NBER Summer Institute Minicourse –

What's New in Econometrics: Time Series

Lecture 9

July 16, 2008

Heteroskedasticity and Autocorrelation Consistent Standard Errors

Outline

1. What are HAC SE's and why are they needed?
2. Parametric and Nonparametric Estimators
3. Some Estimation Issues (*psd*, *lag choice, etc.*)
4. Inconsistent Estimators

1. What are HAC SE's and why are they needed?

Linear Regression: $y_t = x_t'\beta + e_t$

$$\hat{\beta} = \left(\sum_{t=1}^{T} x_t x_t' \right)^{-1} \left(\sum_{t=1}^{T} x_t y_t' \right) = \beta + \left(\sum_{t=1}^{T} x_t x_t' \right)^{-1} \left(\sum_{t=1}^{T} x_t e_t' \right)$$

$$\sqrt{T}(\hat{\beta} - \beta) = \left(\frac{1}{T} \sum_{t=1}^{T} x_t x_t' \right)^{-1} \left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} x_t e_t' \right) = S_{XX}^{-1} A_T .$$

$S_{XX} \xrightarrow{p} \Sigma_{XX} = E(x_t x_t')$, $A_T \xrightarrow{d} N(0, \Omega)$, where $\Omega = \lim_{T \to \infty} \mathrm{Var}\left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} x_t e_t' \right)$

Thus $\hat{\beta} \overset{a}{\sim} N\left(\beta, \frac{1}{T} \hat{V} \right)$, where $\hat{V} = S_{XX}^{-1} \hat{\Omega} S_{XX}^{-1}$

HAC problem : $\hat{\Omega}$ = ???

(Same problem arises in IV regression, GMM, …)

Notation:

$$\Omega = \lim_{T\to\infty} \mathrm{Var}\left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} x_t e_t ' \right)$$

Let $w_t = x_t e_t$ (assume this is a scalar for convenience)

Feasible estimator of Ω must use $\hat{w}_t = x_t \hat{e}_t$,

The estimator is $\hat{\Omega}\left(\{\hat{w}_t\}\right)$, but most of our discussion uses $\hat{\Omega}\left(\{w_t\}\right)$ for convenience.

An expression for Ω: Suppose w_t is covariance stationary with autocovariances γ_j. (Also, for notational convenience, assume w_t is a scalar). Then

$$\Omega = var(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} w_t) = \frac{1}{T} var(w_1 + w_2 + ... + w_T)$$

$$= \frac{1}{T}\{T\gamma_0 + (T-1)(\gamma_1 + \gamma_{-1}) + (T-2)(\gamma_2 + \gamma_{-2}) + ...1\times(\gamma_{T-1} + \gamma_{1-T})\}$$

$$= \sum_{j=-T+1}^{T-1} \gamma_j - \frac{1}{T}\sum_{j=1}^{T-1} j(\gamma_j + \gamma_{-j})$$

If the autocovariances are "1-summable" so that $\sum j\,|\gamma_j| < \infty$ then

$$\Omega = var(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} w_t) \to \sum_{j=-\infty}^{\infty} \gamma_j$$

Recall spectrum of w at frequency ω is $S(\omega) = \frac{1}{2\pi}\sum_{j=-\infty}^{\infty} \gamma_j e^{-i\omega j}$, so that $\Omega = 2\pi\times S(0)$. Ω is called the "Long-run" variance of w.

II. Estimators

(a) Parametric Estimators, $w_t \sim$ ARMA

$$\phi(\mathrm{L}) w_t = \theta(\mathrm{L}) \varepsilon_t,$$

$$S(\omega) = \frac{1}{2\pi} \sigma_\varepsilon^2 \frac{\theta(e^{-i\omega})\theta(e^{i\omega})}{\phi(e^{-i\omega})\phi(e^{i\omega})}$$

$$\Omega = 2\pi \times S(0) = \sigma_\varepsilon^2 \frac{\theta(1)^2}{\phi(1)^2} = \sigma_\varepsilon^2 \frac{(1-\theta_1-\theta_2-\cdots-\theta_q)^2}{(1-\phi_1-\phi_2-\cdots-\phi_p)^2}$$

$$\hat{\Omega} = \hat{\sigma}_\varepsilon^2 \frac{(1-\hat{\theta}_1-\hat{\theta}_2-\cdots-\hat{\theta}_q)^2}{(1-\hat{\phi}_1-\hat{\phi}_2-\cdots-\hat{\phi}_p)^2}$$

Jargon: VAR-HAC is a version of this. Suppose w_t is a vector and the estimated VAR using w_t is

$$w_t = \hat{\Phi}_1 w_{t-1} + \ldots + \hat{\Phi}_p w_{t-p} + \hat{\varepsilon}_t,$$

where $\hat{\Sigma}_\varepsilon = \frac{1}{T}\sum_{t=1}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_t'$

Then $\hat{\Omega} = (I - \hat{\Phi}_1 - \ldots - \hat{\Phi}_p)^{-1} \hat{\Sigma}_\varepsilon (I - \hat{\Phi}_1 - \ldots - \hat{\Phi}_p)^{-1\prime}$

(b) Nonparametric Estimators

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

$$\hat{\Omega} = \sum_{j=-m}^{m} K_j \hat{\gamma}_j$$

$$\text{with } \hat{\gamma}_j = \frac{1}{T} \sum_{t=j+1}^{T} w_t w_{t-j} = \hat{\gamma}_{-j} \quad (j \geq 0)$$

III. Issues

(a) $\hat{\Omega} \geq 0$?

