CENTRE FOR ECONOMETRIC ANALYSIS
CEA@Cass

http://www.cass.city.ac.uk/cea/index.html

Cass Business School
Faculty of Finance
106 Bunhill Row
London EC1Y 8TZ

Theory Testing in Economics and the Error Statistical Perspective

Aris Spanos

CEA@Cass Working Paper Series

WP–CEA–04-2008

Theory Testing in Economics and the Error Statistical Perspective

Aris Spanos
Department of Economics, Virginia Tech

Abstract

1 Introduction

For a domain of inquiry to live up to standards of scientific objectivity it is generally required that its theories be tested against empirical data. The central philosophical and methodological problems of economics may be traced to the unique character of both economic theory and its non-experimental (observational) data. Alternative ways of dealing with these problems are reflected in rival methodologies of economics. My goal here will not be to promote any one such methodology at the expense of its rivals so much, as to set the stage for understanding and making progress on the conundrums in the methodology and philosophy of economics. This goal, I maintain, requires understanding the changing roles of theory and data in the development of economic thought, along side of the shifting philosophies of science which explicitly or implicitly find their way into economic theorizing and econometric practice. Given that this requires both economists and philosophers of science to stand outside their usual practice and reflect on their own assumptions, it is not surprising that this goal has been rather elusive.

1.1 The pre-eminence of theory

Historically, theory has generally held the pre-eminent role in economics with data being given the subordinate role of 'quantifying theories' presumed to be true. In this conception, whether in the classical (19th century) or neoclassical (20th century) historical period or even in contemporary 'textbook' econometrics, data does not so much test as allow instantiating theories: sophisticated econometric methods enabled elaborate ways 'to beat data into line' (as Kuhn would say) to accord with an assumed theory. Since the theory has little chance to be falsified, such instantiations are highly insevere tests of theory. Were the theories known to be (approximately) true at the outset, this might not be problematic, but in fact main-stream economic theories have been invariably unreliable predictors of economic phenomena, and rival theories could easily be made to fit the same data equally well if not better.

1.2 Data-driven modeling

Attempts to redress the balance and give data a more substantial role in theory testing were frustrated by two inveterate problems: (i) the huge gap between economic theories and the available observational data, and (ii) the difficulty of assessing when a fitted model 'accounts for the regularities in the data'. Faced with these difficulties, endeavors to redress the balance tended to focus primarily on data-driven models and assess their adequacy in terms of goodness-of-fit measures. These efforts often gave rise to fitted models with better short term predictions, but shed very little light (if any) on understanding the underlying economic phenomena. Indeed, data-driven correlation, linear regression, factor analysis and principal component analysis, relying on goodness-of-fit, have been notoriously unreliable when applied to observational data, especially in the social sciences. Furthermore, the arbitrariness of goodness-of-fit measures created a strong impression that one can 'forge' significant correlations (or regression coefficients) at will, if one was prepared to persevere long enough 'mining' the data. The (mistaken) impression that statistical spuriousness is both inevitable and endemic, especially when analyzing observational data, is almost universal among social scientists and philosophers. This has led to widely held (but erroneous) belief that substantive information provides the only safeguard against statistical spuriousness. As a result, the prevailing pre-eminence of theory perspective in economics have persistently charged any practices that strive for 'objectivity' in data analysis, by relying on the 'regularities' contained in the data, with 'measurement without theory', 'data-mining' and 'hunting' for statistical significance.

Perhaps unsurprisingly, with Kuhnian relativisms and 'theory-laden' data holding sway in philosophy of science, it became easy to suppose that this is as good as it gets in science in general, making it respectable to abandon or redefine objectivity as opinions shared by a group (McCloskey, 1985), or formalized in subjective Bayesian degrees of belief (Leamer, 1978).

1.3 The 'third way'

Some econometricians set out a 'third way' wishing to avoid both extreme practices, the prevailing theory-driven empirical modeling on one hand, and the data-driven modeling based on goodness-of-fit measures, on the other. Aspiring to give data 'a voice of its own', which could reliably constrain economic theorizing, these practitioners were thrust into the role of inventing new methods while striving to find or adapt a suitable foundation in one or another philosophy of science (Kuhn, Popper, Lakatos); see Hendry (1980). While loosely reflecting on a Popperian conception to criticize and a Lakatosian demand for 'progressiveness', none of these classic philosophical approaches – at least as they were formulated in philosophy of science – provided an appropriate framework wherein one could address the numerous charges (legitimate or baseless) leveled against any modeling practice that did not conform to 'quantifying theories' presumed true. Indeed, spurred by the general impression that all statistical methods which rely on 'regularities' in the data are highly susceptible to the statistical spuriousness problem, the upholders of the pre-eminence of theory viewpoint could not see any difference between solely data-driven methods and the 'third way', scorning the latter as just 'data mining'; Spanos (1989).

