Bent E. Sørensen
December 1, 2011

1 Teaching notes on GMM III (revised Nov 29, 2011).

1.1 Variance estimation.

Most of the material in this note builds on Anderson (1971), chapters 8 and 9. [This book is now available in the Wiley Classics series]. In this revision, I have put the theory, which isn't on the exam in an appendix.

First recall that

$$\Omega = \lim_{J \to \infty} \sum_{j=-J}^{J} E[f_t f'_{t-j}] \ .$$

Notice, that for any L dimensional vector a we have

$$a'\Omega a = \sum_{j=-J}^{J} a' f_t (a' f_{t-j})' \ ,$$

so, since the quadratic form Ω is characterized by the bilinear mapping $a \to a'\Omega a$ (and similar for estimates $\hat{\Omega}$, you see that the behavior of the estimators are characterized by the actions of the estimator on the univariate processes $a'f_t$. In the following I will therefore look at the theory for spectral estimation for univariate processes, and in this section we will ignore that f_t is a function of an estimated parameter. Under the regularity conditions that is normally used, this is of no consequence asymptotically.

Defining the j'th autocorrelation $\gamma(k) = E f_t f_{t-j}$, our goal is to estimate $\sum_{j=-\infty}^{\infty} \gamma(j)$. Define the estimate (based on T observations) of the j'th autocorrelation by

$$c(j) = \frac{\sum_{t=j}^{T} [f_t f'_{t-j}]}{T} \ \ ; \ j = 0, 1, 2, ... \ .$$

Notice that we do not use the unbiased covariance estimate of the autocovariances (this is obtained by dividing by $T - j$ rather than T).

We will use estimators of the form

$$\hat{\Omega} = \sum_{j=-J}^{J} w_j c(j) \ ,$$

where the w_j are a set of weights. (The reason for these and how to choose them is the subject of most of the following). The dependence of f_t on the estimated parameter will be suppressed in the following, but it is always evaluated at our estimate.

The spectral density is

$$f(\lambda) \ = \ \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma(k) cos(\lambda k) \ .$$

NOTE: f now denotes the spectral density as is common in the literature, it is NOT the moment condition!!!

We only need the spectral density at $\lambda = 0$ but the theory makes use of the whole function and you will hear people talk about "spectral estimator."

In most cases, the weights take the form

$$w_j = k(\frac{j}{K_T}) \ ,$$

where $k()$ is a continuous function (a "kernel"), $k(0) = 1$, $k(x) = k(-x)$, normalized such that the implied w^* satisfies $\int_{-\pi}^{\pi} w^*(\lambda|\nu) d\lambda = 1$ for all ν. We will always assume that K_T tends to infinity with T.

The most commonly used kernel was suggested by Bartlett and popularized in a 1987 Econometrica article by Newey and West. It has the form

$$w_j = 1 - abs(j)/K_T$$

for $abs(j) < K_T$, 0 otherwise. It is also sometimes known as a "tent" kernel (try and draw it).

Andrews (1991) shows the consistency of various kernel smoothed spectral density estimates (at 0 frequency), when the covariances are estimated via estimated orthogonality conditions (or as you would usually say: when you use the error terms rather than the unobserved innovations). In this case, some more regularity conditions, securing that the error term varies smoothly with the estimated parameters, are clearly necessary but since those are usually satisfied in practise and no-one typically checks them, we will not go into the details of this.

Andrews shows that the asymptotically optimal kernel is the Quadratic Spectral (QS) kernel which have the form

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

You may want to try and plot it (using, for example GAUSS). I do not want you to try and remember the exact formula, but remember the name.

Andrew find the optimal bandwidth to have the form

$$K_T^* = 1.1447[\alpha(1)T]^{\frac{1}{3}}$$

for the Bartlett kernel, and

$$K_T^* = 1.3221[\alpha(2)T]^{\frac{1}{5}}$$

for the QS kernel. (Notice how slowly they grow with the number of observations T.)

