AUTOMATED DISCOVERY IN ECONOMETRICS

By

Peter C.B. Phillips

July 2004

COWLES FOUNDATION DISCUSSION PAPER NO. 1469

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

Automated Discovery in Econometrics*

Peter C. B. Phillips

May 14, 2004

Abstract

Our subject is the notion of automated discovery in econometrics. Advances in computer power, electronic communication, and data collection processes have all changed the way econometrics is conducted. These advances have helped to elevate the status of empirical research within the economics profession in recent years and they now open up new possibilities for empirical econometric practice. Of particular significance is the ability to build econometric models in an automated way according to an algorithm of decision rules that allow for (what we call here) heteroskedastic and autocorrelation robust (HAR) inference. Computerized search algorithms may be implemented to seek out suitable models, thousands of regressions and model evaluations may be performed in seconds, statistical inference may be automated according to the properties of the data, and policy decisions can be made and adjusted in real time with the arrival of new data. We discuss some aspects and implications of these exciting, emergent trends in econometrics.

Keywords: Automation, discovery, HAC estimation, HAR inference, model building,.online econometrics, policy analysis, prediction, trends.

JEL Classifications: *C32,* C100, C500, C870

*Written in April 2004 for the 20th Anniversary Issue of *Econometric Theory.* Helpful comments by the Co-Editor, Oliver Linton, Benno Pötscher and two referees on the first draft are gratefully acknowledged.

Automated Discovery in Science

> "There's been a profound transformation in economics since the early '70s, particularly in the elevation of empirical research in comparison to pure theory. The explosion of computing power has been integral to that."
> Levitt (2004)

Automated discovery in science is a fairly recent phenomenon. It is commonly associated with our newfound capacity to collect, store and process vast amounts of data in extremely short periods of time. These capabilities come from sheer computational power and storage capability in conjunction with electronic communication, data processing and statistical analysis. Rapid information processing accelerates learning. It also enables algorithms to be implemented that automate judgements and scientific evaluations that would otherwise be made by human participants. The upshot is that empirical and experimental research can now be conducted in an automated fashion with much more limited human involvement than in the past.

Fast processing capability of this type is important in many areas of science and engineering. Familiar examples occur in medical diagnostic imaging, the processing of particle collision data in experimental physics and in the engineering of space guidance systems. In space-craft guidance, for instance, rapid data processing is needed to capitalize on the short window of opportunity that exists for the firing of rocket engines in order to maneuver a space vehicle into a safe landing trajectory. Notwithstanding extensive planning and preparatory analyses, critical final calculations must be performed in real time prior to engine ignition and these involve rapidly processing the current coordinates, velocity and trajectory of the space vehicle. A related example is the use of unmanned robotic vehicles like the Mars exploration rovers. These machines are programmed to function as "geologists", analyzing rocks and soil in their environment in an automated way that does the work of human scientists and makes new scientific discoveries.

Another example in a different field is the use of automated virus detection sofware to search and discover new computer viruses. If anti-virus software is to be effective in the face of mass proliferation of viruses, new viruses must be recognized quickly and incorporated into virus definition files so that they are accessible to users for updating virus definitions on individual machines. Popular anti-virus programs automate such processes. For instance, the engine behind Norton AntiVirus is SARA (Symantec Antivirus Research Automation). This fully automated anti-virus computer system provides ongoing screening of internet files and virus analysis, it implements virus definitions and performs file disinfection, and then adds digital cures to the virus definitions so that they are ready for updating. Likewise, live updating of virus definitions on individual PC's is now implementable in an automated way, just as service packs and security patches may be downloaded automatically to update operating systems software files. IT security systems at major institutions now screen thousands of incoming messages and attachments a minute for viruses and automatically reset file suffixes to protect users from innocently opening attachments and releasing viruses and worms. Even with such automated monitoring and protection mechanisms in place, great damage is still being inflicted by the rising number of virus, worm and spyware attacks.

In a related fashion, the processing capacity of modern computers makes it possible to subject vast amounts of data to statistical analysis with little or no human intervention. Automated regression and statistical search algorithms can often be completed within seconds of the arrival of new data. This means that policy decisions like portfolio investment allocations can be made and adjusted on the fly in real time using such analysis, just as a craft is maneuvered in space by remote control once its latest coordinates and trajectory are processed.

Glymour (2004) identifies these modern computer-led developments in science as bearing the hallmarks of a scientific revolution. In Glymour's view, this revolution breaks the long-standing tradition, often associated with Popper (1959, 1963), of a steady sequential progression of advance and falsification in science:

> "The change is from the textbook scientific paradigm in which one or a very few hypotheses are entertained and tested by a very few experiments, to a framework in which algorithms take in data and use it to search over many hypotheses, as experimental procedures simultaneously establish not one but many relationships."

These changes, which have been affecting many different areas of scientific work in the last decade, are now beginning to be felt in econometrics.

Automation in Econometric Modeling

Methodological and software advances in econometrics in recent years have made the idea of automated modeling a practical reality in many applied econometric problems. Some of these methods are already in use in financial econometric analysis (Pesaran and Timmermann, 1995, 2002 & 2004), in macroeconometrics (Hendry and Krolzig, 2001 & 2002) and in ex ante econometric forecasting and policy analysis (Phillips, 1992, 1995, 1995b; Schiff and Phillips, 2000).