(b) Number of lags (m) in non-parameter estimator or order of ARMA in parametric estimator

(c) form of K_j weights

(a) $\hat{\Omega} \geq 0$?

$$\hat{\Omega} = \hat{\sigma}_{\varepsilon}^{2} \frac{(1-\hat{\theta}_1-\hat{\theta}_2-\cdots-\hat{\theta}_q)^2}{(1-\hat{\phi}_1-\hat{\phi}_2-\cdots-\hat{\phi}_p)^2} \quad \text{(yes)}$$

$$\hat{\Omega} = \sum_{j=-m}^{m} K_j \hat{\gamma}_j \quad \text{(Not necessarily)}$$

MA(1) Example: $\Omega = (\gamma_{-1} + \gamma_0 + \gamma_1)$ and $\hat{\Omega} = (\hat{\gamma}_{-1} + \hat{\gamma}_0 + \hat{\gamma}_1)$

$|\gamma_1| \leq \gamma_0/2$ … but $|\hat{\gamma}_1| > \hat{\gamma}_0 / 2$ is possible.

Newey-West (1987): Use $K_j = 1 - \frac{|j|}{m+1}$ (Bartlett "Kernel")

An alternative expression for $\hat{\Omega} = \sum_{j=-m}^{m} K_j \hat{\gamma}_j$

Let $W = (w_1\ w_2\ \ldots\ w_T)'$ where $W \sim (0, \Sigma_{WW})$, $U = HW$, $U \sim (0, \Sigma_{UU})$ with $\Sigma_{UU} = H\Sigma_{WW}H'$. Choose H so that $\Sigma_{UU} = D$ a diagonal matrix.

A useful result: with W covariance stationary (Σ_{WW} is "special"), a particular H matrix yields (approximately) $\Sigma_{UU} = D$ (diagonal). (The u's are the coefficients from the discrete Fourier transform of the w's.)

Variable	$2\pi \times$Variance (D_{ii})
u_1	$S(0)$
u_2 and u_3	$S(1\times 2\pi/T)$
u_4 and u_5	$S(2\times 2\pi/T)$
u_6 and u_7	$S(3\times 2\pi/T)$
…	…
u_{T-1} and u_T (T odd, similar for T even)	$S([(\mathrm{T}-1)/2]\times 2\pi/T)$

Some algebra shows that

$$\hat{\Omega} = \sum_{j=-m}^{m} K_j \hat{\gamma}_j = k_0 u_1^2 + \sum_{j=1}^{(T-1)/2} k_j \left(\frac{u_{2j}^2 + u_{2j+1}^2}{2} \right)$$

where the *k*-weights are functions (Fourier transforms) of the *K*-weights.

The algebraic details are unimportant for our purposes – but, for those interested, they are sketched on the next slide. Jargon: the term $\left(\frac{u_{2j}^2 + u_{2j+1}^2}{2} \right)$ is called the *j'*th periodogram ordinates. Because $E(u_{2j}^2) = E(u_{2j+1}^2) = 2\pi S\left(\frac{2\pi j}{T} \right)$, the *j'th* periodogram ordinate is an estimate of the spectrum at frequency $(2\pi j/T)$.

<u>The important point:</u> Evidently, $\hat{\Omega} \geq 0$ requires $k_j \geq 0$ for all j.

The algebra linking the data *w*, the *u's*, the periodogram ordinates and the sample covariances parallels the expressions the spectrum presented in lecture 1. (Ref: Fuller (1976), or Priestly (1981), or Brockwell and Davis (1991)).

In particular, let $c_j = \frac{1}{\sqrt{2T}} \sum_{t=1}^{T} w_t e^{i\omega_j t}$ where $\omega_j = 2\pi j/T$

Then $\left(\frac{u_{2j}^2 + u_{2j+1}^2}{2} \right) = |c_j|^2 = \sum_{p=-T-1}^{T-1} \hat{\gamma}_p \cos(p\omega_j),$

where $\hat{\gamma}_p = \frac{1}{T} \sum_{t=p+1}^{T} (w_t - \bar{w})(w_{t-p} - \bar{w})$