One of the things that I learned from the ERROR 06 conference, out of which these exchanges arise, it is that the shortcomings of the standard philosophies of science growing out of the Popper-Lakatos and rival Carnapian traditions are directly related to the difficulties one had to face in attempting to develop a sound foundation for the theory-data confrontation in economics. In coming to understand these shortcomings, in effect, I am challenging philosophers of science to reevaluate the assumptions of their own standard models and hopefully take some steps toward a framework that can steer us clear of some of the current confusion on theory testing using data.

The fact that the 'third way' in empirical modeling in economics is still struggling to achieve the twin goals of developing an adequate methodology and an underpinning philosophical foundations, has contributed significantly to the lack of a shared research agenda and a common language among the proponents; see the collection of papers in Granger (1990). What they seem to share, and importantly so, is the idea that recognizing the flaws and fallacies of both extreme practices (theory-driven and data-driven modeling) provides a foothold for developing more adequate modeling. Minimally there seems to be the core endeavor to accomplish the goals afforded by sound practices of frequentist statistical methods in learning from data: to collect and model data to obtain reliable 'statistical knowledge' that stands independently of theories which they may be called upon to explain or be tested against it with a view to shed light on the phenomenon of interest.

It is interesting to note that experimental economists also found themselves facing the same quandary as the "third way" proponents in striving to find a suitable philosophical foundation, beyond Kuh, Popper and Lakatos, that would provide a sound methodological frame-work for their efforts to test economic theories using experiments; see Smith (2007), Sugden (2008).

1.4 The error statistical perspective

In an attempt to structure the discussion that follows, the main philosophical and methodological problems and issues in the theory-data confrontation in economics may be located in the need to deal adequately with three interrelated modeling stages:

(A) from theory to testable hypotheses: fashioning an abstract and idealized theory T into hypotheses or claims h which are testable in terms of the available data $\mathbf{z}_0:=(z_1, z_2, \ldots, z_n)$, often observational (non-experimental),

(B) from raw data to reliable evidence: transforming a finite and incomplete set of raw data $\mathbf{z}_0$ – containing uncertainties, impurities and noise – into reliable 'evidence' $\mathbf{e}$ pertinent for theory appraisal,

(C) confronting hypotheses with evidence: relating h and e, in a statistically apposite way, to assess whether $\mathbf{e}$ provides evidence for or against h.

It is argued that the link in (A) proved to be particularly challenging in economics because there is often a huge gap between abstract and idealized economic theories and the available observational data. The link in (B) brings out the problem of establishing the reliability of the information contained in the data independently of the theory in question, which includes statistical spuriousness. The link in (C) raises issues that have to do with how statistical inferences pertain to substantive information; the last link that gives rise to learning from data.

The primary objective of this paper is to make a case that the error statistical perspective (see Mayo, 1996) offers a highly promising framework in the context of which we may begin to get clearer on some of these crucial philosophical/methodological problems and make headway in their resolution; see Spanos (2008). This framework enables one to foreground as well as work out solutions to the problems raised by attempts to successfully deal with the links in (A)-(C), by formulating a refined version of the Fisher-Neyman-Pearson (F-N-P) statistical inference framework founded on the notion of a statistical model: a set of internally consistent probabilistic assumptions specifying a data generating mechanism; see Mayo and Spanos (2008).

1.5 Statistical knowledge

To be more specific, it is argued that the key to giving data a substantial role in theory testing, requires an account where 'statistical knowledge' (in direct analogy to Mayo's experimental knowledge) has 'a life of its own', which stems from the statistical adequacy of the estimated model – its probabilistic assumptions are valid for the particular data. The notion of statistical knowledge in the context of error statistics allows the data 'to have a voice of its own', in the spirit of the proponents of the 'third way' (Hendry, 2000), separate from the theory in question, and succeeds in securing the frequentist goal of objectivity in theory testing, providing a natural home and a common language. The notion of statistical adequacy replaces goodness-of-fit as the criterion for: (a) accounting for the regularities in the data, (b) securing predictive ability, and (c) dealing affectively with the problem of spurious statistical results by scorching them as statistically meaningless (due to statistical misspecification), without having to invoke any substantive information. Indeed, it is shown that goodness-of-fit measures are themselves are untrustworthy when the fitted model is misspecified.

Revisiting the theory-data confrontation in economics using the error statistical perspective, puts us in a position to understand some of the problems raised by contemporary methodological discussions in economics/econometrics and sets the stage for making progress in an area where chronic problems have stumped any significant advancement for almost three centuries.

