Specification of Linear and Nonlinear Models Methodological Implications

C.W.J. Granger

Department of Economics, University of California - San Diego, La Jolla, CA 92093-0508, USA (cgranger@ucsd.edu)

Abstract: The methodology of the construction of linear time series models is now quite understood and accepted, but much of it cannot be used with nonlinear models. The main reason is that there is no natural nesting available for many nonlinear models. For some parametric nonlinear classes forms of nesting is possible and a "best of class" or a few good models can be found. To interpret the results an approach called "thick modeling" is suggested. There are several problems with the foundations that are discussed such as definition of trends, linearity, and of integratedness in nonlinear processes. Emphasis is given to purpose and evaluation of models. An appendix reviews that practical question of how to check relative abilities of alternative techniques to forecast volatility of stock market returns. The problems with a meta-analysis are illustrated.

Keywords: Methodology; Nonlinear Model; Evaluation; Volatility Forecasting

1. INTRODUCTION - LINEAR MODELS

This paper will turn out to be rather convoluted. It will start with a fairly simple task of comparing the methodologies used with the specification of linear time series models with those used with nonlinear models. It will generally not consider problems of estimation or testing for such models although there are many important questions in these areas. There is quite enough to discuss about model specification, selection, and evaluation. Initially, a researcher will be thought to start with an appropriate data set, possibly some theoretical models based on economic reasoning and will eventually finish with a final model, or small group of models. The modeler will be thought of as a "producer" and the user of the output of the models as a "consumer." Typically, the producer and the consumer are not the same person.

The easiest place to start a consideration of the methodology of time series is with the univariate, stationary case. For the start, it will assumed that the series has no deterministic components, such as seasonal or trend, and has a zero mean. In this case the appropriate class of models to consider are autoregressive, moving average (ARMA (p, q)) given as

$$X_t = \sum_{j=1}^{p} a_j X_{t-j} + \sum_{j=0}^{q} c_j \varepsilon_{t-j} \qquad (1)$$

where ε_t is a white noise series having covariance $(\varepsilon_t, \varepsilon_q) = 0$, t ? q. Note that p,q are always non-negative integers. It may be noted that for a linear model, one does not require stronger conditions on the "shocks" or "innovations" ε_t, such as ε_t are iid or Gaussian.

There is a useful nesting property of these models as ARMA (p_0, q_0) is within ARMA (p, q) if $p_0 \leq p$ and $q_0 \leq q$. This leads to two possible modeling strategies. This first is that suggest by Box and Jenkins [1970] in which various statistical "tests" and diagrams were used to chose an appropriate pair of stationary variables, denoted p_1, q_1 and the models are tested against neighboring models in size, ARMA (p_1, q_1) against ARMA$(p_1 + 1, q)$ and ARMA (p_1, q_1) against ARMA $(p_1, q_1 + 1)$. The tests could be in terms of goodness of fit or of forecastability, or any pre-determined criterion. If the first comparison is in favor of the simpler model, but the second comparison favors the more complicated one, the one, then further simple ? more complicated tests occur, with a single step in the p or q dimension taken at a time. Sometimes one can end with two or more models with different specifications which cannot be differentiated by testing. Often manipulation of the lag polynomials shows these models to be nearly identical. Later I will argue that having several models is not a bad outcome. The Box-Jenkins approach, which proved to be successful, is an example of "Simple ? General."

The "General ? Simple" methodology advocated by Hendry [1995] starts with an initial model ARMA (P, Q), with greater lags than are likely to really occur, and then whittles it down by testing and decision making, to a single (or small group of) simpler models ARMA (p_0, q_0). For example, one can drop lags that have coefficients which have t-values that are insignificant. I shall return to discussion of this general to simple procedure later in connection with larger models. I have not seen a comparison of the G ? S with the S ? G for univariate models, but would expect the G ? S to produce the more complicated models. However, it does depend on the model selection criterion used, AIC versus BIC. For AR(1) models AIC produces much larger p values than BIC (a medium of 8 from AIC, of 2 for BIC using actual U.S. macro monthly series).