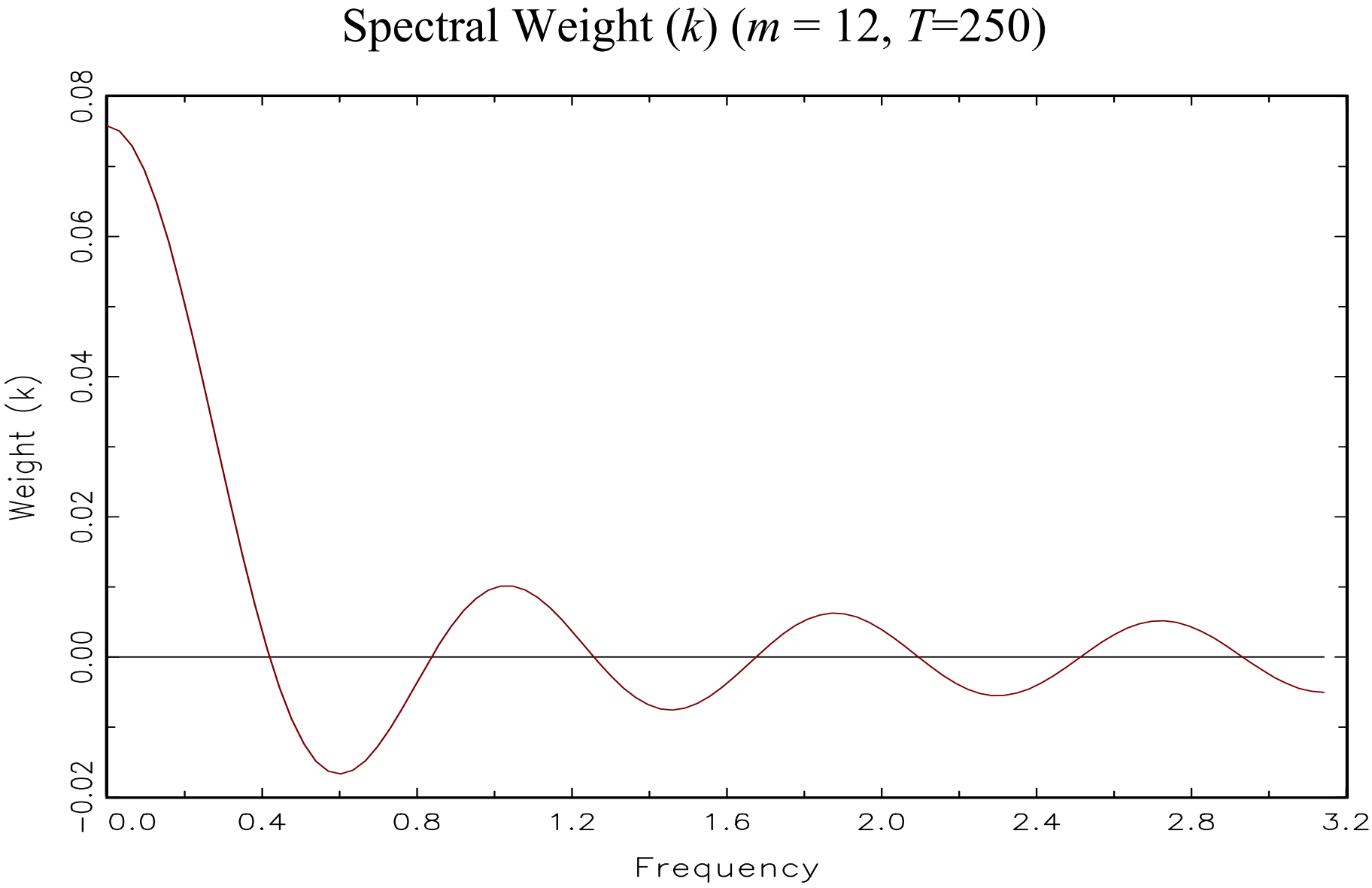
Thus $\left(\frac{u_{2j}^2 + u_{2j+1}^2}{2} \right) = |c_j|^2$ is a natural estimator of the spectrum at frequency ω_j.

$$\hat{\Omega} = \sum_{j=-m}^{m} K_j \hat{\gamma}_j = k_0 u_1^2 + \sum_{j=1}^{(T-1)/2} k_j \left(\frac{u_{2j}^2 + u_{2j+1}^2}{2} \right)$$

is then a weighted average (with weights *k* of the estimated spectra at difference frequencies. Because the goal is to estimate at $\omega = 0$, the weights should concentrate around this frequency as *T* gets large.

