RMM Vol. 2, 2011, 115–145
Special Topic: Statistical Science and Philosophy of Science
Edited by Deborah G. Mayo, Aris Spanos and Kent W. Staley
http://www.rmm-journal.de/

David F. Hendry

Empirical Economic Model Discovery and Theory Evaluation[*]

Abstract:
Economies are so high dimensional and non-constant that many features of models cannot be derived by prior reasoning, intrinsically involving empirical discovery and requiring theory evaluation. Despite important differences, discovery and evaluation in economics are similar to those of science. Fitting a pre-specified equation limits discovery, but automatic methods can formulate much more general initial models with many possible variables, long lag lengths and non-linearities, allowing for outliers, data contamination, and parameter shifts; then select congruent parsimonious-encompassing models even with more candidate variables than observations, while embedding the theory; finally rigorously evaluate selected models to ascertain their viability.

1. Introduction

In 1660, the Royal Society of London was founded as a 'Colledge for the Promoting of Physico-Mathematicall Experimentall Learning', with the intent of bridging the gap between theory and evidence that had persisted since Plato versus Aristotle through to Galileo versus Gilbert (for an excellent discussion, see Goldstein 2010). That gap remains vast to this day in economics, where any attempt to analyze data outside a pre-specified formal model is dismissed as 'measurement without theory', following Koopmans (1947). The most extreme manifestations come from those who espouse 'real business cycle' models, like Kydland and Prescott (l991), essentially asserting that the only role for evidence is to quantify the so-called 'deep parameters' of a mathematical model. Next most extreme are empirical analyses based on claimed 'stylized facts', an oxymoron in the non-constant world of economic data.

Many features of empirical economic models cannot be derived from theory alone: no matter how good that theory may be, it is certain to be an incomplete representation of even key features of 'reality', and will undoubtedly be

[*] This research was supported in part by grants from the Open Society Institute and the Oxford Martin School. I am greatly indebted to Gunnar Bårdsen, Peter Burridge, Julia Campos, Jennifer L. Castle, Jurgen A. Doornik, Guillaume Chevillon, Neil R. Ericsson, Søren Johansen, Katarina Juselius, Hans-Martin Krolzig, Deborah Mayo, Grayham E. Mizon, Bent Nielsen, Duo Qin, James Reade and Aris Spanos for many helpful comments, including on earlier drafts of the book on which this paper is based.

replaced by a better theory in future. Thus, imposing an economic theory on data will almost inevitably mislead. In practice, one needs institutional and empirical evidence on which variables are actually relevant (namely, the complete set of substantive determinants), their lagged responses (if time series or panel data) or dependence (if cross-section or panel data), the functional forms of all connections (any non-linearities), the simultaneity or exogeneity of the 'explanatory' variables or 'instruments', the formation of expectations (where relevant), and the data measurement accuracy, *inter alia*. Most importantly, while some economic theories have begun to address unit roots, almost none incorporate the intermittent occurrence of unanticipated structural breaks (the two central forms of non-stationarities in economies). All these aspects have to be data-based on the available sample while maintaining theory insights, so econometricians must *discover* what matters empirically, then stringently evaluate their findings, hence need methods for doing so.

There are large literatures on the history and philosophy of science examining the process of discovery, primarily in experimental disciplines, but also considering observational sciences. Below, we discern seven common attributes of discovery, namely, the pre-existing *framework of ideas*, or in economics, the theoretical context; going *outside* the existing world view, which is translated into formulating a general initial model; a *search* to find the new entity, which here becomes the efficient selection of a viable representation; criteria by which to *recognize* when the search is completed, or here ending with a well specified, undominated model; *quantifying* the magnitude of the finding, which is translated into 'accurately' estimating the parameters of the resulting model; *evaluating* the discovery to check its 'reality', which becomes testing new aspects of the findings, and perhaps evaluating the selection process itself; and finally, *summarizing* all the available information, so here we seek parsimonious models. *Section 2* discusses the general notions of scientific discovery, then *sections 3 and 4* apply these to discovery in economics and econometrics respectively.

Difficulties of empirical implementation do not detract from the invaluable role that abstract theoretical formulations play in understanding economic behavior, merely that more is required for a useful empirical model than simply imposing a theory, which, of necessity and by design, deliberately abstracts from many complications. Rather than being imposed, theory formulations should be retained when modeling as part of the process of evaluating them, jointly with discovering what additional features are substantively relevant. Empirical model discovery and theory evaluation therefore involves 5 key steps:

(I) specifying the *object* for modeling, usually based on a prior theoretical analysis in economics;

(II) defining the *target* for modeling by the choice of the variables to analyze, denoted $\{\mathbf{x}_t\}$, again usually based on prior theory;

(III) embedding that target in a general unrestricted model (GUM), to attenuate the unrealistic assumptions that the initial theory is correct and complete;

(IV) searching for the simplest acceptable representation of the information in that GUM;

(V) rigorously evaluating the final selection: (a) by going outside the initial GUM in (III), using standard mis-specification tests for the 'goodness' of its specification; (b) applying tests not used during the selection process; and (c) by testing the underlying theory in terms of which of its features remained significant after selection.

To illustrate these steps in turn, consider an observable variable, y, which is postulated to depend on a set of m candidate 'explanatory' variables $\mathbf{z}$, when a sample of T observations is available, denoted $\{\mathbf{x}_t\} = \{y_t, \mathbf{z}_t\}$. An economic analysis suggests that:

$$y = f(\mathbf{z}) \tag{1}$$

Then, (1) is the *object* (I) for the empirical modeling exercise, the relationship about which empirical knowledge is sought. The choice of $\{\mathbf{x}_t\}$ is usually determined by the theoretical analysis depending on its purposes—testing theories, understanding empirical evidence, forecasting possible future outcomes, or conducting economic policy analyses–but as this aspect is subject-matter specific, it is not addressed explicitly here. About the simplest example of (1) is a conditional linear regression $\mathsf{E}[y_t|\mathbf{z}_t] = \beta'\mathbf{z}_t$:

$$y_t = \beta'\mathbf{z}_t + \epsilon_t \tag{2}$$

where β is assumed constant and $\mathbf{z}_t$ is treated as exogenous, with $\epsilon_t \sim \mathsf{IN}[0, \sigma^2_\epsilon]$, denoting an independent normal random variable with mean $\mathsf{E}[\epsilon_t] = 0$ and constant variance σ^2_ϵ.

Second, a common approach is to fit (2) to the data, *imposing* that theory on the evidence. However, the form of $f(\cdot)$ in (1) depends on a range of possible theory choices of (e.g.) the utility or loss functions of agents, the precise formulations of the constraints they face, the information they possess, and the unknown effects of aggregation across heterogeneous individuals with differing choice sets and different parameter values (see e.g., Hildebrand 1994). Moreover, there is no exact specification of a unit of time, so successive observations are generally dependent, and lag responses are not known, again possibly differing across agents. The quality of the observed data is never perfect, so observations may be contaminated, leading to outliers. Nor are the underlying processes stationary, with evolutionary changes ongoing, leading to integrated series, and abrupt shifts inducing various breaks: economies have changed out of all recognition over the past millennium. Thus, many key features of empirical models are bound to be unknown at the outset of an investigation, however good the prior theory. Consequently, *section 5* addresses the basis for (II), namely the derivation of the data-generating process (DGP) in the space of the $m+1$ variables $\{\mathbf{x}_t\}$ being modeled, which is the joint density $\mathsf{D}_{\mathbf{x}}(\mathbf{x}_1 \ldots \mathbf{x}_T)$, called the local DGP (LDGP). All investigators wish to locate the DGP, rather than the LDGP, but given the set of variables chosen for analysis, the best that can be achieved is their LDGP: that LDGP becomes the *target* for any modeling exercise as $\mathsf{D}_{\mathbf{x}}(\cdot)$ contains all the relevant information. The theory-based object and LDGP target are not only related by the choice of $\{\mathbf{x}_t\}$, the DGP is the outcome of the agents' actions about which (1) theorizes, and which may in turn even change those actions. A complete and correct theory would perfectly characterize $\mathsf{D}_{\mathbf{x}}(\cdot)$, which would correspond to the DGP itself (or at least a conditional variant thereof), but absent such omniscience, only the *form* of (2) is given by the theory whereas its *properties* are determined by the LDGP. Other choices of what variables to analyze will create different LDGPs, and the theory of reduction assesses the resulting losses of information (see e.g., Hendry 2009). Poor choices of $\{\mathbf{x}_t\}$ can lead to non-constant LDGPs that are difficult to model or interpret. However, an LDGP can always be formulated with an innovation error, and can often be expressed with constant parameters after 'correcting' for location shifts, so the main evaluation tests discussed below require extending the information set to check if any additions also matter.

Generally the choice of $\{\mathbf{x}_t\}$ entails reductions from the DGP, but even if there was no loss of relevant information, the LDGP generating $\{\mathbf{x}_t\}$ would still require to be *modeled*. Our approach to (III) is to embed the putative LDGP in a much larger formulation that allows for (a) other potentially relevant candidate variables, (b) longer lags than might first be anticipated, (c) a wider range of possible functional forms than linearity, as well as (d) multiple location shifts and possible outliers that may contaminate the available data or induce parameter non-constancy. *Section 6* considers these four extensions in turn. Because change is such a pervasive feature of economies, any substantive mistakes in a specification can seriously distort an empirical analysis. Consequently, all empirical modeling complications must be addressed *jointly* if a useful model is to result. The long history of failed macro-econometric models of all types attests to the past inability to successfully confront this fundamental difficulty. Fortunately, recent developments can allow for much more general formulations, automatically creating extensions for longer lags, polynomial and exponential functional forms, and multiple breaks.

A large GUM forces the need for stage (IV), to efficiently search for the simplest acceptable representation. Here we face what look like two almost insuperable difficulties: (i) the generality of the specification at stage (III) will involve very large numbers of variables, denoted N, posing a challenge for finding the LDGP; (ii) N will almost always be larger than the number of observations, T. As *section 7* explains, automatic selection algorithms save the day. First, theoretical analyses and simulation findings of their properties confirm that the costs of search are small even for large N. Secondly, by conducting a mixture of contracting (general-to-simple) and block expanding searches, all candidate variables can be considered subject to the minimal requirement that fewer, n, matter substantively than T. By embedding the theory specification within the GUM, and not selecting over its components, such an approach allows one to *discover* what matters empirically, find the simplest acceptable representation, and simultaneously evaluate the theory. Thus, the m theory variables $\mathbf{z}_t$ in (2) are *retained* during the search, while the remaining $N - m$ are selected over. This respects, but does not impose, the theory: the final selection may reveal that only some, or perhaps none, of the $\mathbf{z}_t$ are significant, or that their coefficient estimates have uninterpretable signs or magnitudes, disconfirming (2) even if direct estimation had appeared supportive.