2 Theory-data confrontation in economics and the error statistical account

2.1 Philosophy of science and theory testing

Viewing the developments in philosophy of science since the 1930s from the viewpoint of the theory-data confrontation in economics, it becomes clear why the philosophical discourses have left the econometrician hanging. To be more specific, the logical empiricists' perspective of induction, as primarily a logical relationship L(e,h) between evidence **e** – taken as objectively given – and a hypothesis h, essentially assumes away the crucial links (A)-(C) facing a practicing econometrician. Moreover, the inductive logics of logical empiricists had little affinity to the ways practicing econometricians understand the theory-data confrontation in the context of the Fisher-Neyman-Pearson (F-N-P) frequentist statistical inference; see Cox and Hinkley (1974). The post logical empiricist developments in philosophy of science associated with Duhemian ambiguities, underdetermination and the theory-ladeness of observation problem (see Chalmers, 1999), made the problem of theory-data confrontation in economics seem even more hopeless.

2.2 The error statistical account and theory testing

The first promising signs that discussions in philosophy of science could be potentially relevant on issues pertaining to the theory-data confrontation in economics emerged from the 'new experimentalist' tradition, which began focusing attention on modeling the processes that generated the raw data $\mathbf{z}_0$ instead of taking the observational facts $\mathbf{e}$ as objectively given. Using the piece-meal activities involved as well as the strategies used in successful experiments, Hacking (1983) argued persuasively against the theory-dominated view of experiment in science, and made a strong case that in scientific research an experiment can have a 'life of its own' that is largely independent of 'large-scale theory'. From the econometric perspective this was a move in the right direction, but there was still some way to go to relate it to non-experimental data. This link was provided by Mayo (1996), p. 7, who broadened the notion of an experiment:

> "I understand "experiment," ... far more broadly than those who take it to require literal control or manipulation. Any planned inquiry in which there is a deliberate and reliable argument from error may be said to be experimental."

In addition, she fleshed out three crucial insights arising from new experimentalism:

> "1. Understanding the role of experiment is the key to circumventing doubts about the objectivity of observation.
>
> 2. Experiment has a life of its own apart from high level theorizing (pointing to a local yet crucially important type of progress).
>
> 3. The cornerstone of experimental knowledge is its ability to discriminate backgrounds: signal from noise, real effect from artifact, and so on." (ibid., p. 63)

In her attempt to formalize these insights into a coherent epistemology of experiment, she proposed the error statistical account, whose underlying reasoning is based on a refinement of the F-N-P frequentist approach to statistical inference; the name stems from its focus on error probabilities; see Mayo and Spanos (2008).

Contrary to the Popperian and Growth of Knowledge traditions' call for 'going bigger' (from theories to paradigms, to scientific research programs and research traditions), in order to deal with such problems as theory-laden observation, underdetermination and Duhemian ambiguities, Mayo argued that theory testing should be piece-meal and 'go smaller':

> "... in contrast to the thrust of holistic models, I take these very problems to show that we need to look to the force of low-level methods of experiment and inference. The fact that theory testing depends on intermediate theories of data, instruments, and experiment, and that the data are theory laden, inexact and "noisy", only underscores the necessity for numerous local experiments, shrewdly interconnected." (ibid., p. 58)

The error statistical account can deal with link (A) – from theory to testable hypotheses, by proposing a hierarchy of models, primary (theory), experimental (structural) and data (statistical) models, aiming to bridge the gap between theory and data in a piecemeal way that enables the modeler to ' probe and learn from error' at each stage of the modeling (ibid., p. 128). It can also deal with link (B) – from raw data to reliable evidence, by allowing data $\mathbf{z}_0$ to have 'a voice of its own' in the context of a validated statistical model; this provides the observational 'facts' (e) pertinent for theory appraisal. Moreover, it constitutes a philosophical/ methodological framework which has an inherent affinity to the ways practicing econometricians understood the problem in the context of the frequentist inductive inference. Indeed, it addresses link (C) – confronting testable hypotheses with evidence, by providing a coherent framework that deals effectively with the question: 'When do data $\mathbf{z}_0$ provide evidence for or against H?

This question can be answered unambiguously only when the model-based statistical testing associated with F-N-P frequentist inference is supplemented with a post-data assessment based severity evaluations. This provides an inferential construal of tests, based on their capacity to detect different discrepancies with data $\mathbf{z}_0$, and can be used to address the fallacies of acceptance/rejection; see Mayo and Spanos (2006).

The fundamental intuition underlying the error-statistical account of evidence is that: if a hypothesis H 'passes' a test T with data $\mathbf{z}_0$, but T had very low capacity to detect departures from H when present, then $\mathbf{z}_0$ does not provide good evidence for the verity of H; its passing T with $\mathbf{z}_0$ is not a good indication that H is true. Learning from error, according to Mayo (1996), amounts to deliberate and reliable argument from error based on *severe testing*:

> "... a testing procedure with an overwhelmingly good chance of revealing the presence of specific error, if it exists — but not otherwise." (p. 7).