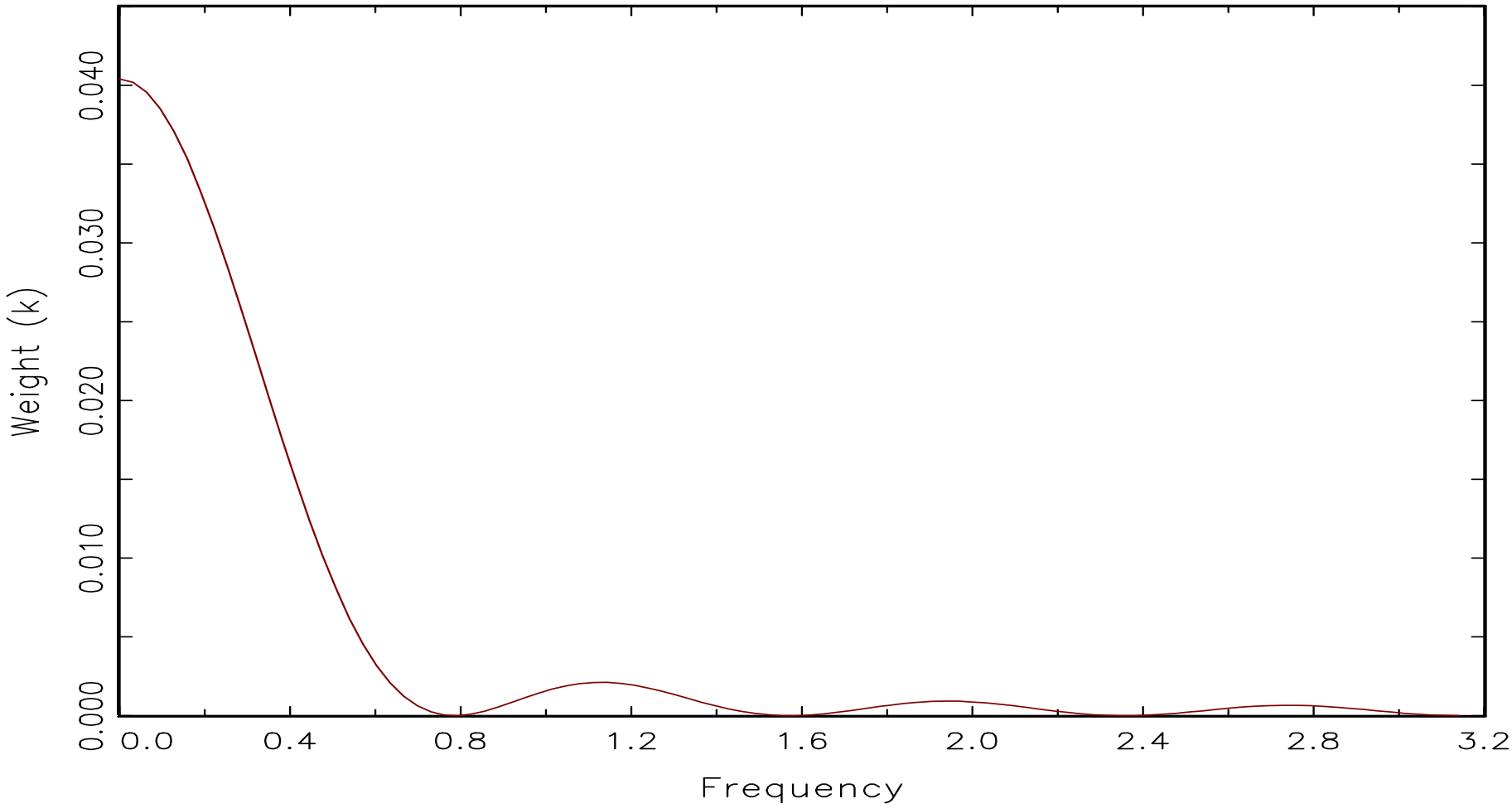
Jargon: Kernel (*K-weights*, sometimes *k*-weights), Lag Window (*K-weights*, RATS follows this notation and calls these "lwindow" in the code), Spectral Window (*k-weights*)

Truncated Kernel: $K_j = 1(|j| \leq m)$



Newey-West (Bartlet Kernel): $K_j = 1 - \frac{|j|}{m+1}$

Spectral Weight (k) (m = 12, T=250)



Implementing Estimator (lag order, kernel choice and so forth)

Parametric Estimators:

AR/VAR approximations: Berk (1974)

Ng and Perron (2001), lag lengths in ADF tests. Related, but different issues.

Shorter lags: Larger Bias, Smaller Variance

Longer lags: Smaller Bias, Larger Variance

Practical advice: Choose lags a little longer than you might otherwise (rational given below).

How Should m be chosen?

Andrews (1991) and Newey and West (1994): minimize mse($\hat{\Omega} - \Omega$)

MSE = Variance + Bias^2

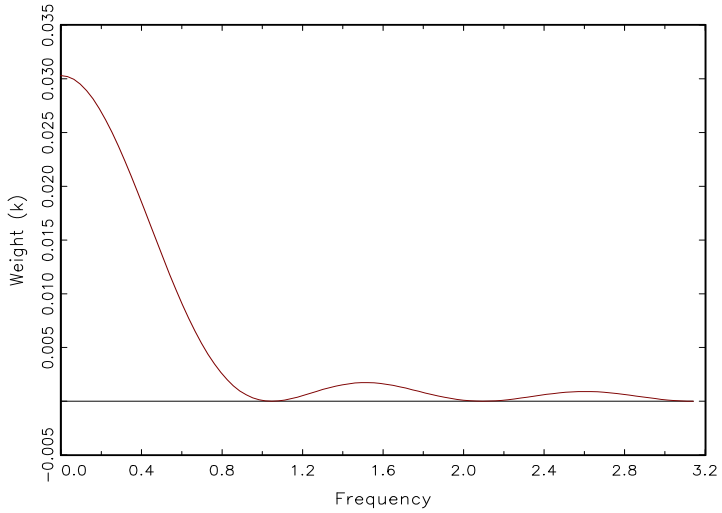
$$\hat{\Omega} = k_0 u_1^2 + \sum_{j=1}^{(T-1)/2} k_j \left(\frac{u_{2j}^2 + u_{2j+1}^2}{2} \right)$$

Variance: Spread the weight out over many of the squared u's: Make the spectral weights flat.