Turning to the multivariate, stationary process, with no deterministic components and all means zero, the vector autoregressive model with p lags for all variables in all equations, VAR(p) is the standard model and the workhorse in the area. Theoretically the autoregressive moving average VARMA (p, q) is more likely to occur, but originally was more difficult to specify and was very difficult to estimate. These days the VARMA (p, q) model is farily straight-forward to use but not easy to interpret, so the VAR remains the favorite. If it is written in the form

$$\underline{A}_0 \underline{X}_t = \underline{A}_p(B)\underline{X}_{t-1} + \underline{\varepsilon}_t \qquad (2)$$

where $\underline{X}_t$ is a vector with m components, B is the backward operator (L is used in many econometrics books), $\underline{A}_p(B)$ is an $m \times n$ matrix where each element is a polynomial of order p in B, the lag operator, and $\underline{\varepsilon}_t$ is a white noise vector, with zero mean, so that $\varepsilon[\varepsilon_t \varepsilon_s'] = 0$, t ? s. Virtually every vector of economic variables will be contemporaneously correlated. This can be captured either by the values taken by the components of $\underline{A}_0$ or through the covariance matrix $\mathrm{cov}[\varepsilon_t \varepsilon_t'] = \underline{\Omega}$. The problem of how to best deal with contemporaneous relations, and how to interpret them, has been a difficult area since multivariate linear models were first considered. If one just stays with the information in the data from the m components in $\underline{X}_t$ there is no unique solution, merely alternative representations. By adding further information, or opinions, the full model can be identified although other people may not accept the model.

If the size and content of $\underline{X}_t$ is agreed and fixed, the methodology of VAR's is, at one level, straight-forward. If a "reduced-form" VAR(p) with $\underline{A}_0 \equiv \underline{I}$ is considered, then the set of alternatives are nested with changing values if p. However, if a "structural-model" form is used, with $\underline{A}_0$ not the unit matrix, then discussion about the number of zero values in this matrix and their location, can occur. For example, a "Wold causality" form can take $\underline{A}_0$ to be triangular, with zeros above the main diagonal. Adding further possible zeros, or reduced ranks in sub-matrices, in the $\underline{A}_p(B)$ matrix in (2) opens up the "identification" debate, for instance. These are old topics that I do not want to discuss. The real uncertainty in the VAR methodology is what variables to include in $\underline{X}_t$, how large should m be, which alternative definitions of inflation or unemployment or interest rates to include, for example.

Further methodological debate has taken place on a single equation taken from a multivariate situation. A typical equation of this kind can be written as

$$Y_t = \mu + \alpha(B)Y_{t-1} + \sum_{j=1}^{R} a_j(B)X_{jt} + \varepsilon_{jt} \qquad (3)$$

where $\alpha(B)$, $\alpha_j(B)$ are all polynomials in B of some predetermined order, p for $\alpha(B)$, q_i for $\alpha_j(B)$. For ease of discussion, I will take all $q_j = q$. X_{jt} are a set of R possible explanatory variables and ε_{jt} will be thought of as a white noise shock with zero mean. With these types of models there is not a clear-cut nesting. Although the very complicated model will nest much simple model $(p_1 > p_0,\ R_1 > R_0)$ one does not have nesting between simple sub-groups as such as two identical models except that one has X_{1t} missing and the other has X_{3t} missing.

For this type of single equation model it is possible to employ a simple ? complex strategy or a complex ? simple one, although the latter has better statistical theoretical justification. It has also been firmly attacked, for example, by Leamer (1983) and by Hoover (2000). Nevertheless, the complex to simple strategy in some form remains to dominant methodology; fairly complicated models are fitted and then simplified, usually by dropping terms with low t-values or parameters and re-estimating. A full-scale reduction is not satisfied by that step but also test for encompassing between models and finally, using a model selection criteria, to help decide on a final "best" model.

I have argued, in Granger and Jeon [2001], that seeking the best single model should not be the aim of the modeling process. With realistic sample sizes there will often be several alternative models which are not significantly different from each other. To attempt to rank them using an arbitrary model selection criterion, such as AIC or BIC, is losing sight of the economic objectives of the modeling process. The process of choosing the "best" single model by the model producer does not allow the preferences of the model consumer to be introduced into the selection procedure. The model itself is not the final objective, there has to be a purpose for the model being constructed, such as the production of forecasts, a policy scenario, or the estimate of a parameter of particular interest. If one has a single model, it will produce a single forecast for the consumer to use, but if there are several alternative models, there will be several forecasts which the consumer can combine. Experience suggests that combinations often produce superior forecasts. See for example Stock and Watson [1999]. Having more than one final model may well be advantageous. I have called the approach "thick-modeling" although a better name would be appreciated. A further advantage of using multiple models is that a more reasonable confidence interval can be obtained than if a single model is used. For a single model, a confidence interval is easily obtained **under the assumption** that the model is correctly specified. For several models, such an assumption is obviously not justified, an interval for the combination will be less optimistic and can be obtained by a bootstrap technique. A recent discussion is given by Aiolfi et al. [2001].

