University of Houston
Fall 2016

Bent E. Sørensen

1 The Linear IV Case.

Generalized Method of Moment (GMM) estimation is one of two developments in econometrics in the 80ies that revolutionized empirical work in macroeconomics. (The other being the understanding of unit roots and cointegration.)

The path breaking articles on GMM were those of Hansen (1982) and Hansen and Singleton (1982). For introductions to GMM, Davidson and MacKinnon (1993) have comprehensive chapter on GMM and I recommend that you read the chapter on GMM in the Hamilton (1994) textbook. This is a good supplement to the teaching notes. For more comprehensive coverage see the recent textbook by Alastair Hall (Oxford University Press 2005).

I think that one can claim that there wasn't that much material in Hansen (1982) that was not already known to specialists, although the article definitely was not redundant, as it unified a large literature (almost every estimator you know can be shown to be a special case of GMM). The demonstration in Hansen and Singleton (1982), that the GMM method allowed for the estimation of non-linear rational expectations models, that could not be estimated by other methods, really catapulted Hansen and Singleton to major fame. We will start by reviewing linear instrumental variables estimation, since that will contain most of the ideas and intuition for the general GMM estimation.

1.1 Linear IV estimation

Consider the following simple model

$$(1) \quad y_t = x_t\theta \ + \ e_t \ , \ t \ = 1, ..., T$$

where y_t and e_t scalar, x_t is $1 \times K$ and θ is a $K \times 1$ vector of parameters. NOTE from the beginning that even though I use the index "t" — indicating time, that GMM methods are applicable, and indeed much used, in cross sectional studies.
In vector form the equation (1) can be written

$$(2) \quad Y = X\theta \ + \ E \ ,$$

in the usual fashion. If x_t and e_t may be correlated, one will obtain a **consistent** estimator by using instrumental variables (IV) estimation. The idea is to find a $1 \times L$ vector z_t that

is as highly correlated with x_t as possible and at the same time is independent of e_t — so if x_t is actually uncorrelated with e_t you will use x_t itself as instruments - in this way all the simple estimators that you know, like OLS, are special cases of IV- (and GMM-) estimation. If Z denotes the $T \times L$ ($K \geq L$) vector of the z-observations then we get by pre-multiplying (2) by Z that

$$(3) \quad Z'Y = Z'X\theta + Z'E \ .$$

If we now denote $Z'Y$ by $\tilde{Y}$, $Z'X$ by $\tilde{X}$, and $Z'E$ by U then the system has the form

$$\tilde{Y} \ = \ \tilde{X}\theta \ + \ U \ ,$$

which corresponds to a standard OLS formulation with L observations. Here the variance Ω of U is

$$\Omega \ = \ var(U) \ = \ Z'var(E)Z \ .$$

Now the standard OLS estimator of θ is

$$\hat{\theta} \ = \ (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y} \ ,$$

which is consistent and unbiased with variance

$$Var(\hat{\theta}) \ = \ (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\Omega\tilde{X}(\tilde{X}'\tilde{X})^{-1} \ .$$

For simplicity let us now consider drop the tilde's, and just remember that the system (of the form (2)) often will have been obtained via the use of instrumental variables. (Most of the GMM-literature uses very sparse notation, which is nice when you are familiar with it, but makes it hard to get started on).
If U does not have a variance matrix that is proportional to the identity matrix the OLS estimator is not efficient. Remember that the OLS estimator is chosen to minimize the criterion function

$$U'U \ = \ (Y - X\theta)'(Y - X\theta) \ .$$

To obtain a more **efficient** estimator than the OLS estimator we have to give different weights to the different equations. Assume that we have given a **weighting matrix** W (the choice of weighting matrices is an important subject that we will return to) and instead choose $\hat{\theta}$ to minimize

$$U'WU \ = \ (Y - X\theta)'W(Y - X\theta) \ ,$$

or (in the typical compact notation)

$$\hat{\theta} \ = \ argmin_{\theta}U'WU \ .$$

In this linear case one can then easily show that $\hat{\theta}$ is the GLS-estimator