Mere fit is insufficient for $\mathbf{z}_0$ to pass H severely, such a good fit must be something very difficult to achieve, and so highly improbable, were H to be in error. More importantly, the fact that $\mathbf{z}_0$ were used to arrive at a good fit with $H(\mathbf{z}_0)$ does not preclude counting $\mathbf{z}_0$ as good evidence for $H(\mathbf{z}_0)$ – it all depends on whether the procedure for arriving at $H(\mathbf{z}_0)$ would find evidence erroneously with very low probability.

As a framework for inductive inference error statistics enhances the F-N-P framework in several different ways, the most crucial being the following (see Mayo and Spanos, 2008):

(i) Emphasizing the learning from data (about the phenomenon of interest) objective of empirical modeling.
(ii) Paying due attention to the validity of the premises of induction.
(iii) Emphasizing the central role of error probabilities in assessing the reliability (capacity) of inference, both pre-data as well as post-data.
(iv) Supplementing the original F-N-P framework with a post-data assessment of inference in the form of severity evaluations in order to provide an inferential construal of tests.
(v) Bridging the gap between theory and data using a sequence of interconnected models: primary, experimental and data models.

(vi) Actively encouraging a thorough probing of the different ways an inductive inference might be in error, by localizing the error probing in the context of the different models.

2.3 Error statistics and model-based induction

The statistical underpinnings of the error statistical approach are squarely within the frequentist inference framework pioneered by Fisher (1922). Fisher initiated the recasting of statistical induction, by turning the prevailing strategy of commencing with the data in search of a descriptive model on its head. He viewed the data $\mathbf{z}_0:=(z_1, z_2, \ldots, z_n)$ as a realization of:

(a) a 'random sample' from (b) a pre-specified 'hypothetical infinite population', formalizing both in purely probabilistic terms. He formalized the notion of a 'random sample' as a set of Independent and Identically Distributed (IID) random variables $\mathbf{Z}_0:=(Z_1, Z_2, \ldots, Z_n)$ giving rise to data $\mathbf{z}_0$, and the 'infinite hypothetical population in the form of a distribution function $f(\mathbf{z}; \boldsymbol{\theta})$, indexed by a set of unknown parameter(s) $\boldsymbol{\theta}$. He then combined the two to define the notion of a parametric statistical model. He defined the problem of specification as the initial choice of the statistical model so as to ensure that the data constitute a 'typical realization':

"The postulate of randomness thus resolves itself into the question, 'Of what population is this a random sample?"' (Fisher, 1922, p. 313), emphasizing the fact that: 'the adequacy of our choice may be tested posteriori' (1922, p. 314).

The revolutionary nature of Fisher's recasting of induction is difficult to exaggerate because he rendered the vague notions of 'uniformity of nature' and 'representativeness of the sample' into testable probabilistic assumptions, and his concept of a statistical model provided the cornerstone for a new form of statistical induction which elucidated the ambiguities and weaknesses of the pre-Fisher perspective; see Spanos (2008).

The frequentist approach to statistics was pioneered by Fisher (1921; 1922; 1925; 1935) and continued with Neyman and Pearson (1928; 1933); the Neyman-Pearson contributions in framing hypothesis testing in model-based inductive terms were particularly crucial for the development of the new paradigm – see also Neyman (1950; 1977), Pearson (1966). Although a detailed discussion of this Fisher-Neyman-Pearson (F-N-P) frequentist paradigm, spearheaded by model-based statistical induction, is beyond the scope of this paper (see Spanos 2006c), it is important to state some of its key features for the discussion that follows:

(i) the chance regularity (stochasticity) in statistical data is inherent – it is a reflection of the probabilistic structure of the observable stochastic processes underlying the data –

(ii) this chance regularity is captured by a prespecified statistical model chosen so as to render the data a truly typical realization of the generic stochastic process specified by this model, and

(iii) ascertainable error probabilities – based on the prespecified statistical model

– provide the corner-stone for assessing the optimality and reliability of inference methods; see Neyman (1977).

Unfortunately, the relationship between statistical and substantive information has not been adequately addressed by the F-N-P approach and remains an issue to this day; see Lehmann (1991) and Cox (1991). In the next section it is argued that the sequence of interlinked models mentioned above (theory, structural and statistical) can be used to shed ample light on this issue.

3 Error statistics and empirical modeling in economics

3.1 'Statistical knowledge' with 'a life of its own'

Spanos (1986, 1988) proposed a modeling framework for econometrics, named probabilistic reduction, which largely overlaps with the error statistical account in Mayo (1996), but is lacking the post-data severity component. The most surprising overlap is the sequence of interlinked models aiming to provide a bridge between theory and data as shown in figure 1.