These methods use various model determination procedures in a largely automated fashion to model and predict single and multiple time series. Work on large multidimensional panels (Bai and Ng, 2001) is also ongoing and can be used in a mechanical way to search for a small number of reference variables that suitably capture the variation in the larger set. Methods of this type have been used in dynamic factor modeling and forecasting exercises (Stock and Watson, 1999). Related work has been underway in the systems engineering literature for a decade or more on subspace algorithms for estimation, prediction and model selection in large linear dynamic systems (Bauer, 2002). This work has recently been extended to allow for unit roots and cointegration (Bauer and Wagner, 2002 & 2003) and is overviewed by Bauer (2004) in the present issue. Algorithms have also been developed for analyzing causal structure among variables by computing conditional independence relations in what are called Bayes net diagrams (c.f. Pearl, 2000). These graph-theoretic approaches are discussed and used in Swanson and Granger (1997) and Hoover (2004) in this issue. In addition to this work, there appears to be scope for using genetic programming algorithms (like those utilized in computer science and optimization theory) to find suitable functional forms in empirical work and thereby assist the model building process. These procedures work by a constructive process of combining elementary mathematical operations through tree structures and mutations to develop more complex functions and are capable of identifying unusual functional forms. Some econometric examples are given in Kaboudan (2000) and Milev (2004).

Just as financial analysts hunt out market opportunities for investment, it is easy for empirical modelers to use modern computing power and tailored software to search systematically over models for ones with apparently superior performance. Statistical testing as in general to specific modeling algorithms (e.g., Perez and Hoover, 1999 ; Hendry, 1995), or direct model selection methods (Phillips, 1996) may be used in this process. The practice is steadily becoming widespread in econometrics. Even in what now seem routine exercises like unit root or cointegration testing, important decisions on lag length parameter settings and variable inclusion need to be made. These decisions often influence the results of inference and so it is reasonable to integrate such decisions into the overall process of finding a suitable model specification rather than to isolate them and treat them separately. Bayesian thinking, of course, tends to encourage coherent model evaluation along such lines. Bayes methods also provide a natural mechanism for smoothing over uncertainties by model averaging and forecast combination, both of which are becoming more common in applied work. Such procedures may, of course, be implemented over subsets of models corresponding to those that are found a posteriori to be most likely.

Model determination exercises of this type inevitably generate some controversy. Particularly since Leamer (1978), there has been ongoing discussion of the validity of specification searches in econometrics and the effect of data mining on inference. In the present issue, this controversy is reflected in the contributions of Hansen (2004), Hoover (2004), Leeb and Pötscher (2004), Paruolo (2004), Perez-Amaral, Gallo, and White (2004), and in the dialogue of Granger and Hendry (2004).

In another recent contribution to this literature, White (2000) examined data-snooping exercises and gave statistical criteria that facilitate the assessment of a chosen model from such a search against certain benchmarks. A central difficulty in this assessment arises from the need to allow for the cross-model statistical dependence that arises from reuse of the same data across models. The complication bears some similarity to the type of dependence that can arise in cross section or panel modeling where there is cross section error dependence but no natural ordering of the data and therefore no natural concepts of weak and strong dependence. A further complication is that practical data-snooping typically involves hunting for a model that works well, not just comparing a fixed number of models and locating the 'best' one. In other words, a specification search is often called off only when a seemingly adequate model specification is found. This means that the number of models examined is itself data-dependent and the endogeneity needs to be accounted for in the statistical analysis.

Recent empirical work by Sala-i-Martin (2003) exemplifies this phenomenon, where literally millions of regressions were run in a hunt for significant explanations of economic growth. At a more subtle level, these data-snooping effects operate across research communities and over time. New empirical work regularly builds on past studies and, however careful individual practitioners may be in the implementation of their methods, data-snooping effects operate in the aggregate across researchers. Practical econometric work seems to be moving inexorably in this general direction and the subject is ripe for theoretical study.

A further issue that complicates matters is the effect that model selection has on subsequent estimation. As discussed in work by Pötscher (1991) and Kabaila (1995), the use of model selection (e.g., in lag length determination) can have a big effect on the distributions of econometric estimators, tests, confidence regions and prediction intervals. Some related research on the finite sample distribution of post-model-selection estimators in the normal linear regression model is given in Leeb and Potscher (2003). More recently, Leeb and Pötscher (2003b) established an impossibility theorem, revealing that model selection searches set up an obstacle that prevents uniformly consistent estimation of the distribution of subsequent estimators. These results are discussed and extended in Leeb and Pötscher's (2004) contribution to the present issue. This work has deep significance for applications, revealing an important limitation of what can be accomplished in estimation when uncertainty about model specification requires the use of model selection.

This research follows in the tradition of an earlier literature on pre-test estimation (e.g. Judge and Bock, 1978). That literature was guided by a similar concern about the effects of prelimary specification tests on estimation. Careful analysis of this problem in linear regression analysis showed that pre-test (of linear restrictions) estimators were inferior in terms of risk to ordinary least squares regression over an infinite range of the parameter space, although they could be far superior in neighborhoods where the restrictions were approximately correct. Further, carefully designed biased estimators (like Stein-rule, positive part and Bayes estimators) were capable of uniformly reducing risk and producing non-trivial gains in multivariate contexts provided risk averaging across dimension was permitted.

Important though these contributions were, their direct impact on applied econometric work has been minor. One reason is that, while prescriptions were available for point estimation, guidelines for hypothesis testing and interval estimation have proved much more difficult. Another is that extensions of these results to more realistic models than linear regression with fixed regressors has also been an obstacle.

In short, this interesting research agenda complicated inference even in simple linear models, did not appear to generalize in a simple manner and did not produce simple algorithms for inference. On the other hand, its indirect impact gave increasing recognition to the role of pre-testing in modeling and showed that empirical workers need clear guidance from theorists about desirable procedures and general rules to follow in applied work that will validate or robustify inference in the face of sequential data analysis. Attention to these issues is ongoing and some relevant recent work, for instance, examines the harm of pretesting and its effects on forecasting (Danilov and Magnus, 2004a and 2004b). None of this research has, or will, put least squares out of business. But it has increased awareness of some of the implications of preliminary specification search on inference.