$$\hat{\theta} \ = \ (X'WX)^{-1}X'WY \ .$$

Let the variance of U be denoted Ω and we find that $\hat{\theta}$ have variance

$$var((X'WX)^{-1}X'WU \;=\; (X'WX)^{-1}X'W\Omega WX(X'WX)^{-1} \;.$$

We want to choose the weighting matrix optimally, so as to achieve the lowest variance of the estimator. It is fairly obvious that one will get the most efficient estimator by weighing each equation by the inverse of its standard deviation which suggests choosing the weighting matrix Ω^{-1}. In this case we find by substituting Ω^{-1} for W in the previous equation that

$$var((X'\Omega^{-1}X)^{-1}X'\Omega^{-1}U \;=\; (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\Omega\Omega^{-1}X(X'\Omega^{-1}X)^{-1} \;=\; (X'\Omega^{-1}X)^{-1} \;.$$

We recognize this as the variance of the GLS estimator. Since we know that the GLS estimator is the most efficient estimator it must be the case that Ω^{-1} is the optimal weighting matrix.
For practical purposes one would usually have to do a 2-step estimation. First perform a preliminary estimation by OLS (for example), then estimate Ω (from the residuals), and perform a second step using this estimate of Ω to perform "feasible GLS". This is asymptotically fully efficient. It sometimes can improve finite sample performance to iterate one step more in order to get a better estimate of the weighting matrix (one may also iterate to joint convergence over Ω and θ — there is some Monte Carlo evidence that this is optimal in small samples).

A special case is the IV estimator (see eq. (3)). If $var(E) = I$, then the variance of $Z'E$ is $Z'Z$. The optimal GMM-estimator is then

$$\hat{\theta} \;=\; (\tilde{X}'(Z'Z)^{-1}\tilde{X})^{-1}\tilde{X}'(Z'Z)^{-1}\tilde{Y} \;,$$

or

$$\hat{\theta} \;=\; (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'Y \;.$$

It is now easy to check that this is the OLS-estimator, when you regress $Z(Z'Z)^{-1}Z'Y$ on $Z(Z'Z)^{-1}Z'X$. This is the classical IV-estimator, which is referred to as the Two-Stage Least Squares in the context of simultaneous equation estimation. The "first stage" is an OLS-regression on the instrument and the "second stage" is the OLS-regression of the fitted values from the first stage regression.

The derivations above illustrate many of the concepts of GMM. Personally I always guide my intuition by the GLS model. For the general GMM estimators the formulas look just the same (in particular the formulas for the variance) except that if we consider the nonlinear estimation

$$(4)\quad Y = h(X,\theta) \;+\; U \;,\;,$$

then "X" in the GLS-formulas should be changed to $\frac{\partial h}{\partial \theta}$. E.g. using the optimal weighting matrix (much more about that later), you find the *asymptotic* variance of the estimated parameter to be

$$var(\hat{\theta}) = (\frac{\partial h'}{\partial \theta}\Omega^{-1}\frac{\partial h}{\partial \theta})^{-1}$$

In GMM jargon the model would usually be formulated as

$$U \;=\; Y - h(X,\theta)\ ,$$

or more often as

$$(**)\;\; U \;=\; f(\tilde{X},\theta)\ ,$$

(where $\tilde{X} = Y, X$ and $f(\tilde{X},\theta) = Y - X\theta$. The later—very compact—notation is the one that is commonly used in the GMM literature and we will follow it here. We again drop the tilde and denote all the variables by X. It is typical for the newer methods (typically inspired from statistics) that the variables are treated symmetrically.
In the language of GMM the whole model is summarized by L **orthogonality conditions**:

$$EU \;=\; 0\ ,$$

or (when you want to be really explicit!):

$$EU(X,\theta) \;=\; 0\ .$$

Here you should think of U as being a theoretical model. It is not quite explicit here whether we think of U as equations that have been pre-multiplied by instrument vectors or not. But in the usual formulation of GMM the dimension L of U is fixed, so e.g. in the OLS model where the dimension of $E = \{e_1, ..., e_T\}'$ depends on T, you would think of the orthogonality conditions as being $U = X'Y - X'X\theta$. In rational expectations models, the theory often implies which variables will be valid instruments; but this is not always so. For the statistical development the terse notation is good; but in applications you will of course have to be more explicit.