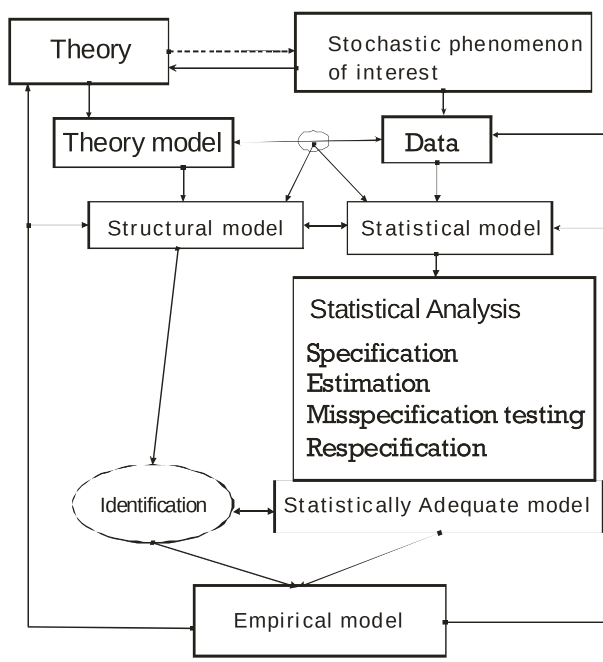

Fig. 1: Sequence of interlinked models in empirical modeling

Although the sequence of models, theory, structural and statistical (see figure 1) were motivated primarily by the problems and issues relating to econometric modeling, there is a direct analogy between these and Mayo's primary, experimental and data models, respectively. The main difference is one of emphasis where Spanos (1986) provides a more formal and detailed account on what a statistical model is and how its validity vis-à-vis the data is secured using thorough misspecification testing and respecification, in the context of the F-N-P statistical framework. A validated statistical model gives rise to statistical knowledge with 'a life of its own' in direct analogy to Mayo's (1996) experimental knowledge; see also Spanos (1995, 1999).

Although a detailed discussion of the more formal aspects of statistical model specification, misspecification testing and respecification (figure 1), designed to secure a statistically adequate model, are beyond the scope of the present paper (see Spanos, 2006a-c), it is important to discuss briefly how the above interlinked models can be used to deal effectively with bridging over the three interrelated modeling stages (A)-(C) mentioned in the introduction.

Using the different models (theory, structural and statistical) in figure 1 one can deal with link (A) – from theory to testable hypotheses, by framing ways to bridge the gap between theory and data that enmesh the structural and statistical models.

From the theory side of the bridge one constructs a theory model (a mathematical formulation) which might involve latent variables which correspond to no available data. In an attempt to connect the theory model to the available data, one needs to transform the theory model into an estimable (in light of the data) form; referred to as a structural model. The construction of structural models needs to take into account the gap between the concepts envisaged by the theory model (intentions, plans), and what the available data actually measure; see Spanos (1995) for further details.

With respect to rendering a theory model estimable, the problem facing an economist is not as different as the situation confronting a physicist wishing to test the kinetic theory of gasses. Due to the unobservability of molecules, the physicist resorts to heating the gas in order to measure the expansion of the gas under constant pressure, in order to test an observable implication of the theory in question.

The structural model contains a theory's substantive subject matter information in light of the available data $\mathbf{z}_0$. How does the structural connect to the statistical model?

3.1.1 Statistical models

The statistical model $\mathcal{M}_\theta(\mathbf{z})$ is built exclusively using information contained in the data and is chosen in such a way so as to meet two interrelated aims:

(a) to account for the chance regularities in data $\mathbf{z}_0$ by choosing a probabilistic structure for the stochastic process underlying $\mathbf{z}_0$ so as to render $\mathbf{z}_0$ a typical realizations thereof, and

(b) to parameterize this probabilistic structure of in the form of an adequate statistical model $\mathcal{M}_\theta(\mathbf{z})$ that would embed (nest) the structural model in its context.

Objective (a) is designed to deal with link (B) – from raw data to reliable evidence, using the notion of a validated statistical model: $\mathcal{M}_\theta(\mathbf{z}) = \{f(\mathbf{z};\boldsymbol{\theta}),\ \boldsymbol{\theta} \in\Theta\}$, $\mathbf{z} \in\mathbb{R}^n_Z$, where $f(\mathbf{z};\boldsymbol{\theta})$ denotes the joint distribution of the sample $\mathbf{Z}_0:=(Z_1, Z_2, \ldots, Z_n)$. Formally, a $\mathcal{M}_\theta(\mathbf{z})$ can be viewed as a reduction from $f(\mathbf{z};\boldsymbol{\theta})$ after imposing a certain probabilistic structure on that reflects the chance regularity patterns in data $\mathbf{z}_0$; Spanos (1989, 1999).