My own practical experience in the field of automated modeling relates primarily to the use of automation in building models for ex ante macroeconometric forecasting. I started out in this field in the early 1990s, keen to apply some automated model determination methods that I had developed in joint work with Ploberger (1994, 1996). These techniques were well suited to finding models in the reduced rank regression class and error correction model class for practical uses such as economic forecasting (Phillips, 1995a & 1995b), policy analysis and impulse response analysis (Phillips, 1998), the latter subject to conventional issues such as shock identification. In such problems, many seemingly innocuous but often practically important decisions are made in building models, such as trend degree specification, intercept inclusion or exclusion, the timing of any structural breaks, and lag length selection. In conventional econometric work, such decisions are commonly made, possibly as part of some group of overall specification tests, prior to further analysis that may involve testing for reduced rank and cointegration or the presence of certain causal patterns in the data. In real time ex ante forecasting, such decisions on the details of model specification can (and, arguably, should) be made jointly, for example by model selection methods, albeit with some consequential effects on subsequent inference as discussed in the last paragraph. It is also possible to weight models according to posterior probabilities and combine them to produce weighted forecasts. In implementation, these model determination exercises take only a few seconds to perform on modern computing equipment, although computing time rapidly increases with the size of the system and with the number of evaluations performed. Hundreds of multivariate regressions are well within the short-time-frame capability (less than five minutes, say) of present equipment, including laptops, and once conducted a final model can be selected on some criterion such as penalized or predictive likelihood. Alternatively, the methods can produce a weighted average of several models based on their posterior probability. All of these decisions can be built into the software algorithm so that users need only specify the variables to be included and, if they wish, insist on certain restrictions, such as specific cointegrating relations involving certain variables.

The development time that went into programming this work ran into months. But, in large part this was a one-time fixed investment. Once the system software was set up, forecasts could be obtained from large multi-equation systems in a matter of seconds, with hundreds of regressions and model evaluations (including unit root and cointegration tests) automatically conducted in this time. Practical experience soon revealed that the forecasting performance of such automated methods was frequently very competitive with that of labor intensive modeling methods, sometimes where entire research teams were involved in building and maintaining models. Some examples and comparisons of this type for the New Zealand economy are given in Schiff and Phillips (1999). In forecasting US GDP and inflation up to 12 quarters ahead for four years from 1995-1998, my findings (Phillips, 1998b) were that automated use of multivariate error correction model methods (with in-built automated selection of intercepts, trends, and lag length) did as well as seasoned macroeconometric forecasters on the level playing field of ex ante economic forecasting.

These comparisons do not necessarily devalue the contribution of more labor intensive methods. Considerations of which variables to include, the quality of the data, and the relevant ideas from economic theory and past empirical studies will always be matters for direct human involvement. Dealing with data-revisions and series updating also requires manual intervention. But the comparisons do point to the reality that is now upon us – that much of the applied work that used to take weeks or even months to complete can now be done in a matter of a few seconds or minutes by automated procedures. Moreover, these procedures have the great advantage that they can be mounted on the web for online use by anyone on a 24/7 basis, much as the simple graphical display of exchange rate and stock price data is now routinely available at financial sites on the web. Some discussion of these possiblilities is given in Phillips (2003). Examples of the use of these methods in practice online are now available at the website **http://www.covec.co.nz/predicta/**. Interactive elements have not been activated on this site for security reasons. Browser interaction, uploading, and open ports all increase website and server vulnerabilities. These threats must be faced as we progress in the development of online econometric technology and the implementation of effective defense barriers becomes a vital part of any such installation, as indeed it is in network operations more generally because of the rising tide of malware on the internet.

A great advantage of online automated econometrics is that it can make available econometric methods to a wide audience in our communities. Business people, politicians, journalists, educators and economic commentators may use these online econometric tools to forecast variables that are of interest to them and to conduct some elementary policy analyses. Newsroom interviews and national level economics discussion can be enlivened by showing the data and producing forecasts online as the discussion proceeds. Policy analyses can be computed that map out the trajectories of key economic indicators under different assumptions about forthcoming Central Bank interest rate decisions and Government taxation policy or even external economic shocks. As methods become more sophisticated it will become possible to program automated facilities so that they are flexible enough to match user needs in decision making. For example, it is possible to allow users to select loss functions that they deem most relevant to the application in hand in evaluating results like inflation and unemployment forecasts. Such real time econometric analysis adds quantitative information to the discussion of economic issues and it can be compelling in clarifying differences in the projected outcomes of various economic policies.

When utilized in such a way, econometric data analysis can be of immediate and transparent benefit to society. Basic data analysis is now becoming familiar to a wider public through televised weather and sports commentating. The value added is particularly apparent in the television coverage of sports events, where statistical data from past events is combined with ongoing data analysis of the current event to provide more informative television coverage. In live tennis broadcasting, for example, data is collected on each point as the match proceeds regarding such things as placement of service, number of shots in the rally, number of winners, number of forehand/backhand errors and so on. This data is analyzed as the match continues and the results are reported in the ongoing commentary and in onscreen visual data displays, showing such statistical information as each player's service location up to the present moment in the match. Viewers may then evaluate the data themselves, as well as listen to the analysis by commentators. Ongoing match data of this type can be combined with past data about earlier matches by the competitors to highlight similarities and differences and to support real time match projections/predictions being made by the commentators. In the past, television coverage relied on human memory to bring these elements into match coverage. Nowadays, sports data bases and ongoing statistical data analysis are available to commentators to enrich coverage in a more detailed, rigorous, and visually engaging manner.