2 GMM and Method of Moments

If we have L orthogonality conditions summarized in a vector function $f(X,\theta)$ that satisfies $Ef(X,\theta) = 0$, the GMM estimator attempts to minimize a quadratic form in f, namely $f'Wf$. Notice that there are L orthogonality conditions (rather than T) – this means that you should think about $Z'(Y - X\theta)$ in the IV setting [rather than $(Y - X\theta)$]. Assume that Z is just columns of ones. Then a relation like $f(X,\theta) = Z'g(X,\theta)$ is just $g_T(X,\theta) =$

$\frac{1}{T}\Sigma_{t=1}^{T} g_t(X,\theta)$. In other words the orthogonality condition is the first empirical moment of the g_t vector. In the case of instruments z_t the orthogonality condition is really $g_T(X,\theta) = \frac{1}{T}\Sigma z_t g_t(X,\theta)$. If the number of orthogonality conditions is the same as the number of parameters you can solve for the θ vector which makes $g_T = 0$ – in this case the weighting matrix does not matter. This does not mean that the method is only applicable for first moments, for example you could have

$$u_t = \begin{pmatrix} x_t - \mu \\ x_t^2 - \sigma^2 - \mu^2 \end{pmatrix} ,$$

which, for a vector of constants as the instruments, corresponds to simple method of moments. More generally, a model often implies that the moments is some non-linear functions of the parameters, and those can then be found by matching the empirical moments with the models implied by the model. (The moments used for the GMM-estimator in Melino-Turnbull (1990) and Ho, Perraudin, and Sørensen (1996) are simply matching of moments). The "Generalized" in GMM comes from the fact that we allow more moments than parameters and that we allow for instruments. Sometimes GMM theory will be discussed as GIVE (Generalized Instrumental Variables Estimation), although this is usually in the case of linear models.

3 Hansen and Singleton's 1982 model

This is by now the canonical example that "everybody" knows and which popularized the approach.
The model in Hansen and Singleton (1982) is a simple non-linear rational expectations representative agent model for the demand for financial assets. The model is a simple version of the model of Lucas (1978), and here the model is simplified even more in order to highlight the structure. Note that the considerations below are very typical for implementations of non linear rational expectations models.

We consider an agent that maximize a time-separable von Neumann-Morgenstern utility function over an infinite time horizon. In each period the consumer has to choose between consuming or investing. It is assumed that the consumers utility index is of the constant relative risk aversion (CRRA) type. There is only one consumption good (as in Hansen and Singleton) and one asset (a simplification here).
The consumers problem is

$$\text{Max } E_t [\sum_{j=0}^{\infty} \beta^j \frac{1}{\gamma} C_{t+j}^{\gamma}]$$

$$\text{s.t. } C_{t+j} + I_{t+j} \leq r_{t+j} I_{t+j-1} + W_{t+j} ; \quad j = 0, 1, .., \infty$$

where E_t is the consumer's expectations at time t and

C_t	:	Consumption
I_t	:	Investment in (one-period) asset
W_t	:	Other Income
r_t	:	Rate of Return
β	:	Discount Factor
γ	:	Parameter of Utility Function

If you knew how C_t and I_t was determined this model could be used to find r_t (which is why it called an asset pricing model), but here we will consider this optimization problem as if it was part of a larger unknown system. Hansen and Singleton's purpose was to estimate the unknown parameters (β and γ), and to test the model.

The first order conditions (called the "Euler equation") for maximum in the model is that

$$C_t^{\gamma-1} = \beta E_t[C_{t+1}^{\gamma-1} r_{t+1}] \ .$$

The model can not be solved for the optimal consumption path and the major insight of Hansen and Singleton (1982) was that knowledge of the Euler equations are sufficient for estimating the model.
The assumption of rational expectations is critical here - if we assume that the agents expectations at time t (as expressed through E_t corresponds to the true expectations as derived from the probability measure that describes that actual evolution of the variables then the Euler equation can be used to form the "orthogonality condition"

$$U(C_t, \theta) \ = \ \beta C_{t+1}^{\gamma-1} r_{t+1} \ - \ C_t^{\gamma-1} \ ,$$