To give some idea as to what this entails, consider the simple Normal model given in table 1, comprising a statistical Generating Mechanism (GM), and the probabilistic assumptions [1]-[4] defining a Normal Independent and Identically Distributed (NIID) process. This model, specified in terms of , can be formally viewed as a reduction from the joint distribution as follow:

$$f(x_1, x_2, ..., x_n; \boldsymbol{\theta}) \quad f(x_1, x_2, ..., x_n; \boldsymbol{\theta}) \overset{\mathsf{I}}{=} \prod_{k=1}^{n} f_k(x_k; \boldsymbol{\theta}_k) \overset{\mathsf{IID}}{=} \prod_{k=1}^{n} f(x_k; \boldsymbol{\theta}), \quad \mathbf{x} \in \mathbb{R}^n.$$

Table 1 - Simple Normal Model

Statistical GM:		$X_k = \mu + u_k, \quad k \in \mathbb{N}.$	
[1]	Normality:	$X_k \backsim \mathsf{N}(.,.),$	$k \in \mathbb{N}.$
[2]	Constant mean:	$E(X_k) = \mu,$	$k \in \mathbb{N}.$
[3]	Constant variance:	$Var(X_k) = \sigma^2,$	$k \in \mathbb{N}.$
[4]	Independence:	$\{X_k, \ k \in \mathbb{N}\}$ independent process	

This is a purely probabilistic construct which depends only on the joint distribution of the stochastic process $\{X_k, \ k \in \mathbb{N}\}$, say $f(\mathbf{x}; \boldsymbol{\theta})$, where the unknown parameters are $\theta := (\mu, \sigma^2)$. Using analogical reasoning one can infer that the data in figure 2 could be realistically viewed as a realization of a NIID process $\{X_k, \ k \in \mathbb{N}\}$, but figure 3 cannot because it exhibits cycles which indicate the presence of positive Markov dependence; see Spanos (1999), ch. 5.

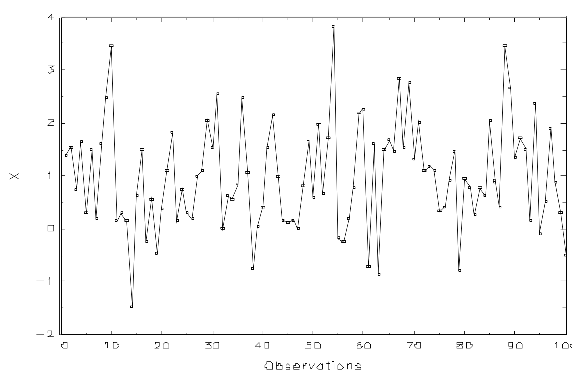

Fig. 2: A realization of a NIID process

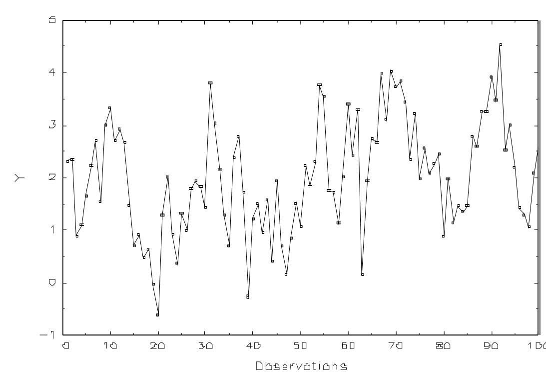

Fig. 3: A realization of a Normal, Markov, stationary process

3.1.2 Spurious correlation/regression

When X_k is a vector with m components, say $\mathbf{X}_k := (X_{1k}, X_{2k}, \ldots, X_{mk})$, the statistical model in table 1 becomes the simple multivariate Normal model, with assumptions [2]-[3] taking the form: $E(\mathbf{X}_k) = \boldsymbol{\mu}$, $Cov(\mathbf{X}_k) = \Omega$. This model underlies several statistical procedures widely used in the social sciences, including correlation, linear regression, factor analysis and principal component analysis.

The spurious statistical inference results that are often considered endemic and inevitable when using observational data, can often be explained away as statistically meaningless when any of the assumptions [1]-[4] are invalid for data $x_k, k=1,\ldots,n$.

Example: trending data. For the data in figures 4-5 assumption [2] is false, they exhibit trends ($E(X_k)$ and $E(Y_k)$ change (increase) with k), and thus the sample correlation:

$$\widehat{\rho}=\frac{\sum_{k=1}^{n}(X_k-\overline{X})(Y_k-\overline{Y})}{\sqrt{\sum_{k=1}^{n}(X_k-\overline{X})^2\cdot\sum_{k=1}^{n}(Y_k-\overline{Y})^2}}$$

provides a terrible (biased, inconsistent, etc.) estimator of the 'population' correlation coefficient:

$$\rho=\frac{E[(X_k-E(X_k))(Y_k-E(Y_k))]}{\sqrt{E[X_k-E(X_k)]^2E[Y_k-E(Y_k)]^2}},$$