If televised economic commentary and data analysis ends up proceeding along similar lines to championship tennis match coverage the resulting public exposure of econometric methods will have its share of drawbacks as well as benefits. Indeed, we may well expect commercial and media usage of econometrics to bring the subject some notoriety, like meteorology, by publicly revealing its limitations when it fails badly. But failures form part of the overall picture and need to be acknowledged. The methods of econometrics are developed to be used and ongoing changes in computerization and automation make these methods eminently more useable. So it is both desirable and inevitable that econometrics will become more widely used and available in society. The societal effects of these changes in the practical side of econometrics may end up being a further manifestation of the scientific revolution to which Glymour (2003) has so aptly drawn attention.

Robustification and HAR Inference

Empirical investigators in economics face many hard realities. One inescapable reality is that the models used in empirical work are inevitably wrong. Even if an empirical model were thought to be correctly specified ab initio, a relevant policy intervention would typically disturb the empirical relationship. As Goodhart (1975) aptly observed,

> "any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes"

Goodhart's law (as this is now called) and the closely related Lucas (1976) critique (see Chrystal and Mizen, 2001, for a discussion of their differences), as well as their many antecedents in the history of econometric thought (as discussed in detail by Hendry and Morgan, 1995) emphasize that in the world of economic activity the observed system and its apparent statistical regularities are not invariant to policy actions (or rules) and other interventions by authorities. This line of thinking has placed a premium on finding autonomous economic relations or economic laws that stem from "deep" theories about the behavior of economic agents.

In practice, of course, all empirical models can only provide simplified representations of economic systems. Simple behavioral models, as Milton Friedman once said, lead to powerful predictions, the permanent income hypothesis being a prominent example. Simplicity comes at the price of reduced realism, so we often prefer to think of the models we use in economics as being sophisticatedly simple (c.f., Zellner et al., 2001; Zellner, 2004). Sophisticated simplicity seeks to buy more realism in a model by marrying 'kernels of economic truth' from economic theorizing about primitive behavior with stylized empirical facts that are known to accord, at least in a general way, with observation. But even very sophisticated models rely on some premises that are unwarranted, and practical empirical models are, as we all know, nothing more than approximations.

In acceptance of the reality that practical models are approximations, econometric methods have been devised to accommodate some generality in the maintained hypothesis. One approach is to impose only weak requirements on a model's supplemental components. For instance, the regression errors may be taken to be stationary or weakly dependent and mildly heterogeneous; or the time series being studied may be assumed to be rather general integrated or fractionally integrated processes. Part of the appeal of the unit root/cointegration revolution has surely been the very general form of the models with which these methods can deal.

Such general maintained hypotheses are sometimes adequate to justify or validate inference. However, reality suggests that even these general hypotheses are wrong – stochastic trends do not always fall in the class of integrated processes and error terms may have some nonstationary elements that are not easily modeled. Ultimately, whatever empirical model we write down, no matter how general it is, is still likely to be misspecified.

The econometric challenge before the empiricist is to find and justify a particular model as a suitable lens for viewing the data, making inferences, producing forecasts and analyzing policy. Several issues present themselves in this process. One is the problem of hunting for an appropriate model specification. Some aspects of the search process of “finding” a suitable model and automated mechanisms for doing so were discussed above. A second issue is robustification. This is our focus in the present section.

In recognition of the fact that empirical models are misspecified in an unknown way, it is now commonplace to attempt to robustify inference. Regression residuals inherit the effects of specification errors in a model, and robust methods of inference use procedures that insulate the estimated standard errors of the regression coefficients from the effects of certain types of model misspecification. The most common procedures utilize consistent covariance matrix estimates that adapt for heteroskedasticity and autocorrelation of unknown form in the errors. These so-called HAC estimates are appealing because they lead to asymptotic tests involving convenient standard normal and chi-squared distributions which remain valid for a wide class of equation errors with weak (short memory) temporal dependence and heterogeneity. Because of their appealing asymptotic properties and their computational convenience, HAC estimates are now in widespread use in empirical work. Some suggestions have also been made to extend their validity to situations where there is unknown forms of cross section dependence in panel data studies (Driscoll and Kraay, 1998).

HAC estimates are typically formulated using conventional kernel smoothing techniques (for an overview, see den Haan and Levin, 1997), although different approaches like wavelets (Hong, 2001; Lee and Hong, 2001; Duchesne, 2003) may also be used. A new approach that involves automated regression on a trend basis is explored in the paper by the author (2004) in the present issue. HAC estimates may also be extended to accommodate long memory dependence, as shown in Peter Robinson’s (2004) paper in the present issue. The asymptotic theory justifying the use of HAC procedures in econometrics has generally closely followed earlier work in statistical theory on the asymptotic properties of kernel estimates of the spectrum. However, particular features of the data may sometimes prevent the immediate use of conventional asymptotic theory. One example is the presence of unbalanced data sets, which commonly appear in applied econometric work, and which are considered by Linton (2004) in the present issue.

Consistent HAC estimation provides asymptotic not finite sample robustness in econometric testing. While the generality that HAC estimation lends to inference is appealing, our enthusiasm for such procedures needs to be tempered by knowledge that finite sample performance can be very unsatisfactory. Distortions in test size and low power in testing are both very real problems that need to be acknowledged in empirical work and on which further theoretical work is needed. The situation is particularly acute when there is strong autocorrelation in the data. In such cases the spectra is peaked at the origin and kernel-based HAC estimates typically underestimate the peak. This tends to produce confidence intervals that are too narrow and liberal-biased tests. Wavelet methods appear to do better in such cases (Hong and Lee, 1999). Some discussion of the failings of conventional approaches to HAC estimation are contained in recent work by Kiefer and Vogelsang (2003). Sul, Phillips and Choi (2003) provide some further recent evidence on this matter, and show that the common use of prefiltering and recoloring in HAC estimation is also not a cure-all and can produce additional bias problems (especially when the data is demeaned or detrended) and even test inconsistency, as in KPSS testing for stationarity (Lee, 1996).