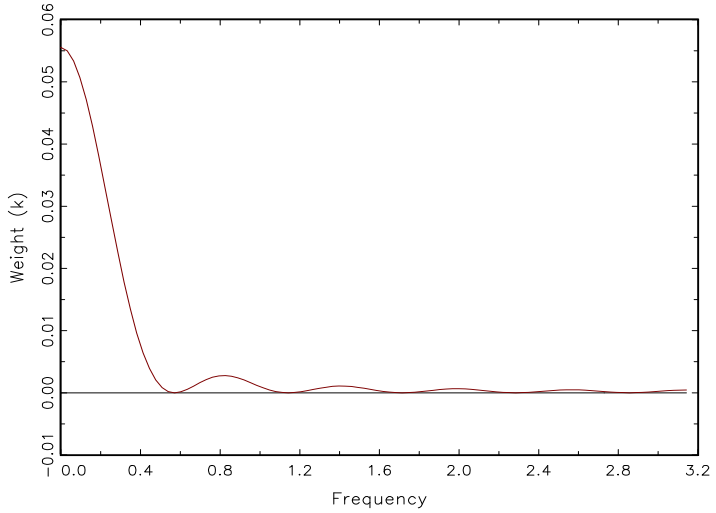
Bias: $E(\hat{\Omega})$ depends on the values $E\left(\frac{u_{2j}^2 + u_{2j+1}^2}{2} \right) \approx S(j \times 2\pi/T)$. So the bias depends on how flat the spectrum is around $\omega = 0$. The more flat it is, the smaller is the bias. The more curved it is, the higher is the bias.

Spectral Window (k_j) for N-W/Bartlett Lag Window ($K_j = 1 - \frac{|j|}{m+1}$), $T = 250$.

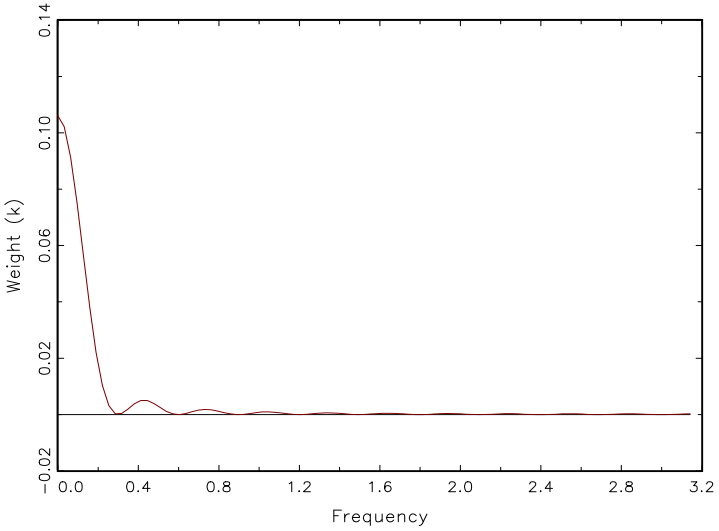
$m = 5$



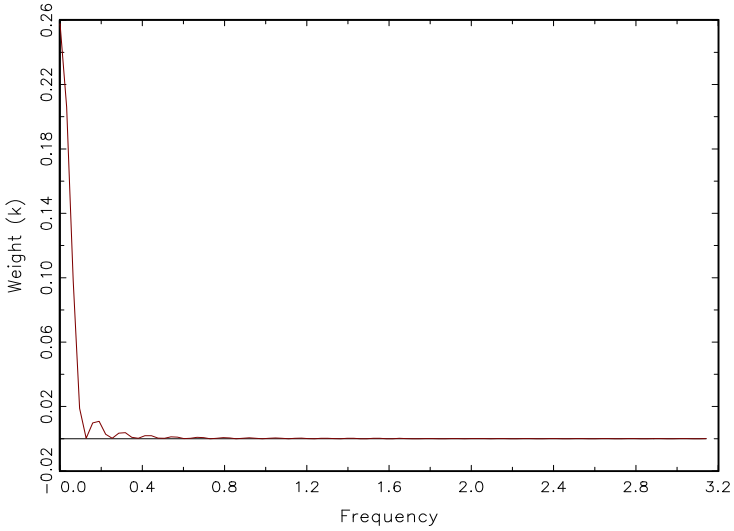
$m = 10$



$m = 20$



$m = 50$



MSE minimizing values for $w_t = \phi w_{t-1} + e_t$, Bartlett Kernel (note … the higher is ϕ, more serial correlation, more "curved" spectrum around $\omega = 0$, more bias for given value of m.)

$$m^* = 1.1447 \times 4^{1/3} \times (\phi^2)^{1/3} \times T^{1/3} = 1.82 \times \phi^{2/3} \times T^{1/3}$$

ϕ	m^*	T		
		100	**400**	**1000**
0.00	0	0	0	0
0.25	$0.72 \times T^{1/3}$	3	5	7
0.50	$1.15 \times T^{1/3}$	5	8	11
0.75	$1.50 \times T^{1/3}$	6	11	15
0.90	$1.70 \times T^{1/3}$	7	12	17
0.95	$1.76 \times T^{1/3}$	8	12	17

Kernel Choice: Some Kernels (lag windows)

