pi^2 S_{ZZ}(0)^2 = \ldots = \frac{1}{B}$$

(keeping only the leading term in the Taylor series expansion).

Summary: relative bias and relative variance:

$$\frac{\operatorname{var}(\hat{\hat{\Omega}})}{\Omega^2} = \frac{1}{B} \quad \text{and} \quad \frac{E(\hat{\hat{\Omega}}-\Omega)}{\Omega} = \frac{1}{2d}\left(\frac{B}{T}\right)^2, \text{ where } d = \frac{3S_{ZZ}(0)}{4\pi^2 S_{ZZ}''(0)}$$

Special case: Z_t is AR(1) with autoregressive parameter $\alpha\neq 0$:

$$d = -\frac{3}{8\pi^2}\frac{(1-\alpha)^2}{\alpha}$$

ii. Min MSE

$$\text{Min}_B \text{ MSE}(\hat{\hat{\Omega}}) = \text{Min}_B \text{ bias}^2(\hat{\hat{\Omega}}) + \text{var}(\hat{\hat{\Omega}})$$

$$= \text{Min}_B \left[\frac{1}{2d}\left(\frac{B}{T}\right)^2 \Omega \right]^2 + \frac{\Omega^2}{B}$$

Solution:

$$B_T^{MinMSE}(\hat{\alpha}) = \left[d\right]^{2/5} T^{4/5}, \text{ where } d = \frac{3S_{ZZ}(0)}{4\pi^2 S_{ZZ}''(0)} = -\frac{3}{8\pi^2}\frac{(1-\alpha)^2}{\alpha}$$

iii. Min Size Distortion

$$\text{Min}_B \Pr[W_T < c] - F(c) \approx \text{Min}_B\, cF'(c)\frac{bias(\hat{\Omega})}{\Omega} + \frac{1}{2}cF''(c)\frac{\text{var}(\hat{\Omega})}{\Omega^2}$$

Solution (for $\alpha > 0$):

$$B_T^{1stOrderSize}(\hat{\alpha}) = \left[\frac{cF''(c)}{2F'(c)} d\right]^{1/3} T^{2/3}$$

where c = 3.84 for 5% tests and F is χ_1^2 cdf.

Optimal HAC Bandwidths for flat spectral kernel: Z_t AR(1) with parameter α

	T = 100				T = 800			
Minimize:	*MSE*		*Size distortion*		*MSE*		*Size distortion*	
α	B	m	B	m	B	m	B	m
.1	43	5	25	8	131	6	62	13
.2	30	7	18	11	90	9	45	18
.3	23	9	14	14	69	12	36	22
.4	18	11	12	17	54	15	30	27
.5	14	14	10	21	43	19	25	33
.6	11	18	8	25	33	24	20	40
.7	8	24	6	32	25	32	16	51
.8	6	35	5	44	17	47	11	70
.9	3	65	3	73	9	85	7	116

Notes: b = bandwidth in frequency domain, m = lag truncation parameter in time domain.

- The rule-of-thumb $m = .75T^{1/3}$ corresponds to m = 4 for T = 100 and m = 7 for T = 800 (however not directly comparable since the rule-of-thumb is for the Newey-West kernel).

6) Monte Carlo Comparisons

<u>Illustrative results:</u>

- Design: $X_t = 1$, u_t AR(1)
- Flat spectral kernel (so that t_{2B} inference is asymptotically valid under fixed-*b* asymptotics)
- Two bandwidth choices: min MSE and minimize size distortion
- Bandwidths chosen using plug-in formula based on estimated α (formula given above, with $\hat{\alpha}$ replacing α)
- Additional MC results: den Haan and Levin (1997), Kiefer and Vogelsang (2002), Kiefer, Vogelsang and Bunzel (2000), Sun (2013).