Robustification of inference is achieved whenever the test statistic is asymptotically pivotal under a general maintained hypothesis for the regression components. For this to be so, it is not necessary to use consistent HAC estimates. It has been known for some time, for instance, that any procedure that scales out the effects of the nuisance parameters in the test statistics will work. Kiefer, Vogelsang and Bunzel (2000) gave an important recent demonstration when they suggested the use of untruncated, inconsistent kernel estimates in testing. In such cases, the limit theory of the resulting test statistics is no longer as convenient as the standard normal or chi-squared but it is free of nuisance parameters. The fact that the limit theory of tests that use inconsistent estimates is non standard seems to play an interesting role in improving the size properties of the resulting tests, essentially because it preserves the finite sample randomness of the denominator in $t-$ and $F-$ ratios in the limit, unlike conventional asymptotic chi-squared tests. The test statistic ends up being closer to its limit distribution by an order of magnitude in the asymptotic sense than that of tests using consistent HAC estimators (Jansson, 2004). While such tests typically have better size than those that use HAC estimators, there is also a clear and compensating reduction in power. Work on finding procedures that improve size properties while retaining power in robust econometric testing is a challenge that is presently ongoing. Some recent efforts in this direction include Jansson (2004), Phillips, Sun and Jin (2003a &2003b), Vogelsang and Kiefer (2003). These robust inferential techniques may be grouped together with conventional HAC procedures as having the same general goals. The term heteroskedastic and autocorrelation robust (HAR) methods can be used to describe them collectively.

The HAR approach seeks to robustify inference in a way that accommodates departures from the model but keeps statistical behavior within the realm of some maintained set of general hypotheses about the processes being observed. The maintained hypothesis may be as general as some form of weak or strong dependence with controlled heterogeneity in the component errors or some encompassing class of stochastic trends. This standpoint seems both flexible and reasonable, and it underpins the HAR approach. But there is another position one can take that justifies this approach.

In particular, one can productively debate whether the formal structure of probabilistic models ever allows for a true data generating process (dgp). Such a debate may appear to belong solely to the philosophical realm of econometric methodology and have little connection with practical methods. Yet the issue touches a fundamental nerve-centre in econometric modeling and affects the interpretation of the statistical methods used in econometrics. Some dimensions of this complex subject have been thoughtfully addressed in recent studies by Keuzenkamp (2000) and Cartwright (2001).

As we look more carefully at the data and as the number and nature of the observations of economic activity change, it appears inevitable that the mechanism of economic data generation changes. This is not just a matter of the mechanism itself evolving over time, a phenomenon that all too frequently does happen as economic institutions and policy goals change (c.f. Goodhart's law), but also that the underlying probabilistic framework is inevitably too underdeveloped to reflect fully the factors involved in the determination of the observed data.

To illustrate, we take an example from modern financial econometrics. In this field, it is now popular to model frequently-observed processes like financial asset prices in terms of a continuous stochastic process. Yet further inspection and the collection of ultra high-frequency data reveal that the data themselves are related to the method of observation, just as in quantum physics the measurements may affect the observations. As we look carefully at intra-day financial data, for instance, we find that the observed process depends on specific features of the market place and that this market microstructure plays a role in every observed data point. Microstructure itself tells a new empirical story which relates to events on a wider probability space, like the placing of orders to buy and sell, time limits on orders, conditional orders, regulation of the trading day and so forth. Such events, which reflect the decisions of many different market participants, the procedural rules of brokers and traders, as well as the institutional regulations of the marketplace (which have historically evolved partly in response to past random shocks), all end up figuring as part of the data generating process. Proper consideration of these events would require the probability space itself to be augmented, in combination with an extension of the modeling apparatus to accommodate all that has been identified in the wider empirical story. Clearly, this process can be continued almost indefinitely, at which point the great omega in the probabily space $(\Omega, \mathcal{F}, P)$ is itself seen to be inadequate to the task and we have to admit that there is no 'true' dgp in a probabilistic sense. Against such a background, models that are now popular in financial econometrics like scalar diffusion equations and affine multi-factor models can only ever be viewed as crude empirical approximations. The same can, of course, be said of empirical econometric models in other applied areas.

In short, the approximate nature of probabilistic models is endemic in economics. But if there is no truth, then what is meant by empirical discovery? In writing his classic textbook, Malinvaud (1966) characterized the aim of econometrics to be "the empirical determination of economic laws". One way of interpreting this description, while admitting the fact that misspecification is endemic, is to say that econometrics is simply concerned with the discovery of empirical relationships. Within these relationships, some underlying economic law (like the law of one price) may reside as a "kernel of truth" and may even be represented in a mathematically precise form as a "primitive dgp" without ever requiring that this be a complete underlying true dgp.

In effect, the maintained hypothesis relating to the residual and other unobserved components in the model is always more complex than the hypotheses we can use to describe it in probabilistic terms. In the context of such a view, econometric practice that seeks generality wherever possible while allowing for specificity where it connects most closely to economic ideas seems most desirable. This sounds like (and indeed it is) a strong argument for the use of semiparametric methods in econometrics and, in part, explains the growing popularity of these methods. Automated inference involving HAR procedures precisely fall within this ambit. Also included are methods that allow directly for the parameter space of a conventional model like a vector autoregression to be infinite dimensional, as in Kuersteiner's (2004) contribution to this colloquium.

Suppose we characterize some phenomena that is to be explained (like the temporal dependence or nonstationarity of economic time series) in terms only of the local behavioral characteristics of those series (like the way their spectra behave locally in the vicinity of the origin - what we call local long-run behavior). Further, we may build into this characterization the possibility that certain of the series have comparable and related local long-run behavior, thereby incorporating some economic ideas (like purchasing power parity) into the primitive dgp. Then, without attempting to prescribe a complete dgp for the data, we may seek empirical confirmation of the characteristics we have modeled purely in terms of the local behavior. While we may not have built a complete stochastic model to study the data in their entirety, we can at least attempt to confirm the validity or usefulness of certain underlying economic ideas by doing some local empirical analysis.