where $E_t U = 0$ (why?), where we now interpret E_t as the "objective" or "true" conditional expectation. Note that $E_t U = 0$ implies that $EU = 0$ by the "law of iterated expectations", which is all that is needed in order to estimate the parameters by GMM. The fact that the *conditional* expectation of U is equal to zero can be quite useful for the purpose of selecting instruments. In the Hansen-Singleton model we have one orthogonality condition and that is not enough in order to estimate two parameters (more about that shortly), but if we can find two or more independent instrumental variables to use as instruments then we effectively have more than 2 orthogonality conditions.
We denote the agents information set at time t by Ω_t. Ω_t will typically be a set of previous observations of economic variables $\{z_{1t}, z_{1t-1}, ...; z_{2t}, z_{2t-1}, ...; z_{Kt}, z_{Kt-1}, ...\}$. (Including C_t, and I_t among the z's. Then any variable in Ω_t will be a valid instrument in the sense that

$$E[z_t U(C_t, \theta)] \ = \ 0$$

for any z_t in Ω_t. Notice that z_t here denotes any valid instrument at time t, for example z_t could be z_{1t-3} - this convention indexing the instruments will prove quite convenient. The $E[.,.]$ operation can be considered an inner product, so this equation is really the origin of the term orthogonality conditions. For those of you who want to see how this can be developed rigorously, see the book by Hansen and Sargent (1991).
Take a few seconds to appreciate how elegant it all fits together. Economic theory gives you the first order condition directly, then you need instruments, but again they are delivered by the model. For empirical economists who want to derive estimation equations from economic principles, it does not get any better that this.
Oh, well maybe there is a trade-off. The reason being that instrumental variables estimators not are very efficient if no good instruments are available (there is active research in this area at the present, see paper with the words "weak instruments"); but for now you may want to compare the Hansen-Singleton (1982) approach to the article "Stochastic Consumption, Risk Aversion, and the Temporal Behavior of Asset Returns", JPE, 93, p 249-265. This is really the same model, but with enough assumptions imposed that the model can be estimated by Maximum Likelihood.

4 Non-Linear GMM.

Assume that economic theory gives us the moment conditions

$$Ef_t(\theta) = 0 \ ,$$

where $f_t(\theta) = f(x_t, \theta)$ is an r dimensional vector of moment conditions and θ is a q dimensional vector of parameters. The *identification* condition is that $Ef_t = 0$ for $\theta = \theta_0$ and otherwise not. (Further you need to assume a compact parameter space or some equivalent assumption as outlined in class.)

Define

$$g_T \ = \ \frac{1}{T}\sum_{t=1}^{T} f_t \ .$$

We will use the notation g_T or $g_T(\theta)$, but from now on the dependence of g_T on the underlying series, x_t, will be implicit. The GMM estimator will be the estimator that makes $g_T(\theta)$ as close to zero as possible. Notice that g_T is the empirical first moment of the series f_t which is why the estimator is called a moment estimator. Also note that the standard idea of moment estimation, which consists of equating as well as possible a series of moments. This would be achieved by choosing $g_T' = [\overline{x}_T - Ex_t, \overline{x^2}_T - E\{x_t^2\}, ..., \overline{x^K}_T - E\{x_t^K\}]$.
We now define the **GMM-estimator** as

$$\hat{\theta} \ = \ \text{argmin}_\theta \ g_T' W_T g_T \ ,$$

where W_T is a weighting matrix that (typically) depends on T such that there exist a positive definite matrix W_0, such that $W_T \to W_0$ $(a.s.)$. The latter condition allows us to let the weighting matrix be dependent on an initial consistent estimator, which is very important since the optimal GMM estimator will be a two step estimator, just as in the GLS-case above.
Let $Dg_T(\theta)$ be the $r \times q$ dimensional matrix of derivatives with typical element $Dg_{Tij} = \frac{\partial g_{Ti}}{\partial \theta_j}$. We will assume Dg_T has full rank. When the underlying data follows continuous distributions this will usually follow with probability 1 from the identification condition. (Below I will often just write Dg in order to simplify notation but, of course, all functions will be evaluated using the T available observations.)