These methods are not a replacement for, but an extension of and an improvement upon, many existing practices in empirical economics. The basic framework of economic theory has offered far too many key insights into complicated behaviors to be lightly abandoned, and has made rapid progress in a large number of areas from auction theory through mechanism design to asymmetric information, changing our understanding. But that very evolution makes it unwise to impose today's theory on data—as tomorrow's theory will lead to such 'evidence' being discarded. Thus, one must walk a tightrope where falling on one side entails neglecting valuable theory, and on the other imposing what retrospectively transpire to be invalid restrictions. Empirical model discovery with theory evaluation seeks to avoid both slips. The available theory is embedded at the center of the modeling exercise to be retained when it is complete and correct; but by analyzing a far larger universe of possibilities, aspects absent from that theory can be captured. There are numerous advantages as Hendry and Johansen (2010) discuss. First, the theory is retained when valid. Second, it should be rejected when it is invalid. Third, it could be rescued if the more general setting incorporated factors the omission of which would otherwise have led to rejection. Fourth, a more complete picture of both the theory and confounding influences can emerge, which is especially valuable for policy analyses. Fifth, well-specified selections avoid reliance on doubtful assumptions about the sources of problems like residual autocorrelation or residual heteroskedasticity—which may be due to breaks or data contamination—so that 'corrections' thereof in fact fail to achieve valid inference. Finally, explaining the findings of rival models by encompassing reduces the proliferation of contending explanations, which would create major uncertainties if unresolved. Consequently, little is lost and much is gained by embedding theory in general formulations.

Having found an acceptable parsimonious selection, a warrant is needed to establish the 'reality' of the discovery, (V), by stringent evaluation of the findings, so *section 8* considers that aspect. *Section 9* clarifies the implications of the analysis for empirical model discovery in economics and concludes.

2. Scientific Discovery

A discovery entails learning something previously unknown. Since one cannot know how to discover what is not known, there is unlikely to be a 'best' way of doing so. That does not preclude some ways being better than others—not looking is rarely a good way. Nevertheless, over the last five centuries the natural and biological sciences have made huge advances, both theoretical and empirical, with sequences of discoveries. From the earliest written records of Babylon through ancient Egypt and the Greece of Pericles, discoveries abounded in many embryonic disciplines from astronomy, geography, mathematics, and philosophy to zoology. While fortune favored prepared minds, discoveries were often fortuitous or serendipitous. The 'scientific method', developed in the Arabic world during the early Middle Ages with the works of scholars like Al-Biruni and Ibn Sina (Avincenna), and formalized in the UK by Roger Bacon (see e.g., Hackett 1997), was a major advance, as it highlighted where gaps in knowledge existed, delivered a systematic approach to filling those gaps, and in due course consolidated the findings in general theories and formulae. Even so, in both the natural and biological sciences, most imaginable ways of discovering have proved successful in some settings: see Mason (1962), Messadié (1991) and compare Popper (1959). Advancing an intellectual frontier essentially forces going from the simple (current knowledge) to the more general (adding new knowledge). As a model-building strategy, simple to general is fraught with difficulties (see e.g., Anderson 1962. Campos, Ericsson and Hendry 2005 provide an overview), so it is not surprising that scientific discoveries are hard earned.

The progressivity of science, cumulating empirical findings that cohere with theoretical ideas, is its most salient attribute. We clearly understand vastly more than the ancient or medieval worlds: electricity lights our homes and streets (see Fouquet and Pearson 2006 on the huge increases in lumens consumed since 1300), computers calculate, planes fly, etc. As noted in Hendry (2009), we can predict what changes to chips will, or will not, speed up calculations, and what aircraft designs will not fly. The path that leads to a scientific discovery is irrelevant to its validity, and could be serendipity, careful testing, or a theory prediction, whereas stringent evaluation and replicability are crucial. Nevertheless, theories are rarely rejected by evidence alone, and are only replaced when 'better' theories develop that explain more and account for some previous anomalies (see e.g., Kuhn 1962; Lakatos and Musgrave 1974).

Luck, and its close relative serendipity, are often cited as sources of discovery: examples include Alexander Fleming's discovery of penicillin (see, e.g., Henderson 1997), Henri Becquerel's discovery of radioactivity, for which he shared the Nobel Prize with Pierre and Marie Curie,[1] and more recently, Arno Penzias and Robert Wilson uncovering the background cosmic microwave radiation.[2] In the first two cases, and even more so with, say, Archimedes 'Eureka' discovery, recognition of the significance of what is found is also crucial (i.e., why the rise in his bath water allowed the assessment of an object's density). However, brilliant intuition can also succeed, as with Michael Faraday's dynamo (see e.g., Holton 1986), as can systematic experimental exploration of all the alternatives, illustrated by Antoine Lavoisier isolating and establishing the properties of oxygen, thereby finally refuting phlogiston theory (see e.g., Musgrave 1976), or Robert Boyle's law of gases (see e.g., Agassi 1977).

[1] URL: http://nobelprize.org/nobel_prizes/physics/laureates/1903/becquerel-bio.html.

[2] URL: http://nobelprize.org/nobel_prizes/physics/laureates/1978/wilson-lecture.html.

Great experiments have clearly advanced scientific understanding (see Harré 1981), but discovery has also been driven both by false theories, as with Johannes Kepler's attempts to characterize the planetary orbits by regular solids, nevertheless leading to his famous laws (see, e.g., Holton 1988), and by 'correct' theories, as with the often-cited example of Isaac Newton's theory of universal gravitation and Arthur Eddington's test of Albert Einstein's theory of relativity by the gravitational effects of the sun bending light (see his own account in Eddington 1928), or Louis Pasteur's germ theory of disease leading to pasteurization and the rejection of Aristotle's notion of 'spontaneous generation' (despite objections that were seemingly valid at the time from a lack of knowledge of thermophilic bacteria: see Walker 2002). Conversely, careful observation led to William Harvey's model of the circulation of blood (see Schultz 2002), to John Snow's tracking down the water borne source of cholera (see e.g., Smith 2002, but contrast McLeod 2000), and to Edwin Hubble's discovering that light from distant astronomical objects was 'redshifted' in proportion to their distance (see, e.g., Nussbaumer and Bieri 2009).

Moreover, the invention of new instruments enabled Galileo Galilei's discovery of the moons of Jupiter by a telescope (see Drake 1980), and of microbes by Robert Hooke and Antonie van Leeuwenhoek using microscopes (see, e.g., Gest 2002; Bennet, Cooper, Hunter and Jardine 2003). The 'natural experiment' of the Second World War reduced, then its termination raised, wheat consumption in the Netherlands, which first dramatically lowered then raised the death rate of young sufferers of celiac disease, and so led to the identification of gluten as the cause (see Fasano 2009). Often 'self testing' was involved, most recently with Barry Marshall drinking *Helicobacter pylori* to demonstrate that they caused peptic ulcers, followed by antibiotics to show the cure.[3] Finally, trial and error on a vast scale was Thomas Edison's route to producing a workable incandescent lamp (see Nelson 1959; Lomas 1999). Other examples abound over time and across countries: science is systematic only in retrospect.

Science is a deductive, not an inductive, discipline in the important sense articulated by Herschel (1830) in his distinction between the context of discovery, which we have just discussed, and the context of evaluation, later re-emphasized by Popper (1963). Empirical findings remain anomalies until situated within a theory; and science seeks an interlinked system of theories that mutually support the interpretations of evidence: radioactivity, dating fossils, plate tectonics, geological time frames, and fMRI scanners are a classic instance. As noted above, theories are abandoned only when a new theory can cover most of the existing ground and explain some new phenomena, albeit that many empirical discoveries have led to changes in theory. The consolidation of evidence also plays a crucial role: the most famous is Einstein's $E = Mc^2$, which summarizes a remarkable amount in a simple formula (see e.g., Farmelo 2002). Without that stage, the mass of data would overwhelm by the huge costs of knowledge con-

[3] Humourously recounted in http://nobelprize.org/nobel_prizes/medicine/laureates/2005/marshall-lecture.pdf.

sumption (Sims 1996 argues that data reduction is a key attribute of science: also see Friedman 1974).

Despite the diversity in how discoveries were achieved, from theory through evidence to luck or chance, there are seven aspects in common to the above examples. First, the *theoretical context*, or more generally, the pre-existing framework of ideas, which may inhibit progress (phlogiston is a classic example), or be a stimulus (as with quantum theory). Secondly, going beyond, or *outside*, the existing state, by greater generality, a new tool, a chance innovation, or a broader idea or perspective. Thirdly, the *search* for something: what is found may not have been the original aim, though it certainly was on some occasions, but there was an objective from the outset to be discovered. Fourthly, *recognition* of the significance of what is found: the discovery usually relied in part on fortune favoring the prepared mind. Fifthly, *quantifying* what is found, by new measurements or experiments. Sixthly, rigorously *evaluating* the discovery to ascertain its 'reality', sometimes by checking replicability, sometimes by testing in new settings. Finally, *parsimoniously summarizing* all the available information.

3. Discovery in Economics

While the literatures on the history and philosophy of science provide invaluable background, social sciences confront uniquely difficult modeling problems of high dimensional, non-linear, inertial yet evolving systems, with intermittent and often unanticipated abrupt changes. Social sciences also make discoveries, but historically in economics, most discoveries have come from theoretical advances rather than empirical findings, as histories of economic thought from Schumpeter (1954) to Blaug (1980) emphasize. Current theoretical approaches tend to derive behavioral equations from 'rational' postulates, assuming optimizing agents with different information sets who face various constraints. Many important developments have been achieved by such analyses, particularly in understanding individual and firm behavior in a range of settings. Nevertheless, the essentially unanticipated financial crisis of the late 2000s has revealed that aspects of macroeconomics have not been well represented by models based on single-agent theories, nor has a timeless theory proved well adapted to the manifest non-stationarities apparent in economic time series. The crucial differences induced by changes in behavior and their feedback onto outcomes must be accounted for.