One reason for the widespread empirical econometric interest in unit roots and cointegration is that in their general semiparametric form these concepts fit well into the framework of ideas relating to local behavior (Phillips, 1991a & 1991b). The same is true of modern work involving fractional processes (e.g., Robinson, 1995; Kim and Phillips, 1999; Robinson and Hualde, 2003; Phillips and Shimotsu, 2004). Correspondingly, a literature that provides semiparametric approaches to the study of unit roots, cointegration and fractional integration has emerged to meet the needs of practitioners and is becoming popular in applied work. In such contexts, HAR principles are employed to conduct inference, so that only local long-run behavior is assumed in treating the nonparametric components. Data-based automation now plays an important role in the implementation of these methods. In this way empirical investigators are freed from some of the consequences of having to build and rely upon a complete stochastic model to study the data.

Empirical Econometrics and its Future

Automated econometric analysis, which has been made possible by the power of modern computing and electronic data availability, has many natural advantages and conveniences. But automation does not of itself lead to scientifically sound conclusions.

The framework is only as good as the algorithms and the statistical justifications that underly it, the economic ideas that are being incorporated, and ultimately the quality of the data. Inevitably there are shortcomings in automated procedures, in the model classes being utilized in the analysis and in the data being used, just as there are when the more conventional tradition of falsification of a single hypothesis at a time is implemented.

The present paper and those published in this Colloquium only touch the surface of some of these questions. Yet the fact that they are being discussed is itself important and indicates that the goals of econometrics are evolving just as our tools are changing. Right now, econometrics is in its infancy in considering this very wide class of problems in automated specification searches, model construction, validation and inference. While consensus is unlikely in the consideration of the many methodological issues that arise in this process, increasing reliance on computerization and some degree of automation in estimation and inference seem certain to be part of the future of econometrics.

Many good empirical economists appear to believe that they

> "... have a talent for taking a big pile of data, thinking economically about it, and sometimes making conclusions come out the other end." Levitt (2004)

Let it also be said that computers have a truly indefatigable talent for a large part of this task – collecting, processing and analyzing data. Utilizing this talent is a pivotal strength of computer automation in econometrics. Our challenge in econometric theory in this emergent age of automated scientific discovery is to provide guidance mechanisms: automated mechanisms for incorporating economic thinking and methods for adapting for the imperfections and simplifications in that thinking into the empirical model construction process.

Recent experience with automated discovery algorithms in econometrics leads me to believe that these methods will play an important role in the future use of applied econometrics. I also believe that they offer our best current hope of reaching out with our methodology to the wide group of potential practitioners in society who are interested in economic and business forecasts and policy analysis but who are not part of our immediate community of econometrically well-trained professionals. Of course, automation means that such users may proceed without any real understanding of the manner in which critical choices (like bandwidth or lag length selection) have been made in the practical implementation of the econometric software, not to mention the implication of these choices. But present empirical econometric practice reveals that such circumstances are already common in applied work by trained economists. Indeed, the practice is inevitable as econometric software options widen and allow users ready access to advanced procedures on a point and click basis.

No driver's licence of econometric training is currently needed to implement packaged software. Users can implement procedures like HAC estimates or the bootstrap while having little knowledge of what these procedures do or what their properties may be. Even less skill will be needed to use automated econometric methods in packaged form or online as they develop and mature. Such shortcomings are unavoidable. But they are matters that theorists who develop data-driven procedures can largely anticipate; and practitioners who write software packages can think about these matters in advance, implementing devices in software that forewarn or signal users about potential problems. The challenges that are involved in these endeavours will keep the econometrics community busy at many different levels in the years ahead.

As we look forward to the next decade of research in econometrics, it is clear that changing computing technology will continue to play a large role in the evolution of econometrics. More importantly, technology now seems poised to advance the services that the discipline of econometrics can provide to society and these services seem likely to be much more broadly based than in the past. In the second half of the last century, econometrics provided a practical mechanism by which relationships between economic variables could be evaluated, quantified and used to assist in the formulation of economic policy and in the making of economic predictions. Much of the perceived practical benefit to society came through the econometrically informed advice given by economists to government (not all of it at the national level), business, and the financial industry. In recent years, econometric tools have been brought to bear in understanding microeconomic behavior, pricing policies, auctions, regulated markets, and the economic effects of social issues like education policy, environmental pollution, crime and deforestation, to name just a few.

Econometric methods and computer software have developed in part to meet the needs of this growing practical research agenda. As this has occured, tool makers have recognised the need for and inherent advantages in automation. We now have the technology that enables most econometric procedures to be performed online using remote servers that are dedicated to the task. Just as software packages brought econometrics to the desktop and laptop during the 1980s and 1990s, online econometrics now seems capable of bringing econometric methodology and data analysis to the vast community of internet users.

The possibilities for automation in the implementation of econometric methods are already substantial and they seem likely to grow enormously in the next decade as user needs increase, as computer technology advances further, and as our understanding of automated inference methods improves. None of us can anticipate the landscape on the road ahead for the discipline. But as econometrics makes its journey forward, increasing automation in econometric methodology seems likely to become a significant factor in its various practical and public manifestations. This new dimension of econometrics is something theorists must learn much more about. “I ran a million regressions” is no longer simply a wisecrack. It is a practical reality that we have to live with and understand.

References

Bai, J. and S. Ng (2001). "Determining the Number of Factors in Approximate Factor Models," *Econometrica, 70, 191-223.*

Bauer, D. (2004). "Subspace algorithms." mimeo, TU Wien.

Bauer, D. and M. Wagner (2002). "Estimating cointegrated systems using subspace algorithms". Journal of Econometrics, 111, 47-84.