The α parameter depends on the (unknown) spectral density function at frequency 0, but Andrews suggest that one assume a simple form of the model, e.g. an AR(1) or an ARMA(1,1), or maybe a VAR(1) in the vector case, and use this to obtain an initial estimate of f(0) which one then uses for an estimate of the α parameter. Notice that the important thing here is to get the order of magnitude right, so it is not necessary that the approximating AR(1) (say) model is the “correct” model. In case you knew the correct parametric model for the long run variance you would obtain more efficiency using this model directly rather than relying on non-parametric density estimators. In any event you can show for example for an AR(1) model with autoregressive parameter ρ that

$$\alpha(1) = \frac{4\rho^2}{(1-\rho)^6(1+\rho)^2} / \frac{1}{(1-\rho)^4} .$$

You should plot the one given here in order to get a feel for it—for example, if ρ is 0, the estimated Ω will not use any autocorrelation of order larger than 0. In general, if there is a lot of autocorrelation, we need to include more lags or we will have a lot of bias while, if there is little autocorrelation, we are better of not including a lot of lags since the noise from those will dominate the bias created by leaving them out. (You should know this pattern and you should know there is a formula, but don't try to memorize the exact form of $\alpha(1)$. More formulas are giving in Andrews (1991), you will need for example $\alpha(2)$ to use the QS kernel. Andrews also gives formulas for both $\alpha(1)$ and $\alpha(2)$, for the case where the approximating model is chosen to be an ARMA(1,1), an MA of arbitrary order or a VAR(1) model. Typically the simple AR(1) model is used.

In a typical GMM application you would run an initial estimation, maybe using the identity weighting matrix, then you would obtain an estimate of the orthogonality conditions (in other word, you would get some error terms) and on those you would estimate an AR(1) model, obtaining an estimate $\hat{\rho}$, and you would then find

$$\hat{\alpha}(1) = \frac{4\hat{\rho}^2}{(1-\hat{\rho})^6(1+\hat{\rho})^2} / \frac{1}{(1-\hat{\rho})^4} \ .$$

which you would plug into your formula for the optimal bandwidth [this would be for the Bartlett kernel, for the QS kernel you would obviously have to find $\alpha(2)$].

Usually you will have multivariate models and you would have to estimate either a multivariate model for the noise (e.g. a VAR(1)), although I personally estimate an AR(1) for each component series and then use the average (i.e. setting the weights w_a in Andrews' article to 1) - this is the way the GMM program that I gave you is set up.

In my experience, the choice between (standard) k-functions matters little, while the choice of band-width (K_T) is important. I am not quite sure how much help the Andrews' formulae are in practice, but at least they have the big advantage that if you use a formula then the reader know you didn't data mine K_T.

Pre-whitening

Since the usual weighting scheme gives the autocorrelations less than full weight it is easy to see, in the situation where they are all positive, that the spectral density estimate is always biased downwards. Alternatively, remember that the spectral density estimate is a weighted average of the sample spectral density for neighboring frequencies, so if the sample spectral density is not "flat", the smoothed estimate is biased. Therefore Andrews and Monahan (1992) suggest the used of so-called "pre-whitened" spectral density estimators. The idea is simple (and not new - see the references in Andrews and Monahan) - if one can perform an invertible transformation that makes the sample spectrum flatter, then one should do that, then use the usual spectral density estimator, and finally undo the initial transformation. This may sound a little abstract but the way it is usually implemented is quite simple: Assume you have a series of "error" terms f_t and you suspect (say) strong positive autocorrelation. Then you may want to fit an VAR(1) model (the generalization to higher order VAR models is trivial) to the f_t terms and obtain residuals, which we will denote f_t^*, i.e.