3.1 Change and Its Consequences

There are two fundamental differences between discoveries in social and physical sciences: the non-permanence of phenomena, and feedbacks of discoveries onto behavior. Light waves were 'bent' in strong gravitational fields millennia ago, are now and will be millennia in the future; many relationships in economics were very different 1000 years ago from now, and probably will be different again in 1000 years. Discoveries in astronomy can also be transient—a comet that visits but once—and even appear unexpectedly, as with supernovae, so the issue is not unique, but it is important as economies are intrinsically nonstationary. Breaks are often viewed as a major problem for empirical models, but are also a serious difficulty for theories in economics, few of which allow for unanticipated sudden large shifts occurring intermittently (see Hendry and Mizon 2010), including the recent financial crisis. Surprisingly, as discussed in §6.3, using automatic empirical modeling methods that can detect and remove multiple location shifts entails that breaks may be handled more easily empirically (albeit *ex post*), than by theories. Moreover, it is essential to do so: unanticipated shifts of distributions are pernicious as they lead to non-ergodic data, so no form of inference will be feasible until effective ways can be found of 'reducing' the problem to one where (say) martingale-difference inputs can be induced, as in the earlier setting of sequential factorization of dynamic processes by Doob (1953).

The other issue, that discoveries in economics change economic reality, has a long history. Economic arguments brought about free trade; and option pricing theory is now widely used in practice. Thus economics itself induces change, often in unexpected ways, and makes the modeling of change a key issue in discovery, by seeking to uncover the less transient aspects of economic behavior. It is simply impossible to treat the economy as a stationary process, even after removing stochastic trends (unit roots) because distributions shift. Unanticipated changes must also impact substantively on how expectations are formed by economic actors, and hence on how to analyze their behavior: it is unclear how agents can form 'sensible' expectations about future events when shifts occur. Hendry and Mizon (2010) demonstrate that so-called 'rational expectations' based on previous conditional expectations are not rational in any sense, as they are neither unbiased nor minimum mean-square error predictors once distributions shift. Indeed, inter-temporal theory calculations in economics also fail in that case, as the law of iterated expectations across time periods does not hold when the relevant integrals are over different distributions.

3.2 Implications for Prior-based Analyses

'Prior distributions' widely used in Bayesian analyses, whether subjective or 'objective', cannot be formed in such a setting either, absent a falsely assumed crystal ball. Rather, imposing a prior distribution that is consistent with an assumed model when breaks are not included is a recipe for a bad analysis in macroeconomics. Fortunately, priors are neither necessary nor sufficient in the context of discovery. For example, children learn whatever native tongue is prevalent around them, be it Chinese, Arabic or English, for none of which could they have a 'prior'. Rather, trial-and-error learning seems a child's main approach to language acquisition: see Clark and Clark (1977). Certainly, a general language system seems to be hard wired in the human brain (see Pinker 1994; 2002), but that hardly constitutes a prior. Thus, in one of the most complicated tasks imaginable, which computers still struggle to emulate, priors are not needed.

Conversely, priors are insufficient to facilitate discovery unless 'correct', and historically, false priors have been a bugbear of progress: witness the intellectual battles of the past, from Copernicus versus an Earth centered Church; or Thompson (1862; 1864) arguing against Darwin that the Earth could not be very old, so evolution could not have had time to happen (despite the overwhelming evidence from geology that he was wrong), a view he did not correct till Thomson (1899); or Fleeming Jenkin 'proving' that evolution must be convergent, not divergent, because 'continuous characteristics blend'—yet anyone could see that the sex of a person was a discrete characteristic and did not blend. As noted above, the pre-existing framework of ideas is bound to structure any analysis for better or worse, but being neither necessary nor sufficient, often blocking, and unhelpful in a changing world, prior distributions should play a minimal role in data analyses that seek to discover useful knowledge.

4. Covert Discovery in Empirical Econometric Research

Histories of econometrics, such as Morgan (1990) and Qin (1993), also focus on its theoretical advances, and while they discuss applied research as well, are not filled with major empirical discoveries that have stood the test of time or altered the course of economic analysis. Although the notion of empirical model discovery in economics may seem to be a marked departure, it is a natural evolution from existing practices: much of previous econometrics has covertly concerned discovery. Despite the paucity of explicit research on empirical model discovery, there are large literatures on closely related material (see Spanos 1990; 2006 for a related view). Classical econometrics focused on obtaining the 'best' parameter estimates, given the correct specification of a model and an uncontaminated sample, yet also delivered a vast range of tests to check the resulting model—to discover if it was indeed well specified. Model selection methods extended that remit to find the subset of relevant variables and the associated parameter estimates, again assuming a correct nesting set, so sought to discover the key determinants of the variables being modeled. Non-parametric methods concern discovering the functional form, and time-series 'model identification' is discovering which model in a well-defined class best characterizes the available data. None of these was framed as discovery, and each approach depended on many assumptions about the validity of their chosen specification, often susceptible to empirical assessment, and by evaluating it later, proceeded from the specific to the general. The following five-fold distinction helps summarize the assumptions of the main different approaches, for the simple case of a regression equation, since generalizations to other model classes are fairly obvious. Even given the prevalence of serendipity in discovery, the 'strategy' of 'data mining' till 'pleasing' results are obtained has little to commend it, as parodied by Leamer (1983) and Spanos (2000).

4.1 Classical Econometrics

Here it is postulated that there is a relation with a constant parameter β:

$$y_t = \beta' \mathbf{g}(\mathbf{z}_t) + \epsilon_t, \quad t = 1, \ldots, T \tag{3}$$

after known data transformations $\mathbf{g}(\cdot)$ (such as logarithms) which make linearity in β reasonable. The aim is to obtain the 'best' estimate of β, assuming the complete and correct variables, $\mathbf{z}$, and an uncontaminated set of observations, $\mathscr{T}$, where $\mathbf{g}(\cdot)$ is known. Auxiliary assumptions often include that $\epsilon_t \sim \mathsf{IID}[0, \sigma^2_\epsilon]$, and perhaps a set of 'instrumental variables' $\{\mathbf{w}_t\}$ (often the $\{\mathbf{z}_t\}$) claimed to be independent of ϵ_t, which determine the choice of estimation method as least squares, instrumental variables, or one of dozens of related methods (see Hendry 1976). Departures from the assumptions of (3) are treated as 'problems' to be solved, such as residual serial correlation or heteroskedasticity, data contamination, outliers, or structural breaks, omitted variables, functional-form misspecification, etc. Most econometrics textbooks provide tests for discovering if these problems are present, often followed by recipes for 'fixing' them, since unless (3) is perfectly pre-specified, all these issues must be resolved from the evidence. Such an approach is covert and unstructured empirical model discovery, with investigators patching their specifications to avoid the most egregious flaws, often reporting estimates as if they were the first attempt.

4.2 Classical Model Selection

Although the starting point is a model like (3), again given the correct initial $\mathbf{z}$, $\mathbf{g}(\cdot)$ and $\mathscr{T}$, now $\mathbf{z}$ includes a set of candidate regressors, which is anticipated to include all the relevant explanatory variables, their functional forms and lags etc., but perhaps also some irrelevant (or 'small') effects. The aim is to find the subset of relevant variables, $\mathbf{z}^*_t$ say, eliminate the irrelevant, then estimate the associated (constant) parameters, β^*. This setting is more general than §4.1, and the need to discover the relevant subset of variables is explicitly recognized, but auxiliary assumptions may include that $\epsilon_t \sim \mathsf{IID}\left[0, \sigma^2_\epsilon\right]$, with a set of 'instrumental variables' (often the $\{\mathbf{z}_t\}$) assumed independent of ϵ_t, determining the choice of estimation method. As with classical econometrics, departures from the assumptions underlying (3) are usually treated as problems, such as residual serial correlation or heteroskedasticity, structural breaks etc., although some selection methods simply ignore all such problems to select the 'best' model on their given criterion function.

4.3 Robust Statistics

Despite the differences at first sight, the aim is to find a 'robust' estimate of β in (3) by also selecting over $\mathscr{T}$, assuming the correct set of relevant variables $\mathbf{z}_t$. The key focus is avoiding data contamination and outliers, so discovering a sample, $\mathscr{T}^*$, where those are least in evidence. However, other difficulties, such as residual serial correlation or heteroskedasticity, structural breaks, functional-form mis-specification etc., still need to be detected, and $\mathbf{z}$ rarely includes a large set of candidate regressors to be selected over jointly with $\mathscr{T}^*$, so is essentially assumed to be $\mathbf{z}^*$.

4.4 Non-parametric Statistics

The objective is again to estimate β in (3) assuming the correct set of relevant variables $\mathbf{z}$, but seeking to discover $\mathbf{g}(\cdot)$ without assuming a specific mathematical function, and possibly also leaving the 'error distribution' unspecified. As before, all forms of mis-specification need to be checked, as even the most 'non-parametric' formulation may provide a poor approximation to the LDGP (let alone the DGP), and $\mathbf{z}$ rarely includes many candidate regressors, so once more is assumed to be $\mathbf{z}^*$. Data contamination can be pernicious as it distorts the function found, but selection over $\mathscr{T}$ jointly with a non-parametric analysis is uncommon.

4.5 Selecting Jointly

To achieve an approach that will be viable when the exact specification, the error properties, the reliability of the data, the functional form, and the constancy of the parameters all need to be chosen jointly, we must return to basics: how were the data generated?

5. Formulating a 'Good' Starting Point

We conceptualize the data-generating process (DGP) as the joint density of all the variables in the economy. It is impossible to accurately theorize about or precisely model such a high dimensional entity, that is anyway also non-stationary. All empirical (and theoretical) researchers reduce the task to a manageable size by implicitly formulating a 'local DGP' (LDGP), which is the DGP in the space of the $m+1$ variables $\mathbf{x}_t$ being modeled. The theory of reduction (see e.g., Hendry 1995; 2009) explains the derivation of the LDGP, which is the joint density $\mathsf{D}_{\mathbf{x}}(\mathbf{x}_1 \ldots \mathbf{x}_T|\theta)$ for a sample $t = 1,\ldots,T$, where the 'parameter' θ may be time varying. The choice of $\{\mathbf{x}_t\}$ is fundamental, and determines the properties of $\mathsf{D}_{\mathbf{x}}(\cdot)$. Knowing the LDGP, one can generate 'look alike data' for $\{\mathbf{x}_t\}$ that only deviate from the actual data by unpredictable noise–so the LDGP $\mathsf{D}_{\mathbf{x}}(\cdot)$ is the target for model selection, once $\{\mathbf{x}_t\}$ is chosen.