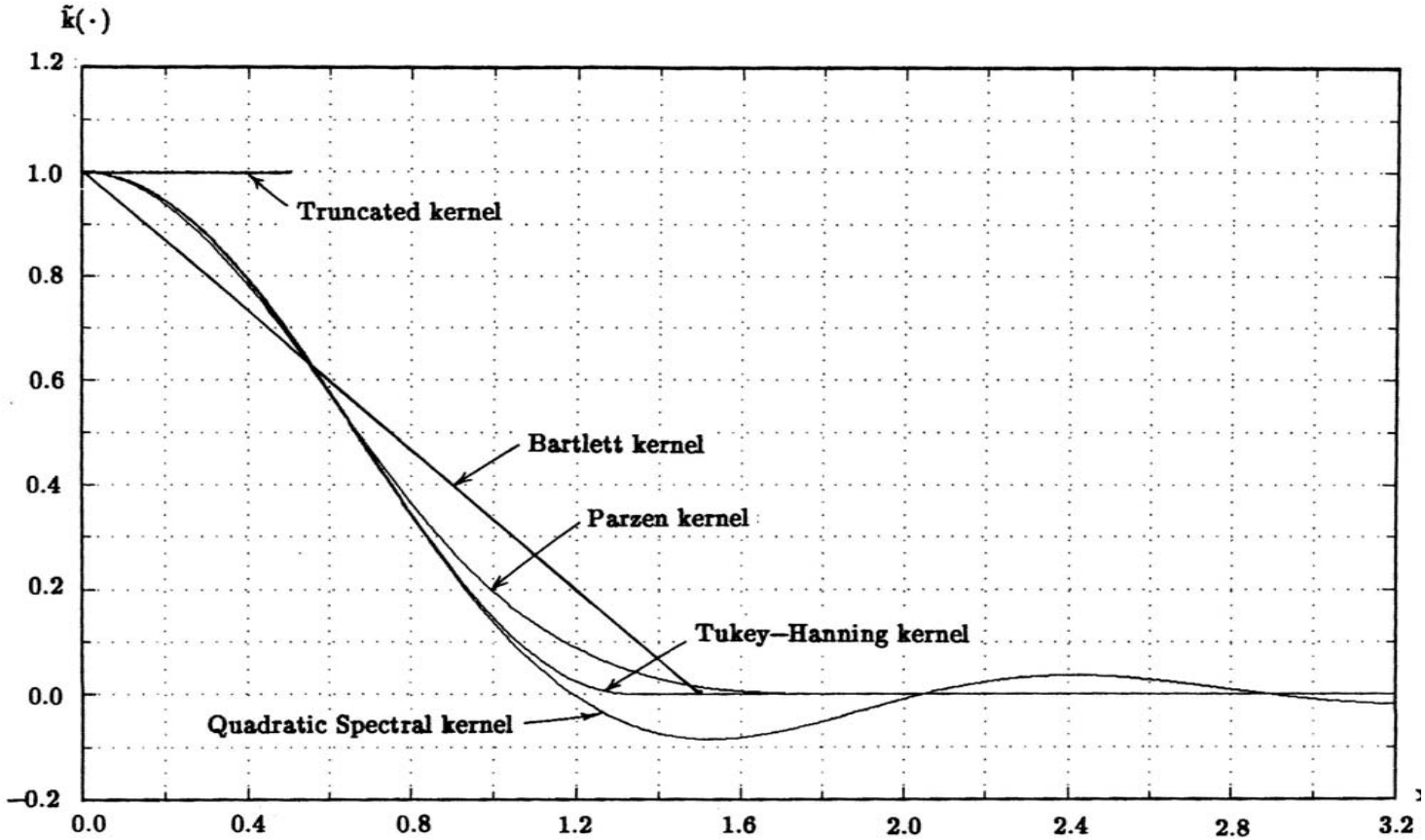


FIGURE 1.—Comparison of kernels.[a]

[a]These kernels have been renormalized as described in the text below equation (2.7).

Does Kernel Choice Matter?

In theory, yes. (QS is optimal)

In practice (for psd kernels), not really.

Does lag length matter? Yes, quite a bit …

$y_t = \beta + w_t$, $w_t = \phi w_{t-1} + e_t$, $T = 250$.

Rejection of 2-sided 10% test

ϕ	$\hat{m}^*$	AR	$\hat{m}^*$(PW)		m^*	$2\times m^*$
0.00	0.10	0.11	0.11		0.10	0.10
0.25	0.13	0.11	0.11		0.13	0.13
0.50	0.15	0.11	0.11		0.15	0.14
0.75	0.21	0.12	0.12		0.21	0.18
0.90	0.33	0.15	0.15		0.33	0.25
0.95	0.46	0.20	0.19		0.46	0.37

(More Simulations: den Haan and Levin (1997) … available on den Haan's web page … and other papers listed throughout these slides)

Why use more lags than MSE Optimal (Sun, Phillips, and Jin (2008))

Intuition: $z \sim \mathrm{N}(0,\sigma^2)$, $\hat{\sigma}^2$ an estimator of σ^2 (assumed independent of z)

$$\Pr\left(\frac{z^2}{\hat{\sigma}^2} < c\right) = \Pr\left(z^2 < \hat{\sigma}^2 c\right) = E\left(1(z^2 < \hat{\sigma}^2 c)\right) = E\left(g(\hat{\sigma}^2)\right)$$

$$\approx E\left(g(\sigma^2)\right) + E(\hat{\sigma}^2 - \sigma^2)g'(\sigma^2) + \frac{1}{2}E\left((\hat{\sigma}^2 - \sigma^2)^2\right)g''(\sigma^2)$$

$$= F_{\chi_1^2}(c) + Bias(\hat{\sigma}^2) \times g' + \frac{1}{2}MSE(\hat{\sigma}^2) \times g''$$

MSE … Bias^2 + Variance

This formula … Bias and MSE … Thus m should be bigger

IV. Making m really big $m = bT$ where b is fixed. (Kiefer, Vogelsang and Bunzel (2000), Kiefer and Vogelsang (2002), Kiefer and Vogelsang (2005))

Bartlett … $m = (T-1)$: $\hat{\Omega}(\hat{w}) = \sum_{j=-T+1}^{T-1} (1 - \frac{|j|}{T})\hat{\gamma}_j$

$\hat{\gamma}_j = \frac{1}{T}\sum_{t=j+1}^{T} \hat{w}_t \hat{w}_{t-j} = \hat{\gamma}_{-j}$, and where $\hat{w}_t = w_t - \overline{w}$.