2. NONLINEAR MODELS FOR STATIONARY PROCESSES

There are very many types of nonlinear models, some surveys can be found in Granger and Teräsvirta [1993], Teräsvirta, et al. [1994], and Franses and van Dijk [2000]. Three broad classes might be denoted as:

i. Parametric, in which a particular functional form is used. For example

$$Y_t = f(X_{t-1}, Y_{t-1}, \alpha) + \varepsilon_t \qquad (4)$$

for some given function $f(\,)$.

i. Semi-parametric (sometimes called semi-non-parametric). For example the neural network model

$$Y_t = \sum_{j=1}^{q} c_j \phi_j (\beta_j' \underline{W}_t) + \varepsilon_t \qquad (5)$$

where $\underline{W}_t$ is a vector of lagged X's, Y's and a constant and $\phi_j(\,)$ is some pre-specified class of functions. In the neural network class they may all be the logistic

$$\phi(x) = 1/(1 + e^{-x}). \qquad (6)$$

iii Nonparametric forms as discussed in Härdle [1990]. For example, in which a purely data determined curve is fitted to a set of data using a moving window of some shape.

The examples given here of nonlinear models mainly illustrate a few types and by no means represent the full range of possible models. It is clear that a modeler needs a different strategy here than in the linear case as the models most certainly do not nest across classes, or even for types within a class. If one had a clear economic theory suggesting a specific nonlinear model that would obviously make a good staring point, but , least in macroeconomics, that is not found to occur [see Granger, 2001].

A plausible methodology would be to chose examples, or a type, in several distinct classes of methods, within each use a general to simple procedure to arrive at a single or small number of alternative models, and then to use the full group of models remaining to produce outputs that can be combined. Again, confidence intervals have to be found by bootstrapping.

There has been a tendency to loose faith in the usefulness of nonlinear models compared to linear ones, to explain conditional means, especially in macroeconomics. A natural reason for this to occur when the effects of cross-sectional and temporal aggregation on nonlinearity are considered, see

Granger and Lee [1999]. However, some recent results by Stock and Watson [1999] show further promise. They use monthly data from 215 U.S. macro series and compare forecasts from various models to those from an AR(4) model. In particular, they consider about 20 different types of univariate neural network models and for horizons of 1, 6, and 12 months, and find that each on the average does less well or hardly better than the AR(4). A different parametric and popular nonlinear model, the smooth transition autoregressive model (STAR) does rather worse. However, a combination of all the neural network and STAR models out forecasts the AR(4) models at all three horizons and by worthwhile amounts. The combination essentially used equal weights, after throwing out poor performers and involved no estimation of parameters. The implication is rather interesting, the individual neural network and STAR models fail to find, on average, useful nonlinearity in the macro-data yet the linear combinations do find it, suggesting that the series contain a subtle nonlinearity. Presumably a "common factor" approach should distill it better, but further study is required.

It should be clear that the "methodology" of nonlinear modeling in general is much less well developed than for a specific sub-fields such as linear models, STAR, neural networks, or non-parametric approaches. There are a variety of tests of linearity, [see for example Lee et al., 1993], which have mixed success. They can have good power against some types of nonlinearity and very little against other forms. There is not yet anything like a complete bibliography of type of nonlinearity, even for conditional means. The question becomes more complicated when one attempts to answer "what is linearity?" which is attempted in section 4(iv).

An important area that I will not discuss is the close rivalry between nonlinear models and linear models with time-varying parameters which can be investigated using state-space and Kalman filter techniques. In many cases it is very difficult to distinguish between the two classes of models, even though their interpretation is quite different.

3. INTRODUCING NONSTATIONARITY

The nonlinear models discussed above all require an assumption of stationarity. The border line between stationarity and nonstationarity includes the unit root processes and these have been well developed within the linear models and are starting to be developed for parametric nonlinear forms.

An example of a unit root process is the random walk

$$X_t = X_{t-1} + \varepsilon_t \qquad (7)$$

which, in operator form is

$$(1-B)X_t = \varepsilon_t \qquad (8)$$

where ε_t is a white noise.