The main reductions are aggregation, marginalization, sequential factorization and conditioning. Marginalizing with respect to variables deemed to be irrelevant *a priori* is a major reduction, possibly hazardous if some of the variables are in fact relevant. Next, sequential factorization of $\mathsf{D}_{\mathbf{x}}(\mathbf{x}_1 \ldots \mathbf{x}_T|\theta)$ to $\prod_{t=1}^{T} \mathsf{D}_{\mathbf{x}_t}(\mathbf{x}_t|\mathbf{x}_{t-1},\ldots,\mathbf{x}_1,\theta_t)$ produces the martingale-difference error $\epsilon_t = \mathbf{x}_t - \mathsf{E}[\mathbf{x}_t|\mathbf{x}_{t-1},\ldots,\mathbf{x}_1]$ (see e.g., Doob 1953): Spanos (1986; 1999) provides an excellent explanation. As $\mathsf{E}[\epsilon_t|\mathbf{x}_{t-1},\ldots,\mathbf{x}_1] = 0\ \forall t$ entails $\mathsf{E}[\epsilon_t|\epsilon_{t-1},\ldots,\epsilon_1] = 0\ \forall t$, then $\{\epsilon_t\}$ is not serially correlated by construction. Conditioning on a subset of contemporaneous variables requires that they be weakly exogenous for the parameters of interest, and preferably super exogenous (see Engle and Hendry 1993) as discussed in §8.1. Failing to handle breaks in the DGP will lead to a non-constant representation, as expectations should be written as $\mathsf{E}_t[\mathbf{x}_t|\cdot]$ when the underlying distributions are non-constant. Although one cannot do better than know $\mathsf{D}_{\mathbf{x}}(\cdot)$, modeling it is a daunting task.

Aggregation over some or all of time, space, commodities, agents, and endowments is also essential to obtain usable economic data, but thereby precludes any claim to 'truth'. Only congruence is on offer in economics, where congruent models match the LDGP in all measured attributes, as congruent triangles match after rotation in 2-dimensions, even if one may actually be the cut-off top of a pyramid. The LDGP is congruent with itself, so non-congruent models are clearly not the LDGP. Empirical congruence in a model is defined by a homoskedastic innovation, ϵ_t, weakly exogenous variables for parameters of interest, that should be constant and invariant for a relevant class of interventions. In addition, selection criteria usually include theory consistent, identifiable structures, with data-admissible formulations on accurate observations that encompass rival models (i.e., account for their results: see Mizon 2003; 2008). Those six requirements exhaustively characterize the null hypotheses to test, and test statistics thereof are essentially independent of the specification tests for model selection based on sufficient statistics (see Mayo 1981), but there are many alternatives against which to seek power to reject false nulls. Congruence is testable and provides necessary conditions for structure, defined as invariance over extensions of the information set across variables, time and regimes. Nevertheless, congruence can be designed by model re-specification, so is far from sufficient to justify a model: *section 8* addresses post-selection evaluation.

Next, one must relate the theory model to that LDGP target, jointly with matching the target to the evidence. Early theories characterized the economy as general equilibrium, but a general sequential dynamic dis-equilibrium would be a better description. Prior reasoning, theoretical analysis, previous evidence, historical and institutional knowledge are all important in avoiding complicated and uninterpretable LDGPs. In wide-sense non-stationary processes, *ceteris paribus* does not apply empirically, so too small a set of variables under consideration may make it impossible to establish constant models that are interpretable by the original theory. Even given a 'good' choice of $\{\mathbf{x}_t\}$, to adequately characterize the resulting LDGP it is crucial not to omit substantively important functions of its variables, such as lags, non-linear transformations, and indicators for breaks, etc. Thus, we embed the target in a general model formulation, which also retains, but does not impose, the theory-based variables. While the lagged reactions in a model corresponding to the sequential factorization need to be data based–as theory specifications of time units are rare–this is easily accomplished and ensures a key attribute for valid statistical inference. Since observations may be contaminated by measurement errors, an approach that is 'robust' against serious data contamination is needed, as is one that also tackles abrupt unanticipated shifts which induce various forms of breaks, so indicators for outliers and location shifts are essential, helping ensure both more constancy and near normality. Appropriate functional form transformations again help with constancy and homoskedasticity. Also, the underlying stochastic processes in economics are usually integrated (denoted I(1)), requiring treatment of stochastic trends as well as using appropriate critical values for inferences reducing to I(0).

As all these aspects must be discovered empirically, model selection is inevitable and ubiquitous, but can be undertaken in a general formulation where valid inference is feasible. To utilize economic analyses when they cannot be imposed empirically, one must embed the theory specification in the general empirical formulation where the theory variables are not selected over, although all other aspects are. Retaining theory variables does not ensure they will be either significant or have their anticipated signs and magnitudes, but if the theory were correct and complete, then Hendry and Johansen (2010) show that the distributions of the estimated parameters of the theory variables would be unaffected by selection. A larger set of variables is less likely to exclude what are in fact important influences, at the possible cost of retaining adventitious effects. These are asymmetric costs: the former is an order one error, the latter of order $1/T$. Thus, it seems preferable to err on the side of profligacy at this stage, and over, rather than under, include, so we consider automatic extensions of initial formulations, then describe the search process in *section 7*.

6. Extensions to Nest the LDGP

Increasing the set of candidate variables to augment $\mathbf{z}_t$ can only be done sensibly by an investigator, as it changes the LDGP that is being modeled. However, three other important extensions of a basic theory model can be created automatically:

(i) functional form transformations for non-linearity;

(ii) longer lag formulation to implement a sequential factorization;

(iii) impulse-indicator saturation (IIS) for parameter non-constancy and data contamination.

Castle and Hendry (2011a) discuss automatically creating approximations to a wide range of functional forms, described in §6.1. Then §6.2 notes creating longer lags, which is straightforward. Hendry, Johansen and Santos (2008), Johansen and Nielsen (2009) and Castle, Doornik and Hendry (2009) discuss automatically generating impulse indicators to saturate the sample with an indicator for every observation. Combining these creates the general unrestricted model (GUM), so we consider *(i)–(iii)* in turn.

6.1 Automatic Non-linear Extensions

Since non-linearity comprises all functions extending linear, there is no guarantee that any specific low dimensional set will nest the LDGP. An optimal approach would include a minimal basis that spanned the relevant space of non-linearity. There are many mathematical series expansions that can approximate any continuous function arbitrarily closely, including polynomials, which is the inital class we consider. Although only low order polynomials will be used, for many variables, their squares and quartics are highly correlated, as are cubes and quintics, so such approximations capture much of the non-linear variation.

A test for non-linearity in a feasible linear GUM is proposed by Castle and Hendry (2010a), using a low-dimensional portmanteau test based on general cubics with exponential functions of the principal components $\mathbf{w}_t$ of the $\mathbf{z}_t$. Let $\widehat{\Sigma}$ denote the $m \times m$ sample correlation matrix of the $T \times m$ candidate variables $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_T)'$, possibly transformed to $\mathsf{I}(0)$ by appropriate differencing. The eigenvalue decomposition of $\widehat{\Sigma}$ is:

$$\widehat{\Sigma} = \widehat{\mathbf{H}}\widehat{\Lambda}\widehat{\mathbf{H}}' \tag{4}$$

where $\widehat{\Lambda}$ is the diagonal matrix of eigenvalues $\{\widehat{\lambda}_i\}$ and $\widehat{\mathbf{H}} = (\widehat{\mathbf{h}}_1, \ldots, \widehat{\mathbf{h}}_m)$ is the corresponding matrix of eigenvectors, with $\widehat{\mathbf{H}}'\widehat{\mathbf{H}} = \mathbf{I}_m$. The sample principal components are computed as:

$$\widehat{\mathbf{W}} = \widehat{\mathbf{H}}'\widetilde{\mathbf{Z}} \tag{5}$$

where $\widetilde{\mathbf{Z}} = (\widetilde{\mathbf{z}}_1, \ldots, \widetilde{\mathbf{z}}_T)'$ are the standardized data, $\widetilde{z}_{j,t} = (z_{j,t} - \overline{z}_j)/\widetilde{\sigma}_{z_j}$ with $\overline{z}_j = \frac{1}{T}\sum_{t=1}^{T} z_{j,t}$, and $\widetilde{\sigma}_{z_j} = [\frac{1}{T-1}\sum_{t=1}^{T}\left(z_{j,t} - \overline{z}_j\right)^2]^{1/2}$, $\forall j = 1, \ldots, m$. Then $u_{1,i,t} = w_{i,t}^2$, $u_{2,i,t} = w_{i,t}^3$, and $u_{3,i,t} = w_{i,t}\exp(-|w_{i,t}|)$ are created. When Σ is non-diagonal, each $w_{i,t}$ is a linear combination of every $z_{i,t}$, so $w_{i,t}^2$ involves squares and cross-products of every $z_{i,t}$ etc. Their test is an F-statistic for the marginal significance of adding the $\{u_{j,i,t}\}$ to the postulated model. There are only $3m$ additional terms for m variables, whereas the number of potential regressors for general cubic polynomials in the $z_{i,t}$ is:

$$N_m = m(m+1)(m+5)/6$$

leading to an explosion in the number of terms as m increases:

m	1	2	5	10	15	20	30	40
N_m	3	9	55	285	679	1539	5455	12300

Thus, one would quickly reach huge N_m, yet $3m = 120$ even at $m = 40$, while allowing for a wide class of functional relations.

They propose proceeding to a non-linear in the parameters formulation, which we denote by $f(\mathbf{z}_t, \theta)$, only when the F-test rejects to avoid possible problems with identifying the parameters of the final specification of the non-linear function (see e.g., Granger and Teräsvirta 1993). Functions like $f(\mathbf{z}_t, \theta)$, are often of an ogive type, such as a smooth-transition regression model (see e.g., Priestley 1981; Chan and Tong 1986 and Teräsvirta 1994) so involve interactions of parameters, some of which will not be identified under a linear null that entails a subset thereof are zero.

Instead of post-testing estimation, however, the $\{u_{j,i,t}\}$ could be included in the GUM initially. After model selection, we then apply an encompassing test against an investigator's preferred functional form $f(\mathbf{z}_t,\theta)$. Once the null of linearity has been rejected for $\{u_{j,i,t}\}$, adding the $f(\mathbf{z}_t,\theta)$ to the final model no longer poses identification issues. A test of the significance of an estimated function, $f(\mathbf{z}_t,\widehat{\theta})$, has three possible outcomes:

a) insignificant, so the LDGP is non-linear, but not of the preferred form;
b) significant, but some of the $\{u_{j,i,t}\}$ are as well, so $f(\mathbf{z}_t,\theta)$ helps provide a more parsimonious but incomplete explanation;
c) significant, and none of the $\{u_{j,i,t}\}$ are, so $f(\mathbf{z}_t,\theta)$ parsimoniously encompasses the approximating model.

Although the last step of such an approach is not general to simple, an advantage is that selecting the $\{u_{j,i,t}\}$ can be combined with impulse-indicator saturation as in §6.3 (to tackle non-normality, outliers, breaks, and possible data contamination) to help avoid non-linear functions inappropriately representing data irregularities as non-linearity (see Castle and Hendry 2011a).