Then the first order condition of the optimization becomes

$$Dg_T(\hat{\theta})' W_T \, g_T(\hat{\theta}) \;=\; 0 \; .$$

Solving non-linear optimization by the Newton algorithms
GAUSS and other programs use a Newton type algorithm to solve non-linear optimization problems. There are many variations of this but most variations involve approximations to how one finds derivatives and things like that. The computer will find the derivative of the criterion function numerically but you will have the option to let a subroutine calculate it if you have an analytical expression, this will often increase computational speeds significantly if the number of parameters is high.

Newton type algorithms work by starting from an initial value θ_0 and for a given value θ_{N-1} finding θ_N which minimize the linearized criterion function:

$$[g_T(\theta_{N-1}) + Dg(\theta_N - \theta_{N-1})]' W_T [g_T(\theta_{N-1}) + Dg(\theta_N - \theta_{N-1})]$$

The solution is (check this!)

$$\theta_N - \theta_{N-1} \;=\; -(Dg' W_T Dg)^{-1} Dg' W_T g_T(\theta_{N-1}) \; ,$$

which is the NEWTON upgrade.

4.1 Asymptotic theory

We will assume that the series $(x_t', z_t')'$ is **ergodic**. A series x_t is ergodic if

$$\frac{1}{T} \sum_{t=1}^{T} h(x_t) \;\to Eh(x_t)$$

for all functions h(.) (for which the mean is well defined). Notice that the right hand side of the above equation is assumed to not be a function of t. It is, more or less, impossible to test

if a series is ergodic. However, it is well known that an *integrated* time-series (e.g., a random walk) is not ergodic. Most macroeconomic series are integrated, or nearly integrated, time series, but most often the model can be rewritten in terms of stationary variables (typically growth rates).

For proof of consistency, see for example, Hansen (1982). The idea is simple enough. When T is large the function $g_T(\theta)$ is close to $Ef_t(\theta)$ and the minimum of g_T will therefore be close to the minimum of Ef_t, i.e., close to θ_0. In order to make these statements precise we need to be specific about what we mean by convergence of *functions* but I will leave this for more specialized econometrics courses.

We will also assume that the series $f_t(\theta)$ satisfies a central limit theorem, i.e. that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} f_t(\theta) \;\Rightarrow\; \mathrm{N}(0,\Omega) \;,$$

where $\Omega = E[f_t f_t']$ if f_t is not autocorrelated, but in general

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

So intuitively, where we in the GLS model had T (or L in the IV case) normally distributed error terms, we here have K asymptotically normally distributed moment (or orthogonality) conditions.

Let $Df = E\frac{\partial f_t}{\partial \theta}(\theta_0)$. One can then show that for any convergent sequence of weighting matrices the GMM-estimator is consistent and asymptotically normal with

$$\sqrt{T}(\hat{\theta} - \theta \;) \Rightarrow \; \mathrm{N}(0,\Sigma) \;,$$

where

$$\Sigma \;=\; (Df' W_0 Df)^{-1} Df' W_0 \Omega W_0 Df (Df' W_0 Df)^{-1} \;.$$

Notice that this formula corresponds exactly to the one obtained in the linear case if you substitute X for Df.

We can sketch the proof (see Hall's book for more detail): By the mean value theorem we can write

$$g_T(\hat{\theta}) = g_T(\theta_0) + Dg_T(\overline{\theta})(\hat{\theta} - \theta_0)$$

Now, pre-multiply this equation by $Dg_T(\hat{\theta})' W_T$ and, by the first-order condition above, the left-hand side is 0 and we get

$$0 \;=\; Dg_T(\hat{\theta})' W_T g_T(\theta_0) + Dg_T(\hat{\theta})' W_T Dg_T(\overline{\theta}) \, (\hat{\theta} - \theta_0) \;,$$

where $|\overline{\theta} - \theta_0| < |\hat{\theta} - \theta_0|$ (which implies that when $\hat{\theta}$ is consistent and converges to θ_0, so will $\overline{\theta}$). We get

$$\hat{\theta} - \theta_0 \;=\; -[Dg_T(\hat{\theta})' W_T Dg_T(\overline{\theta})]^{-1} Dg_T(\hat{\theta})' W_T g_T(\theta_0) \;.$$