rendering any tests of significance statistically meaningless. It is important to emphasize that both the variances and covariance (hence the correlation) are all defined in terms of deviations from the 'true' mean, whatever that happens to be; not the mean one happens to assume. This is because $\overline{X}=\frac{1}{n}\sum_{k=1}^{n}X_k$ and $\overline{Y}=\frac{1}{n}\sum_{k=1}^{n}Y_k$ are 'bad' (inconsistent) estimators of $E(X_k)=\mu_x(k)$ and $E(Y_k)=\mu_y(k)$, respectively. Intuitively, when the means $(E(X_k),E(Y_k))$ are changing with k as in figures 3-4, the sample mean deviations $\sum_{k=1}^{n}(X_k-\overline{X})$ and $\sum_{k=1}^{n}(Y_k-\overline{Y})$ (used in evaluating $\widehat{\rho}$) are highly over-inflated because they are measured from fixed points $\left(\overline{X},\overline{Y}\right)$, giving rise to apparent high correlations, when in fact they yield meaningless numbers that have nothing to do with the probabilistic structure of the underlying process since the assumed probabilistic structure (ID) is false; hence, there is nothing to explain (away) using substantive information. To make matters worse, the reliability of inference exacerbates as $n\to\infty$.

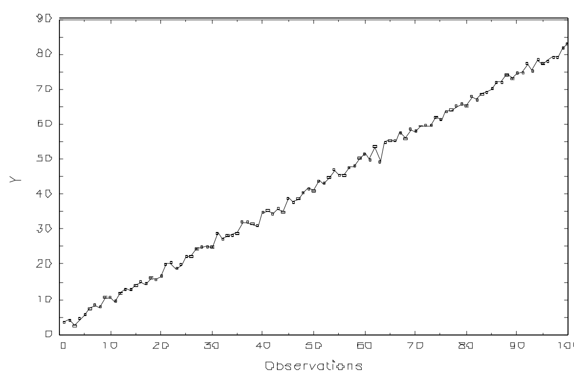

Fig. 4: A realization of a NI, non-ID process

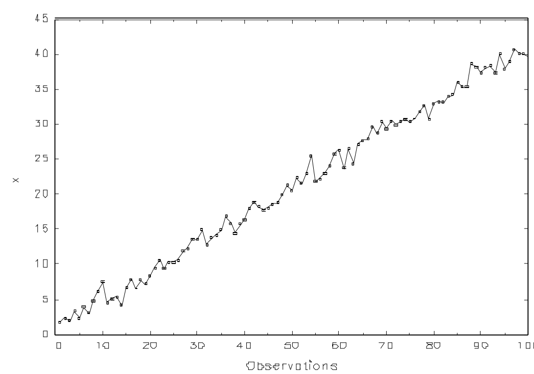

Fig. 5: A realization of a NI, non-ID process

For the same reason (statistical misspeci-fication), when the data in figures 4-5 are used to estimate a linear regression model of the form:

$$y_k=\beta_0+\beta_1x_k+u_k,\quad k=1,2,...,n,$$

least-squares estimation will give rise to a statistically meaningless t-ratio and R^2 :

$$\tau(\beta_1)=\frac{\widehat{\beta}_1}{s\sqrt{\sum_{k=1}^{n}(x_k-\overline{x})^2}},\quad R^2=\frac{(n-2)s^2}{\sum_{k=1}^{n}(y_k-\overline{y})^2},$$

where $s^2=\frac{1}{n-2}\sum_{k=1}^{n}(y_k-\widehat{\beta}_0-\widehat{\beta}_1 x_k)^2$, since both statistics $(\tau(\beta_1), R^2)$ involve deviations from constant sample means $(\overline{x},\overline{y})$ which are inappropriate since their true means are changing with k; see Spanos (1999).

The prespecified statistical model $\mathcal{M}_\theta(\mathbf{z})$ accounts for the statistical (chance) regularities in the data $\mathbf{z}_0$ only when its probabilistic assumptions are valid for $\mathbf{z}_0$. Validation of a statistical model takes the form of thorough Mis-Specification (M-S) testing of the assumptions comprising it (e.g. assumptions [1]-[4]; see Spanos, 1999, ch. 15). A statistically adequate (validated) model has 'a life of its own' in the sense that it accounts for the statistical regularities in the data $\mathbf{z}_0$, independently of any substantive information contained in the structural model in question.

3.1.3 Testing substantive information

The objective (b), associated with the specification of a statistical model $\mathcal{M}_\theta(\mathbf{z})$, has to do with the embedding of the structural model in its context, connecting the two sides of the theory-data gap and allowing one to assess the validity of the substantive information vis-à-vis data $\mathbf{z}_0$. The main objective of the embedding is to render the structural (estimable) model a special case (reparameterization/restriction) of the statistical model. Formally this takes the form of structural parameter identification: the structural parameters $\boldsymbol{\varphi}$ are uniquely defined by the statistical parameters $\boldsymbol{\theta}$. Usually, there are more statistical than structural parameters and that enables one to parameterize the substantive information in the form of overidentifying restrictions:

$$H_0\colon \mathbf{G}(\boldsymbol{\varphi},\boldsymbol{\theta})=\mathbf{0}, \text{ vs. } H_1\colon \mathbf{G}(\boldsymbol{\varphi},\boldsymbol{\theta})\neq\mathbf{0},$$

which can be tested in the context of the statistical model $\mathcal{M}_\theta(\mathbf{z})$ in question; Spanos (1990).