6.2 Creating Lags

Next, automatically create s lags $\mathbf{x}_t \dots \mathbf{x}_{t-s}$ possibly modeled as a system:

$$\mathbf{x}_t = \gamma + \sum_{j=1}^{s} \Gamma_j \mathbf{x}_{t-j} + \epsilon_t \tag{6}$$

Although systems can be handled, we focus here on single equations, so letting $\mathbf{x}_t = (y_t, \mathbf{z}_t)$ formulate the dynamic linear model:

$$y_t = \beta_0 + \sum_{i=1}^{s} \lambda_i y_{t-i} + \sum_{i=1}^{r} \sum_{j=0}^{s} \beta_{i,j} z_{i,t-j} + \epsilon_t \tag{7}$$

6.3 Impulse-indicator Saturation

To tackle multiple breaks and outliers, for T observations add T impulse indicators to the candidate regressor set. To understand how such a 'saturation' can work, consider the simplest setting where $y_i \sim \mathsf{IID}\left[\mu, \sigma_\epsilon^2\right]$ for $i = 1, \dots, T$ when μ is the parameter of interest. Being uncertain of outliers, create the T indicators $1_{\{t=t_i\}}$. First, include half the indicators, and record the significant outcomes: doing so is just 'dummying out' $T/2$ observations for estimating μ. Then omit those, and include the other half, recording significant outcomes again. Finally, combine the two sets of recorded sub-sample indicators, and select the significant ones. Under the null of no outliers or breaks, αT indicators will be selected on average at a significance level α. This 'split-sample' impulse-indicator saturation (IIS) algorithm is the simplest implementation, but multiple unequal splits are feasible: see Hendry et al. (2008).

Johansen and Nielson (2009) extend IIS to both stationary and unit-root autoregressions. When the error distribution is symmetric, for example, adding T impulse-indicators to a stationary dynamic regression with r variables, coefficient vector β (not selected over) and population data second moment Ψ, where $\xrightarrow{\mathsf{D}}$ denotes convergence in distribution:

$$T^{1/2}(\widetilde{\beta}-\beta)\xrightarrow{\mathsf{D}}\mathsf{N}_r\left[\mathbf{0},\sigma_\epsilon^2\Psi^{-1}\Omega_\beta\right] \tag{8}$$

Thus, the rate of convergence of $T^{1/2}$ remains, as does consistency, and the usual asymptotic variance matrix of $\sigma_\epsilon^2\Psi^{-1}$. The efficiency of the IIS estimator $\widetilde{\beta}$ with respect to the OLS estimator $\widehat{\beta}$ is measured by the term Ω_β, which depends on α, and on the form of the error distribution. When $T = 100$ at $\alpha = 1/T$, say, $\alpha T = 1\%$, so despite including 100 extra candidate regressors, the distribution of $\widetilde{\beta}$ in (8) is almost identical to that of $\widehat{\beta}$, with a small efficiency loss of one observation 'dummied out' under the null. However, there is the potential for major gains under alternatives of breaks and data contamination.

6.4 Specification of the GUM

Most major formulation decisions are now made: which m basic variables $\mathbf{w}_t$, after transforming $\mathbf{z}_t$; their lag lengths s; functional forms (cubics of the principal components); and possible location shifts (any number, anywhere) leading to the inestimable GUM:

$$\begin{aligned} y_t = &\sum_{i=1}^{r}\sum_{j=0}^{s}\beta_{i,j}z_{i,t-j}+\sum_{i=1}^{r}\sum_{j=0}^{s}\kappa_{i,j}w_{i,t-j}+\sum_{i=1}^{r}\sum_{j=0}^{s}\theta_{i,j}w_{i,t-j}^2+\sum_{i=1}^{r}\sum_{j=0}^{s}\gamma_{i,j}w_{i,t-j}^3 \\ &+\sum_{j=1}^{s}\lambda_j y_{t-j}+\sum_{i=1}^{T}\delta_i 1_{\{i=t\}}+\epsilon_t \end{aligned} \tag{9}$$

Then there are $K = 4m(s+1)+s$ potential regressors, many of which are perfectly collinear, plus T indicators in total, so one is bound to have $N > T$. That raises the crucial question: how can a feasible model be selected from such a massive, perfectly collinear starting point with $N > T$? The next section answers it, noting that if (9) did not nest the associated LDGP, then neither would any special cases thereof.

7. Model Selection 101

There are too many myths about model selection to disabuse them all here, but an explanation of the elements of a generic simplification approach may clarify their excellent behavior. All aspects of model selection, an essential component of empirical discovery, have been challenged, and many views are still extant.

Even how to judge the status of any new entity is itself debated. Nevertheless, current challenges are wholly different from past ones–primarily because the latter have been successfully rebutted (see e.g., Hendry 2000). All approaches to model selection face serious problems, whether selecting on theory grounds, by fit–howsoever penalized–or by search-based methods. A key insight is that, facilitated by recent advances in computer power and search algorithms, one can adopt an extended general-to-specific modeling strategy that avoids many of the drawbacks of its converse. When N exceeds T, a general-to-specific approach ceases to be applicable as there are too many candidate variables for the GUM to be estimated. Nevertheless, the key notion of including as much as possible jointly remains, albeit that expanding searches are required as well as simplifications. One crucial ingredient is not to undertake a forward search adding just one variable at a time from a null model based on (e.g.) the next 'best' choice on the given criterion. A formal approach to model discovery must take account of all the decisions involved in model specification, evaluation, selection, and reduction, such that the end result is a viable representation of the main variables of interest, with known inference properties.

To construct our explanation, first consider a simple version of the model in (9):

$$y_t = \sum_{i=1}^{N} \beta_i z_{i,t} + \epsilon_t \tag{10}$$

where the regressors are mutually orthogonal in population, $\mathsf{E}[z_{i,t}z_{j,t}] = \lambda_{i,i}$ for $i = j$ and $0\ \forall i \neq j$, with $\epsilon_t \sim \mathsf{IN}[0,\sigma_\epsilon^2]$ when $T >> N$. We take N to be large, say 1000, with $T = 2000$ (see Castle, Doornik and Hendry 2011a). The point of this special case is to show both that 'repeated testing' need not occur when selecting by first estimating (10), and that the number of irrelevant variables (garbage) retained can be controlled. We let n of the β_i be non-zero, where n is usually much smaller than N. The corresponding variables are called relevant; those with $\beta_i = 0$ are irrelevant. However, a 'substantively relevant' variable is one that would be statistically significant at a reasonable significance level α, say $\alpha = 0.01$, when the LDGP is estimated. The *potency* of the procedure on a given test is the proportion of retained relevant variables, which should be close to the power of the corresponding test in the estimated LDGP. The *gauge*, g, of the procedure is the average proportion of retained irrelevant variables, which should be close to α if the procedure works well. However, because selection tries to retain only congruent specifications, insignificant variables may also be retained when they happen to offset what would otherwise be an adventitiously significant mis-specification test. Thus, gauge differs from the 'false discovery rate', and potency from power.

After estimating (10), order the N sample t^2-statistics testing each hypothesis H_0: $\beta_j = 0$ as $\mathsf{t}^2_{(N)} \geq \mathsf{t}^2_{(N-1)} \geq \cdots \geq \mathsf{t}^2_{(1)}$ (squaring to obviate considering signs). The cut-off, k, between variables to be included or excluded is defined by $\mathsf{t}^2_{(k)} \geq c^2_\alpha > \mathsf{t}^2_{(k-1)}$ when c^2_α is the chosen critical value. Variables with larger t^2 values are retained and all other variables are eliminated. Only one decision is needed to select the final model, even for $N = 1000$, so we call this a 1-cut approach. Moreover, 'goodness of fit' is never considered, although the fit of the final outcome is determined indirectly by the choice of c^2_α. For stochastic problems, such a ranking is itself stochastic, so can differ on different samples for t^2-statistics close to the critical value—however, that effect also applies to estimating a correctly specified LDGP equation, so is not a feature of selection *per se*.

So is this approach just a quick route to 'garbage in, garbage out'? It could be:

(a) if (10) bore no relation to the underlying DGP—but that is countered by basing the initial GUM on the best available subject-matter theory, checking that the selection is a congruent representation of the full sample evidence and parsimoniously encompassing the GUM (see Hendry and Richard 1989);

(b) if too loose a significance level was set (say 5% for $N = 1000$), so many irrelevant variables are retained on average (about 50 at 5%)—but an investigator can set c^2_α to maintain an average false null retention at one variable using $\alpha \leq 1/N$, with α declining as $T \to \infty$ to ensure that any finite parameter LDGP will be consistently selected

(c) if the error distribution was highly non-normal, so conventional critical values were inappropriate—but IIS can remove the non-normality sufficiently to sustain Gaussian-based c^2_α (see Castle et al. 2011a);

(d) if the relevant variables were not very significant—but then even the LDGP would not deliver useful results: one cannot expect selection to discover the underlying reality better than knowing it up to a set of unknown parameters (but see Castle, Doornik and Hendry 2011b on detecting many 'small effects').

While (10) is a 'toy model' of realistic selection, it also reveals why past approaches to model selection have such a poor reputation. Consider selecting to maximize 'goodness of fit': that is bound to retain many more variables than the n which are actually relevant, and has led to many 'penalty functions' being proposed to mitigate that difficulty. Unfortunately, such approaches either are modifications of expanding-search algorithms (like stepwise or lasso), or are part of procedures that consider all possible sub-models to select the 'best', both of which suffer serious drawbacks as follows. Apart from the general problem of such methods not checking congruence, forward selection must conduct inference in under-specifications, so critical values for decisions will generally be incorrect. Moreover, they use simple correlations when partial correlations matter in multiple-variable models, so are bound to mis-select when variables are negatively correlated such that both need to be included before either is significant. On the latter, there 2^N possible sub-models, so for large N there is no feasible α that can control spurious significance, and *ad hoc* procedures like 'hold back sub-samples to test against' are adopted as correctives.