For a unit root process ε_t is replaced by any stationary process, a_t. If ε_t has zero mean and if X_t starts at time $t = 1$, with $X_0 = 0$, then

$$X_t = \sum_{j=1}^{t} \varepsilon_{t-j} \qquad (9)$$

and it follows that

$$\text{variance } (X_t) = \sigma_\varepsilon^2 t. \qquad (10)$$

As this is time-varying, X_t is not stationary. It is important to notice that a unit root process is only an example of a nonstationary process and is hardly representative of this extremely wide class of processes. Similarly (9) shows that a random walk is an example of a "persistent" process, as old shocks continue to influence current values, but it is not the only possible example.

Unit root processes are associated with a number of properties: persistence, variance increasing linearly with time, possibly a linear trend in mean and autocorrecations that stay near one for all lags. However, other processes may have many of these properties. The one distinguishing property of a unit root process is that its difference is stationary. The notation I(*d*) was suggested by Box and Jenkins [1970] to denote a process that needs to be differenced *d* times to get to a stationary process (strictly an ARMA process that is both stationary and invertible). Originally *d* had to be an integer but later fractional values were considered.

There is a clear link between the I(1) process and linearity, as differencing is a linear operation. Similarly, the associated property of cointegration was originally defined in a linear way, with a linear combination of two I(1) variables being I(0). To extent these ideas to more general classes of processes, at least a generalization of I(0) is required. The definition which seems to have wide acceptance is that by Davidson [1999]. The definition is somewhat technical stating that the series X_t, 0< t < 8 is I(0) if the process $Z_t(\zeta)$ defined on the unit interval $0 < Z_n < 1$ and given by

$$Z_n(\zeta) = \sigma_n^{-1} \sum_{t=1}^{[n\zeta]} (x_t - Ex_t) \qquad (11)$$

for given 0 < ζ ? 1 where $\sigma_n^2 = \text{var}(\Sigma_{t=1}^{n} x_t)$ is such that $Z_n(\zeta)$ converges weakly to standard Brownian motion B as n ? 8. In other words, the standardized partial sums of the series must satisfy a functional central limit (or FCLT). This definition is theoretically helpful but less so in practice as there seems to be no available test for whether a process obeys the FCLT.

There have been several papers that have looked at the properties of functions of I(1) and I(d), d a fraction, processes using either theory or simulations. For the latter, see Granger and Dittman [2001] and for a substantial investigation of the properties of functions of random walks see Park and Phillips [1999]. There are too many results to summarize but basically if the function is monotonic the majority if the I(1) properties continue to hold, although variance may increase at a different rate. Karlsen, et al. [1999] use ideas based on non-recurrent Markov chains and a difference definition of persistence. The developing area of nonlinear, nonstationary processes involves new mathematics compared to that needed to study unit root processes which itself is quite different from that used for classical stationary series.

The methodological procedures are largely lacking in these newest areas and applications to data still await the first serious attempt.

4. PROBLEMS WITH FOUNDATIONS

It is hardly worth stating that a sound methodology has to be based on firm foundations. I believe that some of the basic concepts of the time series methodology discussed above are not well understood or even defined. Some examples are:

i. Trend. A simple example of a trend can easily be provided, such as a linear function of time or possibly a polynomial of time, but a complete parametric definition is very difficult. In the frequency domain it is generally agreed that the trend contributes a very narrow peak at the zero frequency but so does a constant, which is not a trend. In a sample it is a personal choice about what is a trend and what is not, but a monotonic component will usually be chosen. However, once a sample is extended in length what seemed to be a trend may break and be reclassified. It is not clear if a trend has to posses some smooth property over all time or just over some long time period, relative to the sample length available. What is interesting is that the literature contains papers that claim to discuss tests for trend, although they usually only consider the very special case of linear trend in time, so that the difference has a constant but no trend. The discussion became more involved and further confused by the addition of "stochastic trends" that had previously been known as integrated processes. They are now used as "common features."

ii. Linear and Polynomial Trends. These are examples of deterministic processes. Some models for the seasonal and chaos processes provide other examples. A basic property is that they do not change their generating mechanism over time. However, recently some models for breaking means have required "predetermined" breaks. The obvious question, although it is not often asked, is when were the series generated? Surely not at the start of the sample but perhaps at the start of the economy? Chaos theory suggests that the series is just generated from the generating mechanism at each new instant of time, based on the finite past of the series. Thus it is the generating mechanism that is fixed from the start of the economy. For the linear trend the change in the series is a particular constant and the first term in the series takes some given value. The obvious distinguishing property of a deterministic process is that it is perfectly forecastable in the short-run and highly so in the middle-run. My personal view is that occasionally a deterministic process can provide a useful approximation but that the economy is basically stochastic and so is best modeled using stochastic models.