Bauer, D. and M. Wagner (2003). "Canonical form for unit root processes in the state space framework". mimeo, TU Wien.

Cartwright, N. (1999). *The Dappled World. A Study of the Boundaries of Science.* Cambridge University Press.

Chrystal, K. A. and P. D. Mizen (2001). "Goodhart's law: its origins, meaning and implications for monetary policy". Chapter 8 of *Central Banking, Monetary Theory and Practice: Essays in Honour of Charles Goodhart*, Volume 1 (ed. Paul Mizen), Cheltenham: Edward Elgar, 2003, pages 221-243.Bank of England.

Danilov, D. L. and J. R. Magnus (2004a). "On the harm that ignoring pretesting can cause". *Journal of Econometrics* (forthcoming).

Danilov, D. L. and J. R. Magnus (2004b). "Forecast accuracy after pretesting with an application to the stock market". *Journal of Forecasting*

Den Haan, W.J., and A. Levin (1997), "A practitioner's guide to robust covariance matrix estimation," in *Handbook of Statistics* **15**, G.S. Maddala and C.R. Rao, eds., Elsevier (Amsterdam), pp.299-342.

Driscoll, J. C. and A. C Kraay (1998). "Consistent covariance matrix estimation with spatially dependent panel data". Review of Economics and Statistics, 80, 549-560.

Duchesne, P. (2003). "On testing for serial correlation with a wavelet-based spectral density estimator in multivariate time series". Working paper, University of Montreal.

Glymour, C. (2004). "The automation of discovery". *Daedalus*, Winter, 69-77.

Goodhart, C. A. E. (1975). "Monetary relationships: a view from Threadneedle Street". in *Papers in Monetary Economics*, Vol. 1, Reserve Bank of Australia.

Granger, C. J. and D. F. Hendry (2004). "A dialogue concerning a new instrument for econometric modeling". *Econometric Theory* (this issue)

Hansen, B. E. (2004). "Challenges for econometric model selection". *Econometric Theory* (this issue)

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

Hendry, D.F. and H-M Krolzig (1999). "Improving on 'Data Mining Reconsidered' by K.D. Hoover and S.J. Perez." *Econometrics Journal*, 2, 41–58.

Hendry, D. F. and H-M Krolzig (2001). *Automatic Econometric Model Selection.* London: Timberlake Consultants Press.

Hendry, D.F and H-M. Krolzig (2002). "New Developments in Automatic General-to-specific Modelling." In *Econometrics and the Philosophy of Economics,* edited by B.P. Stigum, MIT Press.

Hendry, D. F. and M. S. Morgan (1995). "Introduction", in *The Foundations of Econometric Analysis*, Cambridge University Press.

Hong, Y. (2001). "Wavelet-based estimation for heteroskedastic and autocorrelation consistent variance-covariance matrices". Working paper, Cornell University.

Hoover, K. D. (2004). "Automatic inference of the contemporaneous causal order of a system of equations". *Econometric Theory* (this issue)

Hoover, K. D. (2001) *Causality in Macroeconomics.* Cambridge University Press.

Hoover, K. D. and S. J. Perez (1999). "Data mining reconsidered: encompassing and the general-to-specific approach to specification search". *Econometrics Journal*, 2, 167-191.

Jansson, M. (2004): "Autocorrelation robust tests with good size and power," *Econometrica*,

Judge G. G. and M. E. Bock (1978). *The statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics.* Amsterdam: North Holland.

Kabaila, P. (1995). "The effect of model selection on confidence regions and prediction regions". *Econometric Theory*, 11, 537-549.

Kaboudan, M. A. (2000). "Genetic programming prediction of stock prices". *Computational Economics*, 16, 207-236.

Keuzenkamp, H. A. (2000). *Probability, Econometrics and Truth.* Cambridge University Press.

Kiefer, N. M. and T. J. Vogelsang and H. Bunzel (2000). "Simple Robust Testing of Regression Hypotheses," *Econometrica,* 68, 695-714.

Kiefer N. M. and T. J. Vogelsang (2002a). “Heteroskedasticty-Autocorrelation Robust Testing Using Bandwidth Equal to Sample Size," *Econometric Theory*, 18, 1350-1366.

Kiefer N. M. and T. J. Vogelsang (2002b). “Heteroskedasticty-Autocorrelation Robust Standard Errors Using the Bartlett Kernel without Truncation," *Econometrica*, 70, 2093-2095.

Kim, C-S. and P. C. B. Phillips (1999). “Fully modified estimation of fractional cointegration models”, Working paper, Cowles Foundation, Yale University.

Kuersteiner, G. M. (2004). “Automatic inference for infinite order vector autoregressions”. *Econometric Theory* (this issue).

Leamer, E. E. (1978). *Specification Searches: Ad hoc Inferences with Nonexperimental Data*. New York: John Wiley and Sons.

Lee,. J. and Y. Hong (2001). “Testing for serial correlation of unknown form using wavelet methods”. *Econometric Theory*, 17, 386-423.

Lee, J. S. (1996) “On the power of stationary tests using optimal bandwidth estimates,” *Economics Letters*, 51, 131-137.

Lee, J. S. (1996) “On the power of stationary tests using optimal bandwidth estimates,” *Economics Letters*, 51, 131-137

Leeb, H. and B. M. Pötscher (2003). “The finite sample distribution of post-model-selection estimators and uniform versus nonuniform approximations”. *Econometric Theory*, 19, 100-142.

Leeb, H. and B. M. Pötscher (2003b). “Can one estimate the conditional distribution of post model selection estimators”. Cowles Foundation Discussion Paper #1444.

Leeb, H. and B. M. Pötscher (2004). “Model Selection and inference: facts and fiction”. *Econometric Theory* (this issue).

Levitt, S. (2004). “Interview: The Really Dismal Scientist”. *Wired Magazine*, page 56.