In non-orthogonal models, sample t^2-statistics can change substantively as different variables are eliminated, so 1-cut selection is unreliable. Nevertheless, the aim is to determine an ordering of the variables that is related to their relevance in the LDGP. There are many ways to undertake selection searches, and to make decisions from the resulting findings, but large improvements in doing so have occurred over the past 15 years. Multiple-path searches improve performance dramatically, as Hoover and Perez (2009) demonstrated, but a more systematic (second generation) method like *PcGets* (see Hendry and Krolzig 2009) is better still. The third generation of tree-search algorithms that examine the whole search space and discard irrelevant routes systematically, such as *Autometrics* (see Doornik 2009a; Doornik and Hendry 2009), yield further improvements, and also handle settings where $N > T$ by including block expanding searches (see Doornik 2009b). Searches follows branches till no insignificant variables remain, then test for congruence and parsimonious encompassing, backtracking if either fails till the first non-rejection is found. If the GUM is congruent, then so are all terminals, comprising undominated, mutually-encompassing representations. If several terminal models are found, they can be combined or one selected (by, e.g., the Schwarz 1978 criterion). Single-path reductions often fail because any mistaken elimination cannot be corrected later.

In practice, all the complications of empirical data need to be tackled jointly using a formulation like (9). Since the regressors cannot now be orthogonal, the initial t^2 values need not represent the importance of the regressor in the LDGP, so a search is required. Whatever the quality of the theory basis and the generality of the initial specification, the most likely state of nature is that some variables are irrelevant, whereas some substantively relevant components are inadvertently omitted as not known. Thus, selection takes place in the context of available theory where even general initial models are likely to be under-specified in some ways while over-specified in others. Such general settings are difficult to analyze, but Monte Carlo simulations, such as those reported in Castle and Hendry (2010b) and Castle et al. (2009; 2011a) suggest that *Autometrics* selections provide a 'good' approximation to the LDGP parameters in terms of their mean-square errors. Castle et al. (2011a) describe the selection process in more detail, explain the selection criteria, and consider the evaluation of congruence. Hendry and Krolzig (2005) discuss (approximate) post-selection bias corrections which take into account that the selection criterion $t^2_{(k)} \geq c^2_\alpha$ only retains large values, and Castle, Fawcett and Hendry (2009; 2011) respectively apply the approach to 'nowcasting' and forecasting. Finally, Hendry and Krolzig (2005) and Castle et al. (2011b) discuss why perfect collinearity between regressors in the GUM is not problematic for multi-path search algorithms, and can be used to resolve the difficulty raised by Campos and Ericsson (1999) that the initial parametrization may determine the parsimony of the final selection.

Returning to the formulation in *section 4*, the aim was to specify a sufficiently general GUM that nested the LDGP chosen for analysis by the investigator, so tackled all the complications of economics data jointly, then select the most parsimonious congruent representation feasible to discover the β^* associated with the relevant functions $\mathbf{g}^*(\mathbf{z}_t^*)\ldots\mathbf{g}^*(\mathbf{z}_{t-s}^*)$, where $\mathbf{g}^*(\mathbf{z}^*)$ denotes the appropriate functions of the selected subset of the basic explanatory variables $\mathbf{z}$, including s lags as necessary, jointly with $\mathbf{d}_t$ and $\mathscr{T}^*$, where $\mathbf{d}_t$ denotes the vector of indicators for breaks and outliers, using the effective sample $\mathscr{T}^*$. Then the finally chosen model is the congruent parsimonious-encompassing representation:

$$\lambda(L)\,y_t = \beta^*(L)'\,\mathbf{g}\left(\mathbf{z}_t^*\right) + \gamma'\mathbf{d}_t + v_t \tag{11}$$

when v_t is the unexplained component, and L denotes the lag operator in (11). Such a task includes establishing the validity of conditioning on any contemporaneous variables, perhaps used as instruments, as well as ensuring that $\{v_t\}$ is an innovation process from valid sequential factorization. A similar formulation applies when $\mathbf{y}_t$ is a vector of target variables to be modeled, although we have not addressed that setting here.

8. Evaluating Selected Models

To avoid blurring the boundaries between discovery and evaluation, we now consider what warrant can be established independently of the empirical model discovery process, additional to congruence and the 'corroboration' of the initial theory specification (see Spanos 1995 on testing theories using non-experimental data). Since 'anything goes' in the former, as *section 2* stressed, stringent evaluation is required in the latter (see e.g., Mayo and Spanos 2006). Such a warrant has to invoke new data, new evidence, new instruments or new tests. William Herschel discovered Uranus because he had a detailed map of the night sky in his brain, and could perceive change against it. However, his finding was only accorded the status of a planet when its orbit was calculated to be round the sun, and its sighting was reliably replicated by others. Perhaps it was no accident that his son, John Herschel (1830) emphasized the distinction between discovery and evaluation.

But independent stringent substantive evaluation is difficult. For example, as discussed in §8.2, accurate forecasting is insufficient. Indeed, even making 'accurate predictions' is not sufficient grounds for accepting a theory model, where prediction denotes stating in advance of a test what its result should be, which need not be in advance of the event: Ptolemaic epicycles predicting eclipses of the moon are a well-known example (see Spanos 2007). Returning to the example of light waves being bent by the sun's gravity, a false theory that this was due to the weight of photons combined with an invalid measurement of that weight could be designed to successfully predict the magnitude of bending as judged by distortions of light from distant stars during eclipses. Moreover, successive successful predictions can combine to refute the very theory each alone corroborated (see e.g., Ericsson and Hendry 1999).

Correct theories can also make false predictions. Two famous cases are the integer masses of elements which measurements always contradicted—until isotopes were discovered; and Pasteur's germ theory discussed above. The Duhem-Quine problem (see e.g., Harding 1976) suggests it is difficult to disentangle which aspect led to any unsuccessful outcome, so failed evaluations require 'detective work'—automatic methods will need supplemented by human insight.

Conversely, the entire framework may be incorrect, including the measurement system and the basis for testing, but as the best available, it is the approach that will be advanced. In the physical and biological sciences, huge advances have been made and embodied in new procedures, new forms of energy, and new materials that have transformed living standards. Major corrections to a framework fit with the 'paradigm shift' notion from Kuhn (1962), where the fads and fashions of one generation of economists (say, e.g., marginalists, Keynesians, neo-classicists etc.) are overthrown in favor of another, perhaps no less 'real' approach, when a conjunction of apparently adverse evidence combines with new thinking. Expanding a frontier perforce offers only partial explanations, and many of the difficulties in economics arise from the non-stationarity of the process being investigated, so empirical discoveries and evaluated theories need not be permanent, notwithstanding the claims in Robbins (1932), or the implicit assumption of most extant economic theories. Above, the relevant theory was embedded in the general model, so a powerful evaluation is if all the additional candidate effects transpire to be irrelevant.

A crucial aspect for a successful warrant seems to be making repeated distinctly-different predictions, none of which is refuted, using a given model consistent with the general theoretical framework. Here economic policy can play a major role: many novel changes in fiscal, monetary, tax and benefit policies are always happening, and at least the first two inevitably involve location shifts with consequences that can be tested against the predictions of models. Few are likely to survive unscathed, but if any did, they would offer a useful basis for future policies, even though other changes will inevitably occur outside the policy process (see e.g., Hendry and Mizon 2005). Matching such responses to predictions requires a form of causal connection between the policy variables and the outcomes, as well as the model capturing such relationships, an issue to which we now turn.

8.1 Testing Super Exogeneity

Parameter invariance under regime shifts is essential to avoid mis-prediction when using policy models. Super exogeneity combines parameter invariance with valid conditioning, so is a key concept for economic policy (see Engle and Hendry 1993). An automatic test thereof in Hendry and Santos (2010) uses impulse-indicator saturation in marginal models of the conditioning variables (i.e., models for the $z_{i,t}$ in (10), say), retaining all the significant indicators, then testing their relevance in the conditional model. No *ex ante* knowledge of timing or magnitudes of breaks is required, and an investigator need not know the LDGP of those marginal variables. The test has the correct size under the null of super exogeneity for a range of sizes of the marginal-model saturation tests, with power to detect failures of super exogeneity when location shifts occur in the marginal models. This provides one 'outside' test to stringently evaluate a selected policy model. A related test for invariance in expectations models is proposed in Castle, Doornik, Hendry and Nymoen (2011c).

Moreover, super exogeneity is close to providing a sufficient condition for causality as follows (see e.g., Hendry 2004). Reconsider the model in (10) when $N = 1$ (for simplicity) and (10) coincides with the DGP:

$$y_t = \beta z_t + \epsilon_t \tag{12}$$

Then if $\partial y_t / \partial z_t = \beta$ constantly for a range on interventions on $\{z_t\}$, then z_t is both super exogenous for β and must be causing the changes in y_t, albeit possibly through an intermediary to which z_t is also related by invariant parameters. Unfortunately, neither success nor failure in forecasting need help evaluation.

8.2 Forecasting

A forecast is a statement about the future, and as its name implies, like casting dice, or fishing lines, or casting spells (see Hendry 2001), forecasting is a chancy affair. Specifically, the lack of time invariance in economics renders both forecast failure or success inappropriate criteria for judging the validity of a theory, or even a forecasting model. Despite numerous papers demonstrating these statements, from Hendry and Mizon (2000) through Clements and Hendry (2005); Castle, Fawcett and Hendry (2010); Castle et al. (2011) to Castle and Hendry (2011b), who emphasize that there need be no connection between the verisimilitude of a model and any reasonable measure of its forecast accuracy, forecasting success is often proclaimed to be the 'gold standard' of model evaluation. That claim cannot be proved. First, Miller (1978) and Hendry (1979) highlight that in a stationary, ergodic world, forecasts from least-squares estimated models must unconditionally attain their expected forecast accuracy independently of the goodness of the specification or its closeness to the DGP. Ptolemaic epicycles again spring to mind. Thus, 'success' need not relate to even congruence. Accuracy with success sounds more stringent, but depends on the intrinsic uncertainty in the entity being forecast—it is easy to forecast low-variance outcomes 'accurately'. Even in a non-constant world, transformations that robustify models against systematic forecast failure are just as effective when models are badly mis-specified.

Secondly, a simple analogy explains why the converse of forecast failure may also be indecisive. Consider Apollo 13's ill-fated journey to the moon in April 1970. The craft was predicted to land at a specific time and date, but was knocked off course *en route* by an unanticipated oxygen cylinder explosion. Consequently, NASA's forecast was systematically and badly wrong—increasingly so as the days go by. But that forecast failure was not due to poor forecasting algorithms, nor does it refute the underlying Newtonian theory of universal gravitation. Rather, it reveals that forecast failure is mainly due to unanticipated location shifts occurring in the forecast period.