iii. The Data Generating Process (DGP). This is a difficult topic that I believe is not well understood. As new data appears at regular intervals the DGP clearly exists and obtaining a model that provides an adequate approximation to it has become virtually the "Holy Grail" of econometric empirical work. It is usually not well defined. Some textbooks make careful efforts to provide useful definitions, such as Davidson [2000] and Spanos [2001], but in all too many cases the topic is handled casually. As the economy evolves with changing institutions, tastes, and technologies, so will the DGP evolve. Thus one would not expect to capture it fully with a constant parameter model. Critics of econometric models, particularly in books on methodology and economic philosophy, are generally confused between the DGP of the raw data of economics and that used to build economic models. The raw data passes through many stages of official analysis, such as seasonal adjustment, as well as substantial cross-sectional and temporal aggregation before being made available to economists for analysis. The econometricians' models are aiming to capture the main features of the DGP of the issued data, which is often far removed from the raw data coming from the basic economic decisions of agents. This comment is more true for macro data than for that arising from a micro survey. A difficulty with the evolution of the economy is that any model is likely to be better at explaining the past than in describing the future, unless the process of evolution can itself be modeled.

iv. Linearity. At first linearity might seem to be an easy concept to define, but in fact it is a much more subtle topic. If one is interested in modeling the conditional mean of Y_t given a particular information set, $\underline{X}_t$ say, then an obvious linear model is

$$E[Y_t \mid \underline{X}_t] = \underline{\beta}'\underline{X}_t, \tag{12}$$

compared to a nonlinear model

$$E[Y_t \mid \underline{X}_t] = \underline{\beta}'\underline{X}_t + g(X_t). \tag{13}$$

Thus, to be "linear" it seems that one needs $g(X_t) \equiv 0$. However, what if $\underline{X}_t$ consists of Z_{t+1} and W_{t-1} where $W_{t-1} \equiv \log C_{t-1}$, C_t being consumption, is the equation still linear? Of course if the residual to the conditional mean model is ε_t, the conditional variance model may be written in the form such as

$$E[\varepsilon_t^2 \mid \underline{X}_t] = \gamma' \mid X_t \mid + \sum_{j=1}^{q} \delta_j' \varepsilon_{t-j}^2. \tag{14}$$

Is this model linear or not? The answer is that it really does not matter; the objective surely is to obtain a model that performs satisfactorily.

A recent theoretical development by Bickel and Bühlmann [1996] has thrown open the whole question of what does linearity mean even in the univariate, stationarity case. If one accepts a clearly linear model to be the MA(8) process, with iid shocks, they shown that the closure of this class is complicated, and that many stationary, nonlinear processes will have exactly the same sample path as a linear process with a positive probability.

5. EVALUATION AND PURPOSE

I strongly believe that considering the modeling process from the perspective of the consumer of the model rather than from its producer provides a better viewpoint when discussing its evaluation. Rather than asking does a model fit well or satisfy various statistical criteria, it is better to ask how well is it performing its required task, or at least doing better than alternatives. It is generally easier to evaluate a model with respect to an alternative rather than in isolation. It is also preferable to evaluate in terms of the effect on the outcomes of economic decisions whenever possible. It is also very much easier to evaluate when you can compare the outcome of the model with some 'true' or actual value. For these reasons evaluation has evolved much further in the areas of forecasting and finance, although even there all the questions have not been resolved by any means. The classical tasks or purposes for the construction of a model are to test a theory, to estimate some parameter, to perform a policy simulation or to make forecasts. Only for the last of these can the model output be compared to an actual situation. As an example where such a comparison was not possible, Magnus and Morgan [1999] conducted an experiment where eight groups of econometricians used different methods on the same data set to estimate the elasticity of demand for food. They found different values that were usually significantly different from each other, but of course there was no true value against which to compare.