Linton, O. (2004). “Nonparametric inference for unbalanced time series data”. *Econometric Theory* (this issue).

Lucas, R. E. (1976). “Econometric policy evaluation: a critique”, in K. Brunner and A. H. Melzer (eds.), “The Phillips Curve and Labor Markets”, *Journal of Monetary Economics*, Supplement, Special Issue, 1, 19-46.

Malinvaud, E. (1966). “The Statistical Methods of Econometrics”. Amsterdam: North Holland.

Milev, J. (2004). “Search for a structural specification of the earnings-returns relation”. Yale University working paper.

Paruolo, P. (2004). “Automated inference and the future of econometrics: a comment”. *Econometric Theory* (this issue)

Pearl, J. (2000). *Causality: Models, Reasoning and Inference.* Cambridge University Press.

Perez-Amaral, T., G. M. Gallo, and H. White (2004). “A comparison of complementary automatic modeling methods: RETINA and PcGets”. *Econometric Theory* (this issue).

Pesaran, M. H. and A. Timmermann (1995). “The robustness and economic significance of predictability of stock returns”, *Journal of Finance*, 50, 1201-1228.

Pesaran, M. H. and A. Timmermann (2000). “A recursive modeling approach to predicting UK stock returns”. *The Economic Journal*, 110, 159-191.

Pesaran, M. H. and A. Timmermann (2004). “Real time econometrics”, *Econometric Theory* (this issue).

Phillips, P. C. B. (1991a). “Optimal inference in cointegrated systems,” *Econometrica* 59, 283–306.

Phillips, P. C. B. (1991b). “Spectral regression for cointegrated time series.” In W. Barnett, J. Powell and G. Tauchen (eds.), *Nonparametric and Semiparametric Methods in Economics and Statistics, 413-435.* New York: Cambridge University Press.

Phillips, P. C. B. (1992). “Bayes Methods for Trending Multiple Time Series with an Empirical Application to the US Economy”. Cowles Foundation Discussion Paper #1025.

Phillips, P. C. B. (1995a). “Automated forecasts of Asia-Pacific economic activity”. *Asia-Pacific Economic Review*, 1 (1): 92 - 102.

Phillips, P. C. B. (1995b). “Bayesian model selection and prediction with empirical applications”. *Journal of Econometrics*, 69: 289 - 331.

Phillips, P. C. B. (1996). “Econometric model determination’. *Econometrica*, 64, 4: 763 - 812.

Phillips, P. C. B. (1998). “Impulse response and forecast error variance asymptotics in nonstationary VARs,” *Journal of Econometrics*, Vol. 83, 1998, pp. 21-56.

Phillips, P. C. B. (1998b). “Econometric Analysis of Nonstationary Data”. *IMF Lectures,* Cowles Foundation for Research in Economics.

Phillips, P. C. B. (2003). “Laws and limits of econometrics’. *Economic Journal, 113, C26-C52.*

Phillips, P. C. B. (2004). “HAC estimation by automated regression”. *Econometric Theory (* this issue).

Phillips, P. C. B. and W. Ploberger (1994). “Posterior odds testing for a unit root with data–based model selection,” *Econometric Theory* 10, 774–808.

Phillips P. C. B. and W. Ploberger (1996). “An Asymptotic Theory of Bayesian Inference for Time Series”, *Econometrica,* 64, 381-413.

Phillips, P. C. B. and K. Shimotsu (2004). “Local Whittle Estimation in Nonstationary and Unit Root Cases”, *Annals of Statistics* (forthcoming)

Phillips, P. C. B., Y. Sun and S. Jin (2003): “Consistent HAC Estimation and Robust Regression Testing Using Sharp Origin Kernels with No Truncation,” Cowles Foundation Discussion Paper No. 1407

Phillips, P. C. B., Y. Sun and S. Jin (2003a). “Consistent HAC Estimation and Robust Regression Testing Using Sharp Origin Kernels with No Truncation,” Cowles Foundation Discussion Paper No. 1407

Phillips, P. C. B., Y. Sun and S. Jin (2003b). “Long run variance estimation using steep origin kernels without truncation”. Cowles Foundation Discussion Paper No. 1437.

Popper, K. (1959). *The Logic of Scientific Discovery.* London: Hutschinson.

Popper, K. (1963). *Conjectures and Refutations.* London: Routledge and Kegan Paul.

Pötscher, B. M. (1991). “Effects of model selection on inference”. Econometric Theory, 7, 163-185.

Robinson, P. M. (1995). “Gaussian semiparametric estimation of long range dependence”, *Annals of Statistics*, 23, 1630–1661.

Robinson, P. M. (2004). “Robust covariance matrix estimation”. *Econometric Theory* (this issue)

Robinson, P.M. and J. Hualde (2003), "Cointegration in Fractional Systems with Unknown Integration Orders", *Econometrica* 71, 1727-1766.

Sala-i-Martin, X. X. (1997). "I just ran two million regressions". *American Economic Review*, 87, 178-183.

Stock, J. H. and M. W. Watson (1999). "Forecasting inflation". *Journal of Monetary Economics*, 44, 293-335.

Sul, D., P. C. B. Phillip and C-Y. Choi (2003). "Prewhitening bias in HAC estimation". Cowles Foundation Discussion Paper #1436, Yale University.

Swanson, N. R. and C. W. J. Granger (1997). "Impulse response functions based on a causal approach to residual orthogonalization in vector autoregressions". *Journal of the American Statistical Association*, 92, 357-367.

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

Zellner A. (2004). Statistics, Econometrics and Forecasting. Cambridge: Cambridge University Press.

Zellner, A., H. A. Keuzenkamp, and M. McAleer (2001). *Simplicity, Inference and Modelling: Keeping it Sophisticatedly Simple.* Cambridge: Cambridge University Press.