9. Conclusion: Empirical Model Discovery in Economics

We can now clarify what empirical model discovery entails in economics. Most specifications of models intended to be matched against data are derived from a pre-existing theory. For example, (1) in §1 was probably derived from an optimization problem postulated for a single agent, subject to some implicit *ceteris paribus* conditions. There are many styles of implementation, from tight theoretical specifications to be calibrated quantitatively, to using (1) as a guide to the set of variables and the possible form of relationship. To be applicable to the real world, economic theory has to explain the behavior of the agents who create the DGP. The LDGP is then a reduction of the DGP to the subset of variables believed still to capture the relevant behavior, as discussed in *section 5*. However, since so many features of any model are unknown until the data are investigated, it seems crucial to re-frame empirical modeling as a discovery process, part of a progressive research strategy of accumulating empirical evidence, based on extending the chosen set $\mathbf{x}$ in the theory to a general model (§6), then selecting an appropriate representation (§7) which retains and evaluates that theory yet jointly investigates the relevance of longer lags, functional forms, outliers and shifts, *inter alia*. Globally, knowledge augmentation perforce proceeds from the simple to the general; but locally, it need not do so. 'Keep it simple stupid', the so-called KISS principle, at best applies to the final selection, and certainly not to the initial model.

Thus, automatic empirical model discovery can be seen to require the same seven stages noted above as common aspects of discovery in general. The first involves the prior theoretical derivation both of the relevant set $\mathbf{x}$ defining the LDGP, and of the relationships between its components, specifying preferred functional forms, lag reactions, etc., for the retained formulation.

The second is the automatic creation of a more general model than initially envisaged by an investigator. To reiterate, the choice of the $m+1$ variables to analyze is fundamental, as it determines what the target LDGP is, whereas the other extensions (lags, non-linear functions, impulse-indicator saturation, etc.) determine how well that LDGP is approximated. Failure to include substantively important influences at either step will usually lead to non-constant relationships that are hard to interpret—and probably be dominated by an investigator willing to consider a broader universe of determinants. When a theory model is simply imposed on the evidence, little can be learned—reaching outside is essential to reveal phenomena that were not originally conceived. However, empirical model discovery is not an inductive approach, since the prior theory still plays a key role in structuring the framework, and is the vehicle for thinking about the basic set $\mathbf{x}_t$ of determinants and how they might matter. In macroeconomic models confronting wide-sense non-stationary data, no theory is able to cover all aspects, but ideas remain important as they can be embedded in and guide the general model. As discussed, if correct, theory variables should be retained in the final selected model, perhaps augmented by features where the theory was incomplete, which had such features been omitted, might have led to its rejection, as illustrated in Hendry and Mizon (2011).

Thirdly, having created a general initial model, efficient selection is essential to find the representations to which it can be reduced without loss of relevant information. Model selection then plays a key role, and if the first two stages created an initial set of $N > T$ candidate variables, selection has to be automatic since the scale is too large for humans. An efficient selection method should have a small probability of retaining irrelevant variables, and a probability of retaining relevant variables similar to the LDGP when conducting inference at the same significance levels. By embedding the theory model as a retained feature, selection after orthogonalizing with respect to all other variables ensures either the same estimator distributions when the theory is complete and correct, while stringently evaluating that theory against a wide range of alternatives, or learning that it is not valid and concluding with an improved model.

Next, the algorithm has to recognize when the search is completed, namely when a congruent parsimonious-encompassing representation has been found characterizing the target LDGP. Such selections are called terminal models. Multiple mutually-encompassing terminals are possible, especially when variables are highly collinear and relatively loose significance levels are used, but at tight significance usually one selection appears. However, should forecasting be an objective as in §8.2, there may be advantages to combining terminals.

Fifthly, to appropriately quantify the outcome requires near unbiased parameter estimates with small mean-square errors (MSEs), and a near unbiased estimate of the equation standard error. Approximate bias corrections are shown in Hendry and Krolzig (2005) to substantially reduce the MSEs of any adventitiously retained irrelevant variables. Near normality is important both for accurate inferences during selection, and for the bias corrections which are derived under the null of a Gaussian distribution. Here, IIS plays an important role in removing breaks and outliers to facilitate normality. Throughout, an appropriate estimator is essential, so it is important to test exogeneity in the final model as in §8.1.

The sixth step is evaluating the resulting discovery. Since all the theory and data evidence will have been employed in the first five stages, new data, new tests or new procedures are needed for independent evaluation of the selection. When the initial general model is estimable from the available sample, then its evaluation by mis-specification testing is one of the first activities in empirical modeling. When that is infeasible, an initial reduction to $N < T$ is required before such tests are conducted, although they could reject the null of congruence at that point. We do not use 'hold-back' samples, both because the lack of time invariance makes it unclear what is learned if the results differ, and even when the DGP is constant, doing so is inefficient: see e.g., Lynch and Vital-Ahuja (1998); Hendry and Krolzig (2004).

The final step is to summarize the findings parsimoniously in a model that is undominated at the significance level α used. This is almost automatic given the selection criterion of a congruent parsimonious-encompassing representation—

but not quite. Having simplified the model to a size a human can grasp, various further simplifications may suggest themselves, including: combining indicators in a single dummy (see e.g., Hendry and Santos 2005); combining lags of variables into more interpretable forms (see e.g., Hendry 1995), or combining other groups of variables (see e.g. Campos and Ericsson 1999); replacing unrestricted non-linear functions by an encompassing theory-derived form, such as an ogive (see Castle and Hendry 1995); and so on, again requiring human intervention. All seven stages can be interweaved for a practical approach as we have described, enhancing the scope and capabilities of empirical researchers, not replacing them. Other disciplines are experimenting with automated empirical discovery (see e.g., King et al. 2009) with some success, and the concepts and methods involved may themselves prove a fruitful ground for analyses by philosophers of science.

References

Agassi, J. (1977), "Who Discovered Boyle's Law?", *Studies in History and Philosophy of Science* A(8), 189–250.

Anderson, T. W. (1962), "The Choice of the Degree of a Polynomial Regression as a Multiple-decision Problem", *Annals of Mathematical Statistics* 33, 255–265.

Bennett, J., M. Cooper, M. Hunter and L. Jardine (2003), *London's Leonardo*, Oxford: Oxford University Press.

Blaug, M. (1980), *The Methodology of Economics*, Cambridge: Cambridge University Press.

Bontemps, C. and G. E. Mizon (2003), "Congruence and Encompassing", in: Stigum, B. P. (ed.), *Econometrics and the Philosophy of Economics*, Princeton: Princeton University Press, 354–378.

— and — (2008), "Encompassing: Concepts and Implementation", *Oxford Bulletin of Economics and Statistics* 70, 721–750.

Campos, J. and N. R. Ericsson (1999), "Constructive Data Mining: Modeling Consumers' Expenditure in Venezuela", *Econometrics Journal* 2, 226–240.

—, — and D. F. Hendry (2005), "Editors' Introduction", in: Campos, J., N. R. Ericsson and D. F. Hendry (eds.), *Readings on General-to-Specific Modeling*, Cheltenham: Edward Elgar, 1–81.

Castle, J. L., J. A. Doornik and D. F. Hendry (2009), "Model Selection When There Are Multiple Breaks", working paper 472, Economics Department, University of Oxford.

—, — and — (2011a), "Evaluating Automatic Model Selection", *Journal of Time Series Econometrics* 3(1), DOI: 10.2202/1941-1928.1097.

—, — and — (2011b), "Model Selection in Equations with Many 'Small' Effects", discussion paper 528, Economics Department, University of Oxford.

—, —, — and R. Nymoen (2011c), "Mis-specification Testing: Noninvariance of Expectations Models of Inflation", working paper, Economics Department, University of Oxford.

—, N. W. P. Fawcett and D. F. Hendry (2009), “Nowcasting Is Not Just Contemporaneous Forecasting”, *National Institute Economic Review* 210, 71–89.

—, — and — (2010), “Forecasting with Equilibrium-correction Models during Structural Breaks”, *Journal of Econometrics* 158, 25–36.

—, — and — (2011), “Forecasting Breaks and during Breaks”, in: Clements, M. P. and D. F. Hendry (eds.), *Oxford Handbook of Economic Forecasting*, Oxford: Oxford University Press, 315–353.

— and D. F. Hendry (2010a), “A Low-dimension, Portmanteau Test for Non-linearity”, *Journal of Econometrics* 158, 231–245.

— and — (2010b), “Model Selection in Under-specified Equations with Breaks”, discussion paper 509, Economics Department, University of Oxford.

— and — (2011a), “Automatic Selection of Non-linear Models”, in: Wang, L., H. Garnier and T. Jackman (eds.), *System Identification, Environmental Modelling and Control*, New York: Springer, forthcoming.

— and — (2011b), “On Not Evaluating Economic Models by Forecast Outcomes”, *Istanbul University Journal of the School of Business Administration* 40, 1–21.

— and N. Shephard (2009) (eds.), *The Methodology and Practice of Econometrics*, Oxford: Oxford University Press.

Chan, K. S. and H. Tong (1986), “On Estimating Thresholds in Autoregressive Models”, *Journal of Time Series Analysis* 7, 179–194.

Clark, H. H. and E. V. Clark (1977), *Psychology and Language: An Introduction to Psycholinguistics*, New York: Harcourt Brace Jovanovich.

Clements, M. P. and D. F. Hendry (2005), “Evaluating a Model by Forecast Performance”, *Oxford Bulletin of Economics and Statistics* 67, 931–956.

Doob, J. L. (1990[1953]), *Stochastic Processes*, New York: John Wiley Classics Library.

Doornik, J. A. (2009a), “Autometrics”, in: Castle and Shephard (2009), 88–121.

— (2009b), “Econometric Model Selection with More Variables Than Observations”, working paper, Economics Department, University of Oxford.

— and D. F. Hendry (2009), *Empirical Econometric Modelling Using PcGive: Volume I*, London: Timberlake Consultants Press.

Drake, S. (1980), *Galileo*, Oxford: Oxford University Press.

Eddington, C. (1928), *Space, Time, and Gravitation*, Cambridge: Cambridge University Press.

Engle, R. F. and D. F. Hendry (1993), “Testing Super Exogeneity and Invariance in Regression Models”, *Journal of Econometrics* 56, 119–139.

Ericsson, N. R. and D. F. Hendry (1999), “Encompassing and Rational Expectations: How Sequential Corroboration Can Imply Refutation”, *Empirical Economics* 24, 1–21.

Farmelo, G. (2002) (ed.), *De Motu Cordis*, London: Granta Publications. (It Must Be Beautiful: Great Equations of Modern Science).

Fasano, A. (2009), “Surprises from Celiac Disease”, *Scientific American* 301, 54–61.

Fouquet, R. and P. J. G. Pearson (2006), “Seven Centuries of Energy Services: The Price and Use of Light in the United Kingdom (1300–2000)”, *Energy Journal* 27, 139–178.