This implies

$$\sqrt{T}(\hat{\theta} - \theta_0) \;=\; -[Dg_T(\hat{\theta})' W_T Dg_T(\overline{\theta})]^{-1} Dg_T(\hat{\theta})' W_T \sqrt{T} g_T(\theta_0)$$

This has the form

$$A_T \, \sqrt{T} g_T(\theta_0) \;,$$

where $\sqrt{T} g_T(\theta_0)$ converges in distribution to $\mathrm{N}(0, \Omega)$ and A_T converges in probability to $(Df' W_0 Df)^{-1} Df' W_0$ which gives us the formula above. (Dg converges to Df by ergodicity.) Again, you need to look at more specialized articles to make sure this is all kosher but, in general, the assumption that econometric theorists impose to prove the theorems are rarely of such a form that practitioners can verify them.

The reasoning behind the GLS estimator also carries over and the *optimal GMM-estimator* is the one where $W_T \to \Omega^{-1}$ in which case the asymptotic covariance of the GMM-estimator is

$$\Sigma_0 \;=\; (Df' \Omega^{-1} Df)^{-1} \;.$$

In order to obtain an estimate $\hat{\Sigma}_0$ you need an estimate $\hat{\Omega}$ and then you use

$$\hat{\Sigma}_0 = (Dg' \hat{\Omega}^{-1} Dg)^{-1} \;,$$

where Dg, of course, is evaluated at $\hat{\theta}$.

Notice that this is the optimal estimator for a **given set of instruments**. The problem of finding the best instruments is much harder and no satisfactory solution exists to that problem in general (although often for special cases, like the OLS model). I will comment on this (very important) issue below.

Also notice, that consistency is found without making assumption on the error terms and without specifying the model such that the error terms are independent. Sometimes authors will claim that this is a big strength of GMM, but if you error are not approximately normal you will often have problems and, in particular, if you have a lot of autocorrelation in your residuals you will not get very precise estimates. (The profession seems to go in circles as to whether it is consider a strength to not have to make distributional assumptions ["OLS is the Best Linear Unbiased estimator" or the opposite "OLS is the ML estimator"]. Some people can get "religious" about this issue, but my more pragmatic attitude is that the important thing is to get error terms that are approximately uncorrelated.)

4.2 Hypothesis Testing in a GMM-framework

There exists equivalents of the standard Wald-, LM-, and ML-test in the case of GMM estimation. Note: This is only true in the case where the optimal weighting matrix has been applied. In a case where you apply a non-optimal weighting matrix then there is no equivalent of the ML-test available. (Ho, Perraudin, and Sørensen (1996) is an example of a paper that applies a non-optimal weighting matrix).

Consider a test for s nonlinear restrictions

$$R(\theta) \;=\; 0 \;,$$

where R is an $s \times 1$ vector of functions.

Let DR be $\frac{dR}{d\theta}$ (and we assume that DR is evaluated at the optimal GMM-estimator in the unrestricted model), then the **Wald test** is

$$TR(\theta)'[DR\hat{\Sigma}DR']^{-1}R(\theta) \;,$$

or

$$TR'[DR(Dg'W_TDg)^{-1}Dg'W_T\hat{\Omega}_T^{-1}W_TDg(Dg'W_TDg)^{-1}DR']^{-1}R \;,$$

where W_T is the weighting matrix and Dg is the derivative of g_T with respect to the parameters. Here Dg (and DR if this is dependent on the parameters) are evaluated at the unrestricted estimator of θ. Let us define

$$\hat{\Sigma} \;=\; (Dg'W_TDg)^{-1}\,Dg'W_T\hat{\Omega}_T^{-1}W_TDg\,(Dg'W_TDg)^{-1} \;,$$

In the formula for the Wald-test $\hat{\Sigma}$ is our estimator of the variance of $\hat{\theta}$ and when we pre- and post-multiply this by DR we get an estimate of the asymptotic variance of $R(\hat{\theta})$.