$$f_t \;=\; \hat{A} f_{t-1} + f_t^* \ .$$

More specifically the process of finding the f_t^*s from the f_t is denoted pre-whitening. It is easy to see that in large samples this implies (approximately)

$$(I - \hat{A}) \frac{1}{T} \sum_1^T f_t \;=\; \frac{1}{T} \sum_1^T f_t^* \ ,$$

so we see that

$$Var\{\frac{1}{T} \sum_1^T f_t\} \;=\; (I - \hat{A})^{-1} Var\{\frac{1}{T} \sum_1^T f_t^*\} (I - \hat{A}')^{-1} \ ,$$

and to find your estimate of $Var\{\frac{1}{T} \sum_1^T f_t\}$ you find an estimate of $Var\{\frac{1}{T} \sum_1^T f_t^*\}$ and use this equality. This is denoted "re-coloring". The reason that this may result in less biased estimates is that f_t^* has less autocorrelation and therefore a flatter spectrum around 0. On the other hand the pre-whitening operation may add more noise and one would usually only use pre-whitening in the situation where strong positive auto-correlation is expected. Also be aware that in this situation the VAR estimation is not always well behaved and you may risk that $I - \hat{A}$ will be singular. Therefore Andrews suggests that one use a singular value decomposition of $\hat{A}$ and truncate all eigenvalues larger than .97 to .97 (and less than -.97 to -.97) - see Andrews and Monahan (1992) for the details.

Andrews and Monahan supply Monte Carlo evidence that shows that for the models they consider, pre-whitening results in a significant reduction in the bias, at the cost of an increase (sometimes a rather large increase) in the variance. In many applications you may worry more about bias than variance of your t-statistics, and pre-whitening may be preferred.

An alternative endogenous lag selection scheme [I won't ask questions about this].

In a recent paper Newey and West (1994) suggest another method of choosing the lag length endogenously. Remember that the optimal lag-length depends on

$$\alpha(q) \;=\; 2 \left(\frac{f^{(q)}}{f(0)} \right)^2 \ .$$

Newey and West suggest estimating $f^{(q)}$ by

$$\hat{f}^{(q)} = \frac{1}{2\pi} \sum_{r=-n}^{n} |r|^q c(r)$$

which you get by taking the definition and plugging in the estimated autocorrelations and truncating at n. Similarly they suggest

$$\hat{f}(0) = \frac{1}{2\pi} \sum_{r=-n}^{n} c(r) \ .$$

Note that this is actually the truncated estimator (which have all weights equal to unity for the first autocorrelations and 0 thereafter) of the spectral density that we want to estimate but they suggest only to use this estimate in order to get

$$\hat{\alpha}(q) \ = \ \left(\frac{2\hat{f}^{(q)}}{\hat{f}(0)} \right)^2 \ ,$$

and then proceed to find the actual spectral density estimator using a kernel which guarantees positive semi-definiteness. Newey and West show that one has to choose n of order less than $T^{2/9}$ for the Bartlett kernel and order less than $T^{2/25}$ for the QS kernel. Note that there still is an arbitrary constant (namely n) to be chosen, but one may expect that the Newey-West lag selection scheme will be superior to the Andrews scheme in very large samples, (if you let n grow with the sample) since it does not rely on an arbitrary approximating parametric model. In Newey and West (1994) they perform some Monte Carlo simulations, that show that their own lag selection procedure is superior to Andrews' but only marginally so. In the paper Andersen and Sørensen (1996) we do, however, find a stronger preference for the Newey-West lag selection scheme in a model with high autocorrelation and high kurtosis.

2 Theory Sketch

Now it is easy to show that

$$\int_{-\pi}^{\pi} cos(\lambda h) f(\lambda) d\lambda \ = \ \frac{1}{2}\gamma(h) \ ,$$

since $\int_{-\pi}^{\pi} cos(\lambda h) cos(\lambda j) d\lambda = \pi\delta_{hj}$ (where δ_{hj} is Kronecker's delta [1 for $h = j$, 0 otherwise]). You can easily see that the spectral density is flat (i.e. constant) if there is no autocorrelation at all, and that $f(\lambda)$ becomes very steep near 0, if all the autocovariances are large and positive (the latter is called the "typical spectral shape" for economic time series by Granger and Newbold). In any event, since we want to estimate only $f(0)$, this is the all the intuition you need about this.