Friedman, M. (1974), “Explanation and Scientific Understanding”, *Journal of Philosophy* 71, 5–19.

Gest, H. (2002), “The Remarkable Vision of Robert Hooke (1635–1703): First Observer of the Microbial World”, *Perspectives in Biology and Medicine* 48, 266–272.

Goldstein, R. N. (2010), “What’s in a Name? Rivalries and the Birth of Modern Science”, in: Bryson, B. (ed.), *Seeing Further: The Story of Science and the Royal Society*, London: Harper Press, 107–129.

Granger, C. W. J. and T. Teräsvirta (1993), *Modelling Nonlinear Economic Relationships*, Oxford: Oxford University Press.

Hackett, J. (1997) (ed.), *Roger Bacon and the Sciences: Commemorative Essays*, New York: Brill.

Harding, S. G. (1976) (ed.), *Can Theories Be Refuted? Essays on the Duhem-Quine Thesis*, Dordrecht: D. Reidel Publishing Company.

Harré, R. (1981), *Great Scientific Experiments*, Oxford: Oxford University Press.

Henderson, J. W. (1997), “The Yellow Brick Road to Penicillin: A Story of Serendipity”, *Mayo Clinic Proceedings* 72, 683–687.

Hendry, D. F. (1976), “The Structure of Simultaneous Equations Estimators”, *Journal of Econometrics* 4, 51–88.

— (1979), “The Behaviour of Inconsistent Instrumental Variables Estimators in Dynamic Systems with Autocorrelated Errors”, *Journal of Econometrics* 9, 295–314.

— (1995), *Dynamic Econometrics*, Oxford: Oxford University Press.

— (2000), *Econometrics: Alchemy or Science?*, Oxford: Oxford University Press. New Edition.

— (2001), “How Economists Forecast”, in: Hendry, D. F. and N. R. Ericsson (eds.), *Understanding Economic Forecasts*, Cambridge, Mass.: MIT Press, 15–41.

— (2004), “Causality and Exogeneity in Non-stationary Economic Time Series”, in: Welfe, A. (ed.), *New Directions in Macromodelling*, Amsterdam: North Holland, 21–48.

— (2009), “The Methodology of Empirical Econometric Modeling: Applied Econometrics through the Looking-glass”, in: Mills, T. C. and K. D. Patterson (eds.), *Palgrave Handbook of Econometrics*, Basingstoke: Palgrave MacMillan, 3–67.

— and S. Johansen (2010), “Model Selection When Forcing Retention of Theory Variables”, unpublished paper, Economics Department, University of Oxford.

—, — and C. Santos (2008), “Automatic Selection of Indicators in a Fully Saturated Regression”, *Computational Statistics* 33, 317–335. Erratum, 337–339.

— and H.-M. Krolzig (2004), “Sub-sample Model Selection Procedures in General-to-specific Modelling”, in: Becker, R. and S. Hurn (eds.), *Contemporary Issues in Economics and Econometrics: Theory and Application*, Cheltenham: Edward Elgar, 53–74.

— and — (2005), “The Properties of Automatic Gets Modelling”, *Economic Journal* 115, C32–C61.

— and G. E. Mizon (2000), "On Selecting Policy Analysis Models by Forecast Accuracy", in: Atkinson, A. B., H. Glennerster and N. Stern (eds.), *Putting Economics to Work: Volume in Honour of Michio Morishima*, London School of Economics: STICERD, 71–113.

— and — (2005), "Forecasting in the Presence of Structural Breaks and Policy Regime Shifts", in: Andrews, D. W. K. and J. H. Stock (eds.), *Identification and Inference for Econometric Models*, Cambridge: Cambridge University Press, 480–502.

— and — (2010), "On the Mathematical Basis of Inter-temporal Optimization", discussion paper 497, Economics Department, University of Oxford.

— and — (2011), "Econometric Modelling of Time Series with Outlying Observations", *Journal of Time Series Econometrics* 3(1), DOI: 10.2202/1941-1928.1100.

— and J.-F. Richard (1989), "Recent Developments in the Theory of Encompassing", in Cornet, B. and H. Tulkens (eds.), *Contributions to Operations Research and Economics. The XXth Anniversary of CORE*, Cambridge, Mass.: MIT Press, 393–440.

— and C. Santos (2005), "Regression Models with Data-based Indicator Variables", *Oxford Bulletin of Economics and Statistics* 67, 571–595.

— and — (2010), "An Automatic Test of Super Exogeneity", in: Watson, M. W., T. Bollerslev and J. Russell (eds.), *Volatility and Time Series Econometrics*, Oxford: Oxford University Press, 164–193.

Herschel, J. (1830), *A Preliminary Discourse on The Study of Natural Philosophy*, London: Longman, Rees, Browne, Green and John Taylor.

Hildenbrand, W. (1994), *Market Demand: Theory and Empirical Evidence*, Princeton: Princeton University Press.

Holton, G. (1986), "The Advancement of Science, and Its Burdens", *Daedalus* 115, 77–104.

— (1988), *Thematic Origins of Scientific Thought*, Cambridge: 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.

Johansen, S. and B. Nielsen (2009), "An Analysis of the Indicator Saturation Estimator as a Robust Regression Estimator", in: Castle and Shephard (2009), 1–36.

King, R. D. et al. (2009), "The Automation of Science", *Science* 324(5923), 85–898.

Koopmans, T. C. (1947), "Measurement without Theory", *Review of Economics and Statistics* 29, 161–179.

Kuhn, T. (1962), *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press.

Kydland, F. E. and E. C. Prescott (1991), "The Econometrics of the General Equilibrium Approach to Business Cycles", *Scandanavian Journal of Economics* 93, 161–178.

Lakatos, I. and A. Musgrave (1974) (eds.), *Criticism and the Growth of Knowledge*, Cambridge: Cambridge University Press.

Leamer, E. E. (1983), "Let's Take the Con Out of Econometrics", *American Economic Review* 73, 31–43.

Lomas, R. (1999), *The Man Who Invented the Twentieth Century: Nicola Tesla, Forgotten Genius of Electricity*, London: Headline Books.

Lynch, A. W. and T. Vital-Ahuja (1998), "Can Subsample Evidence Alleviate the Data-snooping Problem? A Comparison to the Maximal R2 Cutoff Test", discussion paper, Stern Business School, New York University.

Mason, S. F. (1977[1962]), *A History of the Sciences*, New York: Collier Books. 2nd edn.

Mayo, D. (1981), "Testing Statistical Testing", in: Pitt, J. C. (ed.), *Philosophy in Economics*, Dordrecht: D. Reidel Publishing Co., 175–230.

— and A. Spanos (2006), "Severe Testing as a Basic concept in a Neyman-Pearson Philosophy of Induction", *British Journal for the Philosophy of Science* 57, 323–357.

McLeod, K. S. (2000), "Our Sense of Snow: The Myth of John Snow in Medica Geography", *Social Science & Medicine* 50, 923–935.

Messadié, G. (1991), *Great Scientific Discoveries*, Edinburgh: Chambers.

Miller, P. J. (1978), "Forecasting with Econometric Methods: A Comment", *Journal of Business* 51, 579–586.

Morgan, M. S. (1990), *The History of Econometric Ideas*, Cambridge: Cambridge University Press.

Musgrave, A. (1976), "Why Did Oxygen Supplant Phlogiston? Research Programmes in the Chemical Revolution", in: Howson, C. (ed.), *Method and Appraisal in the Physical Sciences*, Cambridge: Cambridge University Press, 181–209.

Nelson, R. R. (1959), "The Simple Economics of Basic Scientific Research", *The Journal of Political Economy* 67, 297–306.

Nussbaumer, H. and L. Bieri (2009), *Discovering the Expanding Universe*, Cambridge: Cambridge University Press.

Pinker, S. (1994), *The Language Instinct*, London: Penguin Books.

— (2002), *The Blank Slate*, London: Penguin Books.

Popper, K. R. (1959), *The Logic of Scientific Discovery*, New York: Basic Books.

— (1963), *Conjectures and Refutations*, New York: Basic Books.

Priestley, M. B. (1981), *Spectral Analysis and Time Series*, London: Academic Press.

Qin, D. (1993), *The Formation of Econometrics: A Historical Perspective*, Oxford: Clarendon Press.

Robbins, L. (1932), *An Essay on the Nature and Significance of Economic Science*, London: Macmillan.

Schultz, S. G. (2002), "William Harvey and the Circulation of the Blood: The Birth of a Scientific Revolution and Modern Physiology", *News in Physiological Sciences* 17, 175–180.

Schumpeter, J. (1954), *History of Economic Analysis*, New York: Oxford University Press.

Schwarz, G. (1978), "Estimating the Dimension of a Model", *Annals of Statistics* 6, 461–464.

Sims, C. A. (1996), "Macroeconomics and Methodology", *Journal of Economic Perspectives* 10, 105–120.

Smith, G. D. (2002), "Commentary: Behind the Broad Street Pump: Aetiology, Epidemiology and Prevention of Cholera in Mid-19th Century Britain", *International Journal of Epidemiology* 31, 920–932.

Spanos, A. (1986), *Statistical Foundations of Econometric Modelling*, Cambridge: Cambridge University Press.

— (1990), "Towards a Unifying Methodological Framework for Econometric Modelling", in: Granger, C. W. J. (ed.), *Modelling Economic Series*, Oxford: Clarendon Press, 335–364.

— (1995), "On Theory Testing in Econometric Modelling with Non-experimental Data", *Journal of Econometrics* 67, 189–226.

— (1999), *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*, Cambridge: Cambridge University Press.

— (2000), "Revisiting Data Mining: 'Hunting' with or without a License", *Journal of Economic Methodology* 7, 231–264.

— (2006), "Econometrics in Retrospect and Prospect", in: Mills, T. C. and K. D. Patterson (eds.), *Palgrave Handbook of Econometrics*, Basingstoke: Palgrave MacMillan, 3–58.

— (2007), "Curve-fitting, the Reliability of Inductive Inference and the Error-statistical Approach", *Philosophy of Science* 74, 1046–1066.

Teräsvirta, T. (1994), "Specification, Estimation and Evaluation of Smooth Transition Autoregressive Models", *Journal of the American Statistical Association* 89, 208–218.

Thompson, W. K. (1862), "On the Age of the Sun's Heat", *Macmillan's Magazine* 5, 288–293.

— (1864), "On the Secular Cooling of the Earth", *Transactions of the Royal Society of Edinburgh* 23, 167–169.

— (1899), "The Age of the Earth as an Abode Fitted for Life", *Science* 9(228), 665–674.

Waller, J. (2002), *Fabulous Science*, Oxford: Oxford University Press.