The LM-test can be implemented in different ways. I strongly recommend you check with a trusted source (like the article in the handbook (Newey and McFadden: Large Sample Estimation and Hypothesis Testing. In Handbook of Econometrics IV, eds. Engle and McFadden, North-Holland, (1994) or Gallant (1987)). For example, there is a formula in Ogaki (1992,) that I cannot quite get to agree with Gallant's formula and a much simpler looking formula in Davidson and MacKinnon (1993 [their older most advanced book]), that I cannot see how they get. They may be OK, but I recommend you be careful. The formula given here should agree with Gallant (1987). This version has the form

$$LM \;=\; Tg_T'W_TDg(Dg'W_TDg)^{-1}DR'(DR\hat{\Sigma}DR')^{-1}DR(Dg'W_TDg)^{-1}Dg'W_Tg_T$$

where Dg and G_T are evaluated at $\hat{\theta}$.

One way to motivate this version of the LM test is notice that if the restriction $R(\theta)$ is true the $DR(\hat{\theta})d\theta$ (evaluated at the restricted estimator) should be approximately zero where $\hat{\theta}$ is evaluated at the constrained minimum. The idea (of this version of the LM-test) is that you choose $d\theta$ as the update in a NEWTON algorithm, i.e.,

$$\theta_N - \hat{\theta} \;=\; (Dg' W_T Dg)^{-1} Dg' W_T g_T(\hat{\theta}) \;,$$

The idea of the LM test is that if the model fits well, the NEWTON step away from the restricted parameter value will be small or, at least, orthogonal to DR. Now you find the LM test-statistic by evaluating

$$[DR(\theta_N - \hat{\theta})]' V^{-1} DR(\theta_N - \hat{\theta}) \;,$$

where $V = DR\hat{\Sigma}DR'$ is the variance of $DR(\theta - \hat{\theta})$ (ignoring the small sample variance in DR), and

$$DR(\theta_N - \hat{\theta}) \;=\; DR(Dg' W_T Dg)^{-1} Dg' W_T g_T(\theta_{N-1}) \;,$$

using the expression for the Newton-step found above.

Finally the **LR-test** (of course it should strictly speaking be "LR-type test" for Likelihood-Ratio type) is

$$LR \;=\; 2 * T[J_T(\theta_2^r) - J_T(\theta_2^u)] \;,$$

where J_T is the objective function **(NB)** evaluated at the *optimal* weighting matrix and where the superscripts u and r of course indicates that the estimators were found in the unrestricted and the restricted models respectively.

The Wald-, LM-, and LR-test can all be shown to converge in distribution to a χ^2-distribution with s (number of restrictions) degrees of freedom in the case where the restrictions are true.

Hansen (1982) suggested the following **test for mis-specification**: Consider

$$J_T \;=\; T g_T(\hat{\theta}_2)' \hat{\Omega}_T^{-1} g_T(\hat{\theta}_2) \;.$$

If the model is correctly specified this statistic is asymptotically χ^2 distributed with degrees of freedom equal to $r - q$, where q is the number of parameters estimated. So a value that is far out in the tail indicates that the whole model is mis-specified. By the whole model I do not mean that *all* parts of the model are mis-specified; but rather that *some* part of the model is mis-specified - it could be that it was just the instruments that were not pre-determined. This test is known as the **test for overidentifying restrictions** or sometimes as the "Hansen J-test".
Note that you cannot test unless you have more moment conditions than parameters (an

"overidentified model"), in the case the model is exactly identified the J_T will be identically 0.
In Hansen and Singleton (1982) the model was rejected by the J-test, and my subjective impression is that from then on it became acceptable for a while to present an econometric estimation that rejected the model, as one that accepted the model. (This is the "scientific method" that Summers reject for macroeconomics. It seems that Summers won in that dimension, because at present it is basically impossible to publish an article that rejects the model.)

I often find the J-test useless. Models are never exactly true so the result of the J-test will usually be that it accepts the model (due to lack of power) if the number of observations is low, and rejects the model if the number of observations is high.