Alternative linear models are often compared in terms of their forecasting ability, particularly when there is plenty of data, as then questions of data-mining can be separated from the evaluation exercise. The same strategy is not always available with nonlinear models. To form forecasts with a univariate nonlinear AR model is very difficult beyond the first step, as one not only needs the correct functional form in the model but also the distribution of the residuals. Rather than relying on the one-step model to form multi-step forecasts, a pragmatic procedure is to reformulate and estimate a new model for each forecast horizon, and this practice is becoming common. The same approach cannot be used for many nonlinear moving average and bilinear models as they are not invertible, so that forecasts cannot be formed directly. Nonparametric models can be evaluated using forecasts within the range of the original data set but not outside that range and so evaluation is limited in practice. These difficulties with univariate models are naturally multiplied in nonlinear multivariate models and there is very little discussion of this evaluation problem.

When considering evaluation the best starting point is to consider the model's purpose. It is surprising that most empirical studies do not explicitly state the reason that the model is being constructed. Only by knowing the purpose can the study be evaluated. Models may be specified differently for alternative purposes. A forecasting model could look quite different from a data-mining exercise or a policy consideration. Only for papers in economic or econometric theory need there be no stated purpose as they will be trying to provide tools for workers at a later stage in the research process, they may be considered as intermediate goods. As such they should be evaluated by their potential users rather than by their peers.

6. OTHER TOPICS

6.1 The Truth

Searching for the truth is virtually never mentioned as a purpose by empirical researchers, although it is of considerable concern to the philosophers. I doubt if most econometricians have given much thought to the topic. In the "general to simple" framework it is probably believed that the truth falls within the "general" class and hopefully will remain in or close to the final small group of selected models. Here the truth will correspond to the DGP and the models will approximate to it. As the sample gets larger it will be hoped that this approximation gets better so that asymptotically the final model will be close to the truth. Of course this is all based on faith, on the quality of the original set of models and of the theory they are based on. In practice the truth could be well away and very little convergence is occurring. A different modeler could be start with quite a different initial set and be converging to her own small set of final models some distance away. Both could be distant from the truth. The interesting question arises, if one modeler by chance selects the true model, how would they know? It is certainly not a question of fitting well or even of fitting the best of a group of models as one can always find a model with those properties. My model fitting better than your best one, or even encompassing it, is good but still not convincing. A possibility to consider is that for the true model all decisions based on it will, in the long run, be superior to decisions based on any other model. By considering all decisions you are using an economic criterion and not tying it down to a particular utility function or class of such functions. As the economy is considered stochastic, superiority will not occur at each instant but only on the average.

6.2 Empty Boxes

When a new class of models is introduced it is not clear how useful in practice it will prove to be. Some years later, after further development and methods for its testing have become available, the search for realistic examples should take place. It can then be decided if it is potentially an important class with a number of applications. Unfortunately some models seem to correspond to virtually no real-world economic data and so can be described as belonging to an "empty box" [Craft, 1987]. Candidates for this classification based on their lack of practical usefulness with economic data include catastrophe theory, chaos, and fractionally integrated models, and I suspect that there may also be good examples from cross-sectional and panel econometrics. The fractionally integrated case is of some interest as the absolute returns of speculative prices have the correct second moments (autocorrelations) and so seemed to provide a good example. It was later realized that they had the wrong first moment, as a theoretically forecast nonlinear trend in mean is not seen in the data.

7. CONCLUSIONS

To summarize the main points of this paper: although the methodology for the specification and construction of linear models is developed, it is still rather controversial. For nonlinear models there is a great need for further methodological progress. Because nonlinear models do not nest easily it is difficult to compare competing classes of models. A possible solution is to concentrate on the purpose of modeling and thus on model evaluation. Rather than the traditional "thin modeling" approach where an attempt is made to find the best individual model, it may be preferable to keep a group of good models, record there outputs according to the purpose, and finally to combine in some fashion. This "thick modeling" is more flexible and will produce more realistic confidence intervals. It is not controversial to request that economic criteria be used for the evaluation rather then merely statistical criteria, where-ever possible. The practicality of this needs further development; an example is given in Granger and Pesaran [2000]. Some of the newly developed models in nonlinear stationary but particularly nonlinear and nonstationary cases are likely to prove to be "empty boxes" but this does not mean that they are not worth discussing at their beginning stages. It is possible that the recently developed method of "data-snooping" will be helpful in both rapidly deciding if a new method is useful and helping with the choice of an appropriate type of nonlinear model [for a particular data set see White, 2000]. The appendix contains discussion of a particular example of the methodological problems that arise when evaluating forecasts.