The Sample Spectral Density

Define

$$I(\lambda) \;=\; \frac{1}{2\pi}\sum_{k=-T}^{T} c(k)cos(\lambda k) \;.$$

$I(\lambda)$ is that sample equivalent of the spectral density and is denoted the *sample spectral density.* It is fairly simple to show (you should do this !) that

$$I(\lambda) \;=\; \frac{1}{2\pi T}|\sum_{t=1}^{T} f_t e^{i\lambda t}|^2 \;.$$

The importance of this is that it shows that the sample spectral density is positive. We do not want spectral estimators that can be negative (or not positively semi-definite in the multivariate case).

Anderson (1971), p. 454 shows that

$$EI(0) = \int_{-\pi}^{\pi} k_T(\nu)f(\nu)d\nu \;,$$

where

$$k_T(\nu) \;=\; \frac{\sin^2\frac{1}{2}\nu T}{2\pi T\sin^2\frac{1}{2}\nu}$$

is called Fejer's kernel. Notice that the expected value is a weighted average of the values of $f(\lambda)$ in a neighborhood of 0. If the true spectral density is flat then the sample spectrum is unbiased but otherwise not in general. Anderson also shows (page 457) that if the process is normal then

$$Var(I(0)) = 2[E\{I(0)\}]^2$$

(for non-normal processes there will be a further contribution involving the 4th order cumulants).

If $\sum|\gamma(k)| < \infty$ then on can show that

$$\lim_{T\to\infty} EI(\lambda) = f(\lambda) \;,$$

and for normal processes on can show that

$$\lim_{T\to\infty} Var I(0) = 2f(0)^2 \ ,$$

(and again there is a further contribution from 4th order cumulants for non-normal processes).

One can also show that (for normal processes)

$$\lim_{T\to\infty} Cov\{I(\lambda)I(\nu)\} = 0 \ ,$$

for $\lambda \neq \nu$, so that the estimates for even neighboring λs are independent. This independence together with the asymptotic unbiasedness is the reason that one can obtain consistent estimates of the spectral density by "smoothing" the sample spectrum.

For a general (and extremely readable) introduction to smoothing and other aspects of density estimation (these methods are not specific for spectral densities), see B. Silverman: "Density Estimation for Statistics and Data Analysis", Chapman and Hall, 1986.

Consistent estimation of the spectral density

One can obtain consistent estimates of the spectral density function by using weights, i.e. for a sequence of weights w_j

$$\hat{f}(\gamma) = \frac{1}{\pi} \sum_{r=-T+1}^{T-1} \cos(\gamma r) w_r c(r) \ .$$

If you define

$$w^*(\lambda|\nu) = \frac{1}{\pi} \sum_{r=-T+1}^{T-1} \cos(\lambda r)\cos(\nu r) w_r \ ,$$

it is easy to see that

$$\hat{f}(\nu) = \int_{-\pi}^{\pi} w^*(\lambda|\nu) I(\lambda) d\lambda \ .$$

We will only use these formula's for $\nu = 0$, but the important thing to see is that our estimate of the spectral density is a *smoothed* estimate of the sample spectral density. Also note that the usual way to show that a set of weights result in a positive density estimate is to check that the implied $w^*(.|0)$ function is positive.

Anderson (page 521) shows that

$$\lim E\hat{f}(0) = \int_{-\pi}^{\pi} w^*(\lambda|0) f(\lambda) d\lambda \ .$$

This means that the kernel smoothed estimate is not in general consistent for a fixed set of weights. Of course if the true spectral density is constant the smoothed estimate will be consistent (since the weights will integrate to 1 in all weighting schemes you would actually use), but the more “steep” the actual spectral density is, the more bias you would get. We will show how one can obtain an asymptotically unbiased estimate of the spectral density by letting the weights be a function of T, but the above kind of bias is still what you would expect to find in finite samples, which is why it is worth keeping in mind.