Simulated GMM
You may sometimes be in the situation where you cannot find an analytic expression for f_t. However, you might be able to *simulate* f_t. It is most easily explained by an example. Consider, for example, an $MA(2)$ process

$$y_t \;=\; u_t + b_1\, u_{t-1} + b_2 u_{t-2}\ , \quad (*)$$

where the error terms are $N(0,\sigma^2)$ distributed. The parameter vector here is $\theta' = \{b_1, b_2, \sigma^2\}$. For this model, I would actually use the Kalman-filter to evaluate the likelihood function for this particular model, but this is just an example, so imagine I couldn't find the likelihood function or the conditional likelihood function. Then I might simulate some moments for the y_t process. For example, I might use a random number generator to draw $N = 100,000$ observations $u_1, ..., u_N$ and calculate $y_3, ..., y_N$ using equation (*). I would set this up as a subroutine named, e.g., $SIM(\theta)$ in GAUSS [meaning that you call the subroutine SIM as a function of a given set of parameters]. Then, in the same subroutine I could calculate, say, $m_1(\theta) = \Sigma_{i=3}^N y_i$, $m_2 = \Sigma_{i=3}^N y_i^2$ $m_3 = \Sigma_{i=3}^t y_i^4$. Note that we would have to call the routine for a given parameter vector $\theta = \{b_1, b_2, \sigma^2\}$. Assume now that x_t, $t = 1, ..., T$ is your actual data. Now you would define $g_T' = \{m_1 - \Sigma_{t=1}^T x_t, m_2 - \Sigma_{t=1}^T x_t^2, m_4 - \Sigma_{t=1}^T x_t^4\}$. and you would minimize

$$\hat{\theta} \;=\; \text{argmin}_\theta\ g_T' W_T g_T\ .$$

This would give you a consistent estimate of θ. Note that this might be slow because for each step θ_N in the Newton algorithm, you need to call $SIM(\theta_N)$ in order to calculate the moments. As a matter of fact, for this model this would not be a problem since a modern computer can do this very quickly. In principle, *any* model that can be simulated (which more or less is the universe of models can be put into the SIM routine and some moments returned). In practice, you would have trouble with a large General Equilibrium (GE) model— GE models typically would need to be simulated which means that you would add

a layer of non-linear simulations for each θ_N...but as computers get faster you might be able to do it for a small GE model if you program cleverly. (Since there would be billions of calculations they better be streamlined.)

You need to choose N much larger than T—otherwise you need to take the extra variance that comes from simulating the moments into account when calculating std. errors.

Choice of moments. What matter much more for efficiency than the choice of weighting matrix is the choice of moments. In the case, as in the Hansen-Singleton model, where "choice of moments" means "choice of instruments" theory give little guidance. You can show that as T gets larger you should use more instruments, but in practice you have one T and you have to use common sense (use instruments that are not to correlated, don't use too many, ...). In the case, such as the MA(2) example, where you actually choose moments, you can more or less guess which moments will be good. For example, the ones I chose above, were pretty bad. An MA(2) model is characterized by non-zero first and second autocorrelations and higher order correlations being zero. So good moments would be the empirical variance, first, second, third, and maybe fourth order autocorrelations, rather than the higher moments I chose above.
It is possible to be quite systematic about this. Gallant and Tauchen suggested a method called **Efficient Method of Moments** (EMM) that can be used if you have a model with a likelihood function that you cannot write down such as a stochastic volatility model but you have a model that captures similar features of the data such as a GARCH model, you can actually estimate the GARCH model even if it is misspecified and then use the first derivatives of the likelihood function as the moment conditions. More precisely, if $s(\theta, x)$ is the score function (the derivative of the likelihood function) as a function of the data, you use simulated method of moments to match this to the model, but drawing a long series of observations y_i and calculate $s(\theta, y)$ and then your moment conditions are $g_T = s(\theta, x) - s(\theta, y)$. For the particular models that I mentioned this turns out to actually work very well—see the very comprehensive Monte Carlo papers by Andersen and Sorensen (1996) and Andersen, Chung, and Sorensen (1999). I also have some hand-written notes on EMM that you can have if you are interested.

5 Variance estimation.

Most of the underlying math in this note builds on Anderson (1971), chapters 8 and 9. [This book is now available in the Wiley Classics series].

Note that this material in principle is independent of GMM estimation. For example,

Phillips and Perron has a, quite well known, unit root test that uses the material below. Because Hansen and Singleton in the famous paper used this type of variance estimation it is sometimes (as in the Davidson-MacKinnon textbook) treated as part of GMM estimation from the beginning, but in my view that is making a confusing soup out of two different issues. 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 $\gamma(k) = Ef_t f_{t-j}$, 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

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

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

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

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.

6 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

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

(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

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

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.