Null Rejection Rate

		χ^2 c.v.		t c.v.	
φ	T	B_T^{MinMSE}	$B_T^{\text{1stOrderSize}}$	B_T^{MinMSE}	$B_T^{\text{1stOrderSize}}$
0.00	100	0.055	0.055	0.050	0.049
	400	0.052	0.052	0.051	0.050
0.50	100	0.094	0.088	0.075	0.066
	400	0.068	0.064	0.061	0.055
0.90	100	0.216	0.212	0.141	0.132
	400	0.111	0.107	0.083	0.073
0.95	100	0.310	0.309	0.195	0.190
	400	0.149	0.144	0.102	0.092

Table 1: Null rejection rates for tests based on χ^2 and t critical values, and on two different bandwidth formulas. 50,000 Monte Carlo repetitions.

7) Panel Data and Clustered Standard Errors

Clustered standard errors are an elegant solution to the HAC/HAR problem in panel data.

- Although the original proofs of clustered SEs used large N and small T (Arellano [2003]) in fact they are valid for small N if T is large (Hansen [2007], Stock and Watson [2008]), but using t or F (not normal or chi-squared) inference.
- The standard fixed effects panel data regression model

$$Y_{it} = \alpha_i + \beta' X_{it} + u_{it},\ i = 1,\ldots,N,\ t = 1,\ldots, T,$$

  where $E(u_{it}|X_{i1},\ldots, X_{iT}, \alpha_i) = 0$ and u_{it} is uncorrelated across i but possibly serially correlated, with variance that can depend on t; assume i.i.d. over i
- The discussion here considers the special case $X_t = 1$– the ideas generalize

Clustered SEs with $X_t = 1$

$$Y_{it} = \alpha_i + \beta + u_{it},\ i = 1,\ldots,N,\ t = 1,\ldots, T,$$

The fixed effects (FE) estimator is

$$\hat{\beta}^{FE} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}Y_{it}$$

Thus

$$\sqrt{NT}(\hat{\beta}^{FE} - \beta) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}u_{it}\right)$$

$$= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}v_i\ ,\ v_i = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}u_{it}$$

For fixed N and large T, $v_i \xrightarrow{d}$ N(0,Ω), $i = 1,\ldots, N$ (i.i.d.). Thus the problem is asymptotically equivalent to having N observations on v_i, which is i.i.d. N(0,Ω).

$X_t = 1$ case, continued:

Clustered variance formula: $\hat{\Omega}^{cluster} = \frac{1}{N}\sum_{i=1}^{N}(\hat{v}_i - \bar{\hat{v}})^2$, $\hat{v}_i = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\hat{u}_{it}$

By standard normal/t arguments: $\hat{\Omega}^{cluster} \xrightarrow{d} \frac{\Omega\chi^2_{N-1}}{N} = \frac{\Omega\chi^2_{N-1}}{N-1} \times \frac{N-1}{N}$

and $t = \frac{\hat{\beta}^{FE} - \beta_0}{\sqrt{\hat{\Omega}^{cluster}}} \xrightarrow{d} \sqrt{\frac{N}{N-1}}\, t_{N-1}$

- Note the complication of the degrees of freedom correction – this is because the standard definition of $\hat{\Omega}^{cluster}$ has N, not N-1, in the denominator.
- Extension to multiple X: The F-statistic testing p linear restrictions on β, computed using $\hat{\Omega}^{cluster}$, is distributed $\frac{N}{N-p}F_{p,N-p}$
- For N very small, the power loss from t_{N-1} inference can be large – so for very small N it might be better to use HAC/HAR methods, not clustered SEs (not much work has been done on this tradeoff, however).

8) Summary

- Applications of HAC/HAR methods are generic in time series. GLS is typically not justified because it requires strict exogeneity (no feedback from u to X)
- Choice of the bandwidth is critical and reflects a tradeoff between bias and variance.
- The rule-of-thumb $m = .75T^{1/3}$ uses too few autocovariances (m is too small) – overweights variance at the expense of bias
- However, inference becomes complicated when large m (small B) is used, because this increases the variance of $\hat{\Omega}$.
- In general (including for N-W weights), fixed-b inference is complicated and requires specialized tables (e.g. Kiefer-Vogelsang inference).
- However, in the special case of the flat spectral kernel, asymptotically valid fixed-B inference is based on t_{2B}. Initial results for size control (and power) using this approach are promising.