KV (2002) show that some rearanging yields : $\hat{\Omega}(\hat{w}) = 2T^{-1}\sum_{t=1}^{T}\left(T^{-1/2}\sum_{j=1}^{t} \hat{w}_j \right)^2$

But, $T^{-1/2}\sum_{j=1}^{[sT]} w_j \xrightarrow{d} \Omega^{1/2} W(s)$, $T^{-1/2}\sum_{j=1}^{[sT]} \hat{w}_j \xrightarrow{d} \Omega^{1/2}(W(s) - sW(1))$, and

$T^{-1/2}\sum_{j=1}^{[\cdot T]} \hat{w}_j \Rightarrow \xi(\cdot)$, where $\xi(s) = \Omega^{1/2}(W(s) - sW(1))$.

Thus

$$\hat{\Omega}(\hat{w}):\ \hat{\Omega}(\hat{w}) = 2T^{-1}\sum_{t=1}^{T}\left(T^{-1/2}\sum_{j=1}^{t}\hat{w}_j\right)^2 \xrightarrow{d} \Omega 2\int_0^1 (W(s) - sW(1))^2\, ds$$

So that $\hat{\Omega}(\hat{w})$ is a "robust," but inconsistent estimator of Ω.

Distributions of "*t*-statistics" (or other statistics using this $\hat{\Omega}$) will not be asymptotically normal, but large sample distributions easily tabulated (see KV (2002)).

Table of *t* cricitical values

TABLE I

ASYMPTOTIC CRITICAL VALUES OF t^*

1.0%	2.5%	5.0%	10.0%	50.0%	90.0%	95%	97.5%	99.0%
–6.090	–4.771	–3.764	–2.740	0.000	2.740	3.764	4.771	6.090

Source: Line 1 of Table I from Abadir and Paruolo (1997, p. 677) scaled by $1/\sqrt{2}$.

F-critical values .. See KVB (DIVIDE BY 2 !)

Power "loss" (From KVB)