8. REFERENCES

Aiolfi, M., C.A. Favero, and G. Primiceri, Recursive "thick" modelling of excess returns and portfolio allocation, Working paper, Bocconi University, Italy, 2001.

Bickel, P.J., and P. Bühlmann, What is a linear process? Proceedings of the National Academy of Science, USA, 93, 12128-2131, 1996.

Box, G.E.P., and G.M. Jenkins, *Time Series Analysis, Forecasting and Control,* Holden Day, San Francisco. 1970.

Crafts, N.F.R., Empty boxes, The New Palgrave, A Dictionary of Economics, 12, 133-134, 1987.

Davidson, J., When is a time series I(0)? Evaluating the memory properties of nonlinear dynamic models, Working papers, Cardiff Business School, 1999.

Davidson, J., *Econometric Theory,* Cambridge University Press, 2000.

Franses, P.H., and D. van Dijk, *Nonlinear Time Series Models in Empirical Finance,* Cambridge University Press, 2000.

Granger, C.W.J., An overview of nonlinear macroeconomic empirical models, *Journal of Dynamic Macroeconomics*, (in press) 2001.

Granger, C.W.J., and I. Dittmann, Properties of nonlinear transformations of fractionally integrated processes, Working Paper, Department of Economics, University of California, San Diego, 2001.

Granger, C.W.J. and Y. Jeon, Thick modeling, Working Paper, Department of Economics, University of California, San Diego, 2001.

Granger, C.W.J., and T.-H. Lee, The effects of aggregation on nonlinearity, *Econometric Reviews*, 18, 259-270, 1999.

Granger, C.W.J. and T. Teräsvirta, *Modeling Nonlinear Economic Relationships*, Oxford University Press, 1993.

Granger, C.W.J., and M.H. Pesaran, A decision theoretic approach to forecast evaluation, In: *Statistical Finance: An Interface*, W.-S. Chan, et al., (eds), Imperial College Press, 261-278, 2000.

Härdle, W., *Applied Nonparametric Regression,* Cambridge University Press, 1990.

Hendry, D. F., *Dynamic Econometrics*, Oxford University Press, 1995.

Hoover, K. and S.J. Perez, Three attitudes towards data mining, *Journal of Economic Methodology*, 7, 197-210, 2000.

Karlsen, H.A., T. Myklebust, and D. Tjøsthiem, Nonparametric estimation in a nonlinear cointegration type model, Working paper, Department of Mathematics, University of Bergen, 1999.

Leamer, E.B., Let's take the "con" out of economics, *American Economic Review,* 23, 31-43, 1983.

Lee, T.-H., H. White, and C.W.J. Granger, Testing for neglected nonlinearity in time series models, *Journal of Econometrics,* 56, 269-290, 1993.

Magnus, J.R. and M.S. Morgan, *Methodology and Tacit Knowledge,* J. Wiley, 1999.

Park, J.-Y., and P.C.B. Phillips, Asymptotics for nonlinear transformations of integrated time series, *Econometric Theory*, 15, 269-298, 1999.

Poon, S.-H. and C.W.J. Granger, Forecasting financial market volatility: a review, Working Paper, University of California, San Diego, 2001.

Spanos, A., *Introduction to Econometrics,* Cambridge University Press, 2001.

Stock, J.H., and M.W. Watson, A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series, Chapter 1 of *Cointegration, Causality, and Forecasting,* A festschrit in honour of Clive W.J. Granger, R.F. Engle and H. White, (eds), Oxford University Press, 1999.

Teräsvirta, T, D. Tjøstheim, and C.W.J. Granger, Aspects of modeling nonlinear time series, Chapter 48 of *Handbook of Econometrics*, volume IV, R.F. Engle and D. McFadden, (eds), Elsevier, 1994.

White. H., A reality check for data snooping, *Econometrica*, 68, 1097-1127, 2000.

APPENDIX

EVALUATION OF VOLATILITY FORECASTING TECHNIQUES

Two important, and practical, questions in finance are (i) can volatility be forecast; and (ii) which of the various available models of volatility provide the best forecast? The first problem encountered is that it is unclear exactly what is being forecast. If I were to forecast next week's volume for a particular stock, then at the end of the week a value for volume would be available in the financial page of, say, the *New York Times*. This is even true for some stock indices where the concept of volume is rather unclear. However the financial press does not publish figures for volatility, except for high minus low price over some period. When evaluating a forecast having an "actual" against which to compare the forecast is rather important but in this area researchers construct their own estimates of the actual, usually the variance or standard deviation or mean absolute deviation of returns over the forecast period. Although these measures are certainly highly related they will have different properties. A further decision that the investigator needs to make is the time period over which volatility is measured, should it be five minutes, an hour, a day, a week? With the mass of high quality, high frequency data available in financial markets all of these periods are possible as well as many others. Researchers seem to chose a period, or several periods, for their studies for statistical convenience. A better criterion would be relevance for actual financial decision makers.