For the asymptotic theory the smoothness of the function k near 0 is important, define k_q as

$$\lim_{x \to 0} \frac{1 - k(x)}{|x|^q} = k_q \ ,$$

where q is the largest exponent for which k_q is finite. Various ways of choosing the function k to generate the weights result in different values of q and k_q. Under regularity conditions (most importantly $\sum_{r=-\infty}^{\infty} |r|^q \gamma(k) < \infty$) you find that for $K_T \to \infty$ such that the q-th power grows slower than T, $K_T^q/T \to 0$, then

$$\lim K_T^q [E\hat{f}(\nu) - f(\nu)] = \frac{-k_q}{2\pi} \sum_{r=-\infty}^{\infty} |r|^q \cos(\nu r)\gamma(k) \ .$$

Note that this implies that the smoothed estimate is consistent, and the most important is the *rate* of convergence, which is faster the larger K_T^q (subject to being less than T). It is easy to verify that $q = 1$ for the Bartlett kernel, and $q = 2$ for most other kernel schemes used. For the variance one can show that

$$\lim_{T\to\infty} \frac{T}{K_T} var\{\hat{f}_T(0)\} = 2f^2(0) \int_{-1}^{1} k^2(x)dx$$

(for the estimate at points not equal to zero or π the factor 2 disappears - this is due to the fact that the spectral density is symmetric around 0, so at 0 a symmetric kernel will in essence smooth over only half as many observations of the sample spectral density). So we notice that the variance does not go to zero at the usual parametric rate $\frac{1}{T}$, but only at the slower rate K_T/T. So in order to get low variance you would like K_T to grow very slowly, but in order to obtain low bias you would like K_T to grow very fast. You can also see that asymptotically the kernel with higher values of q will totally dominate the ones with lower values of q since you for the same order of magnitude of the variance get a lower order of magnitude of the bias. In practice this may no be so relevant, however, since the parameter q only depends on the kernel near 0, which only really comes into play in extremely large samples.

The only kernels that allow for a q larger than 2 are kernels that do not necessarily give positive density estimates, which people tend to avoid (although Lars Hansen have used the truncated kernel, which belongs to those). Among the kernels that have $q = 2$ Andrews show that the optimal kernel is the one which minimizes $k_q^2(\int_{-1}^{1} k^2(x)dx)^4$. (See Andrews (1991), Theorem 2, p. 829). This turns out to minimized by the Quadratic Spectral (QS) kernel.

The usual way the bias and the variance is traded off is by minimizing the asymptotic Mean Square Error. For simplicity define

$$f^{(q)} = \frac{1}{2\pi} \sum_{r=-\infty}^{\infty} |r|^q \gamma(r) \ .$$

It is simple to show that the MSE is

$$\frac{K_T}{T} f^2(0) \int_{-1}^{1} k^2(x) dx \ + \ \left(\frac{1}{K_T^q} \right)^2 k_q^2 [f^{(q)}]^2$$

Now in order to minimize the MSE, differentiate with respect to K_T, set the resulting expression equal to 0, solve for K_T and obtain

$$K_T \ = \ \left(\frac{2q k_q^2 [f^{(q)}]^2}{f(0)^2 \int k^2} \right)^{\frac{1}{2q+1}} T^{\frac{1}{2q+1}}$$

For example for the Bartlett kernel you can find $k(0) = 1$ and $\int k^2 = 2/3$. Andrews define

$$\alpha(q) = \frac{2[f^{(q)}]^2}{f(0)^2}$$

and the optimal bandwidth

$$K_T^* = \left(\frac{q k_q^2}{\int k^2(x) dx} \right)^{\frac{1}{2q+1}} (\alpha(q) T)^{\frac{1}{2q+1}}$$

so you find

$$K_T^* = 1.1447 [\alpha(1) T]^{\frac{1}{3}}$$

for the Bartlett kernel, and

$$K_T^* = 1.3221 [\alpha(2) T]^{\frac{1}{5}}$$

for the QS kernel.