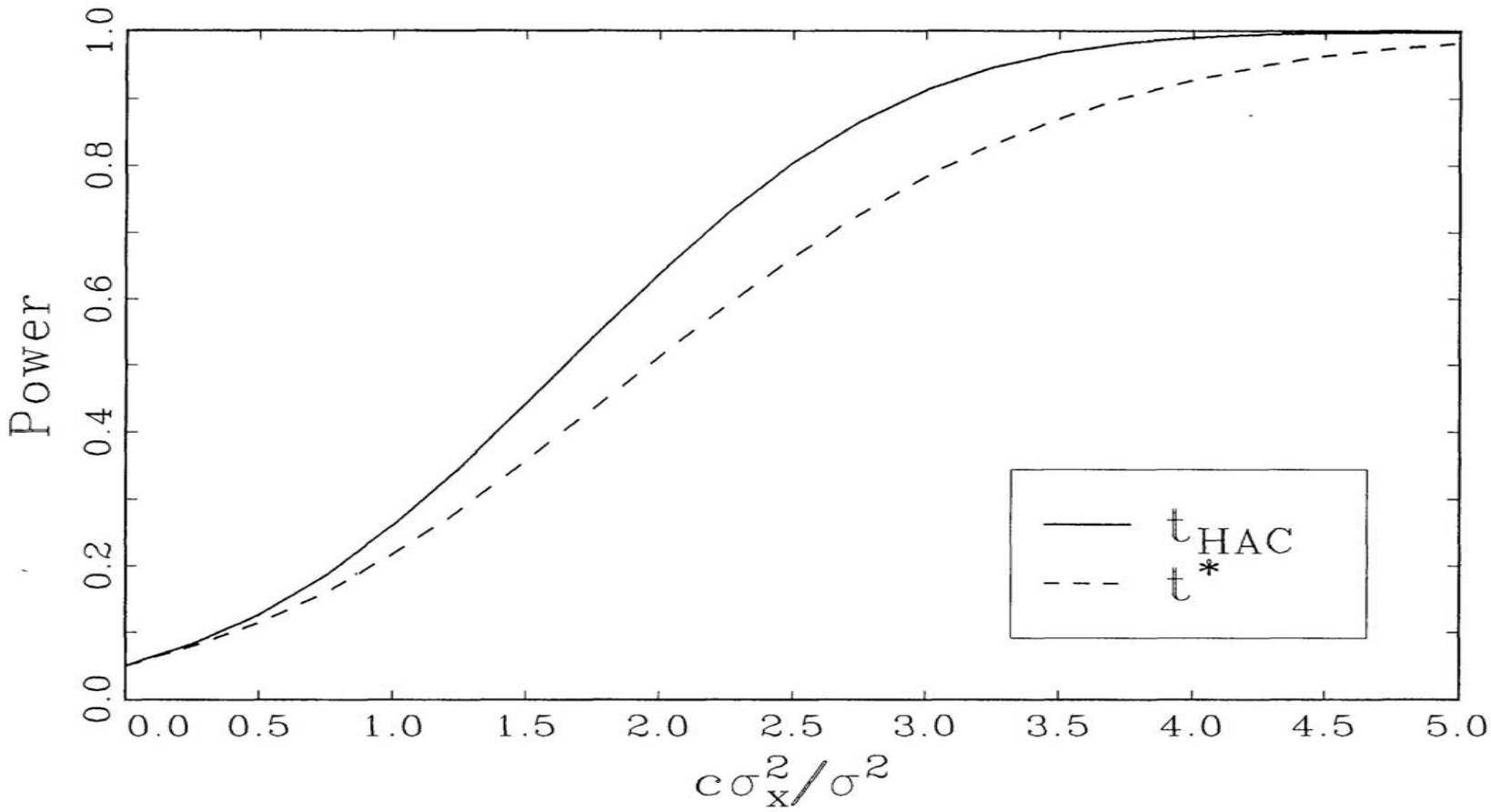


FIGURE 2.—Local asymptotic power, 5% nominal size, $k = 1$, $y_t = \beta x_t + u_t$, H_0: $\beta \leqq \beta_0$, H_1: $\beta > \beta_0 + cT^{-1/2}$.

Size Control (Finite Sample AR(1))

TABLE V

FINITE SAMPLE NULL REJECTION PROBABILITIES AR(1)-HOMO MODEL, $T = 256, 512$
2,000 REPLICATIONS, NOMINAL LEVEL 0.05; ASYMPTOTIC CRITICAL VALUES USED

Model	ρ	F^*	QS	QS-PW	Model	ρ	F^*	QS	QS-PW
	−0.5	0.050	0.069	0.059		−0.5	0.062	0.089	0.071
	−0.3	0.051	0.068	0.061		−0.3	0.049	0.064	0.062
AR(1)-	0.0	0.044	0.057	0.058	AR(1)-	0.0	0.053	0.053	0.056
HOMO	0.3	0.052	0.066	0.061	HOMO	0.3	0.057	0.075	0.067
$q = 1$	0.5	0.054	0.081	0.064	$q = 2$	0.5	0.070	0.095	0.077
$T = 256$	0.7	0.067	0.101	0.078	$T = 256$	0.7	0.095	0.137	0.106
	0.9	0.123	0.191	0.141		0.9	0.170	0.289	0.197
	0.95	0.184	0.297	0.207		0.95	0.263	0.442	0.311
	−0.5	0.063	0.104	0.080		−0.5	0.072	0.120	0.095
	−0.3	0.048	0.075	0.071		−0.3	0.056	0.086	0.078
AR(1)-	0.0	0.052	0.060	0.064	AR(1)-	0.0	0.057	0.064	0.068
HOMO	0.3	0.066	0.082	0.079	HOMO	0.3	0.067	0.090	0.086
$q = 3$	0.5	0.073	0.101	0.096	$q = 4$	0.5	0.084	0.132	0.106
$T = 256$	0.7	0.100	0.173	0.125	$T = 256$	0.7	0.122	0.202	0.146
	0.9	0.213	0.386	0.263		0.9	0.250	0.477	0.342
	0.95	0.330	0.565	0.406		0.95	0.394	0.682	0.502

(a) Advantages ...
(i) Some what better size control (in Monte Carlos and in theory (Jansson (2004)).

(ii) More "Robust" to serial correlation (Müller (2007). Müller proposes estimators that yield "t-statistics" with *t*-distributions.

(b) Disadvantages ...
(i) Wider confidence intervals ... In the one-dimensional regression problem a value of $\beta-\beta_0$ chosen so that β is contained in the (true, not size-distorted) 90% HAC confidence interval with probability 0.10. The value of β is contained in the KVB 90% CI with probability 0.23. (Based on an asymptotic calculation)

(ii) Robustness to volatility breaks ... No ....

KVB Cousins: (Some related work)

(a) (Müller (2007): Just focus on low frequency observations. How many do you have?

Periodogram ordinates at $\frac{2\pi j}{T}, j = 1, 2, \ldots (T\text{-}1)/2$

Let p denote a low frequency "cutoff" … frequencies with periods greater than p are $\omega \leq \frac{2\pi}{p}$. Solving … number of periodogram ordinates with periods $\geq p$: $\frac{T}{p}$

60 years of data, $p = 6$ years, number of ordinates = 10. (Does not depend on sampling frequency).

Müller : do inference based on these obs. Appropriately weighted these yield t-distributions for t-statisics.

(b) Panel Data: Hansen (2007). "Clustered" SEs (over T) when T is large and n is small ($T \to \infty$ and n fixed). This too yields "t-statistics" being distributed t (with $n-1$ df) and appropriately constructed "F-stats" having Hotelling t-squared distribution.

(c) Ibragimov and Müller (2007). Allows heterogeneity.

HAC Bottom line(s)

(1) Parametric Estimators are easy and tend to perform reasonably well. In some applications a parametric model is natural (MA(q) model for forecasting q+1 periods ahead). In other circumstances VAR-HAC is sensible. (den Haan and Levin (1997).) Think about changes in volatility.

(2) Why are you estimating Ω?

(a) Optimal Weighting matrix in GMM: Minimum MSE $\hat{\Omega}$ seems like a good idea. (Analytic results on this ?)

(b) Inference: Use more lags than you otherwise would think you should in Newey-West (or other non-parametric estimators). Worried? Use KVB estimators (and their critical values for tests) or Müller (2007) versions (with t or F critical values).