In a recent survey of the volatility forecasting literature, Poon and Granger [2001] considered 79 papers concerned with evaluation, rather than with problems such as the specification, testing, estimation, or asymptotic properties of volatility models for which there are many further papers. It was found that several different version of "actual" were used and many different time periods and forecast horizons making comparison of studies difficult. There is a wide agreement about the first question; it is generally that volatility, however measured, is somewhat forecastable and several papers discussed only this topic.

The question of which method is best is very complicated as there are many candidates which can divided into four main classes:

i. HISTORIC, in which the "actual" measure, a standard deviation for example, is used to form a series of values and simple time series models constructed, such as a random walk, autoregressive, exponential moving average or possible more complicated forms such as a nonparametric representation.

ii. GARCH, the generalized autoregressive conditional heteroskedastic class of models which have been much discussed in financial econometrics. There are numerous further generailzations including the exponential GARCH, Switch GARCH, and various nonlinear forms. Generally this class is fairly easy to estimate and has well known properties.

iii. IMPLIED VOLATILITY, being derived as an implication of the Black-Schole's option pricing formula. For its use, appropriate options are needed, and so it is not available for all speculative markets and estimation is not simple.

iv. STOCHASTIC VOLATILITY (SV), which is a class of pre-specified statistical models with a Bayesian aspect. Considered as an alternative to GARCH, they are more difficult to estimate, requiring two sets of stochastic inputs and their assumed distribution.

All but the IMPLIED are based just on the sequence of past returns suitably manipulated. IMPLIED uses a wider information set as option prices are also involved. Only IMPLIED has a strong theoretical foundation, although the foundation does require certain assumptions, such as normality, which are known to be incorrect.

If one has an actual a_n and a pair of forecasts f_n, g_n the result is a pair of forecast errors

$$ef_n = a_n - f_n, eg_n = a_n - g_n. \qquad (A.1)$$

The usual method of evaluation is to use an estimate of the average cost

$$\frac{1}{N}\sum_{n=1}^{N} c(ef_n) v \frac{1}{N}\sum_{n=1}^{N} c(eg_n) \qquad (A.2)$$

where *c*( ) is some appropriate cost function. The obvious question is how does one decide which cost function is appropriate. This question must surely depend on how the volatility forecast will be used by a financial decision maker and merely guessing at a function is just academics using their own preferences. In the papers surveyed in Poon and Granger [2001] roughly fourteen different cost functions are used, the most popular being mean squared error, root mean squared error (RMSE), and mean absolute error, all of which are symmetric, so that negative errors are given the same weight as positive errors of the same size. Again, this lack of uniformity of approach across researchers makes comparisons difficult and suggest a lack of precision in the research. Most of the papers were content to simply report that Method A achieved a lower average cost score than Method B, rather than attempting a statistical test of significance between the two score values. A simple way to proceed is to combine the two forecasts, giving

$$c_n = \alpha f_n + \beta g_n + \mu ec_n \qquad (A.3)$$

where α, β, μ are chosen by regression, and to compare the *t*-statistics for ?, ?. Only six papers in the survey considered combinations, and for four of these the combined forecast outperformed its constituents, in the sense of getting a lower score.

Despite all the difficulties in comparison, a meta-analysis of the results of forty papers involving direct comparison of techniques from different classes, found that IMPLIED generally proved to be superior, with GARCH and HISTORIC roughly similar in second place. SV either did well or very poorly but was found only in a handful of studies. The implication is not that IMPLIED should be used, as a combination may do better, and IMPLIED is not always available.

There are further possible biases in the survey; supporters of a method may have been selective about what they chose to try to publish, editors and referees could have been biased in various directions and this is reflected in what papers get published and the authors of the survey could have biases due to their own publishing history or intentions. To undertake a definitive empirical study in this field is going to be difficult until the area reaches agreement about cost functions, how actual should be measured and a few relevant time periods, at the very least.