How these restrictions can be properly appraised 'evidentially' is the concern of link (C) – confronting testable hypotheses with evidence. When the substantive information in h passes a severe test in the context of a validated statistical model (**e**), one can deduce the substantive adequacy of the structural model vis-à-vis data $\mathbf{z}_0$. Then, one can impose h in order to fuse the substantive and statistical information in the context of an empirical model (see figure 1). The empirical model inherits its statistical and substantive meaningfulness from the statistical and structural models, respectively.

3.1.4 Statistical knowledge

Of crucial importance is that a statistically adequate model:

(i) is independent (separate) from the substantive information, and therefore,

(ii) it can be used to provide the broader inductive premises for evaluating its adequacy.

The independency stems from the fact that the statistical model is erected on purely probabilistic information by capturing the 'chance regularities' exhibited by data $\mathbf{z}_0$. In figures 1-2 the chance regularities have nothing to do with any substantive subject matter information since data $\mathbf{z}_0$ are viewed as a realization of a generic stochastic processes , irrespective of what substantive variable Z_t quantifies. In this sense, a statistically adequate model provides a form of statistical knowledge, analogous to what Mayo calls experimental knowledge, against which the substantive information will be appraised. As remarked by Mayo (1996), pp. 129-130, the sequence of models (theory, structural, statistical) performs double duty:

> "...it provides a means for spelling out the relationship between data and hypotheses – one that organizes but does not oversimplify actual scientific episodes – and organizes the key tasks of a philosophy of experiment."

The notion of 'statistical knowledge', like experimental knowledge, is instrumental in discriminating backgrounds: "...signal from noise, real effect from artifact, and so on." (ibid., p. 63). The notion of statistical knowledge formalizes the sense in which data $\mathbf{z}_0$ has 'a voice of its own', separate from the one ideated by a theory. This is crucial for theory testing because it ensures the reliability and scientific objectivity of the theory-data confrontation.

3.2 Probing for different 'local' errors

The typology of different models, as exemplified in figure 1, was designed to delineate questions concerning different types of errors and render their probing much more effective than the traditional approach of lumping them together in one overall error term. These models can be used to illustrate the error statistical encouraging of a thorough probing of the different ways an inductive inference might be in error, by localizing the error probing in the context of the different models; see Mayo (1996). The reliability of evidence is assessed at all levels of modeling by using error-statistical procedures based on learning from error reasoning. When any of these potential errors are present, the empirical evidence will be rendered untrustworthy.

In the context of a *statistical model* the primary sources of error are inaccurate data and statistical misspecification. Let us briefly discuss the two.

(I) Inaccurate data: data $\mathbf{z}_0$ are marred by systematic errors imbued by the collection/compilation process.

The inaccuracy and inadequacies of economic data as of the late 1950s has been documented by Morgenstern (1950/1963). Since then the accuracy of economic data has been improving steadily, but not enough attention has been paid to issues concerning how the collection, processing and aggregation of data in economics might imbue systematic errors that can undermine their statistical analysis; see Abadir and Talmain, (2002).

(II) Statistical misspecification: some of the probabilistic assumptions comprising the statistical model (premises of induction) are invalid for data $\mathbf{z}_0$.

A misspecified statistical model gives rise to unreliable inferences because the actual error probabilities are very different from the nominal ones — the ones assumed to hold if the premises are true. Applying a .05 significance level t-test, when the actual type I error is .95, renders the test highly unreliable; see Mayo (1996), Spanos (2005).

In the context of a *structural (estimable) model* the relevant source of error is:

(III) Incongruous measurement: data $\mathbf{z}_0$ do not measure the concepts $\boldsymbol{\xi}$ envisioned by the particular theory (e.g. intentions vs. realizations – see Spanos (1995).

This is a fundamental issue because typically theory models are built on intentions on behalf of economic agents but the available data measure realizations of ongoing convoluted processes. Moreover, the overwhelming majority of economic data are collected by government agencies and private institutions for their own purposes and not the modelers themselves.

In the context of the theory (and empirical) model the primary source of error is:

(IV) Substantive inadequacy: the circumstances envisaged by the theory differ 'systematically' from the actual Data Generating Process (DGP). This might involve investigating the validity of ceteris paribus clauses causal claims, as well as search for missing confounding factors (see Guala, 2005, Hoover, 2001a). Substantive adequacy concerns the extent to which the empirical model 'captures' the aspects of the reality it purports to explain, shedding light on the phenomenon of interest, i.e. 'learning from data'. It concerns the extent to which the inferences based on the empirical model are germane to the phenomenon of interest.This issue is particularly relevant when modeling with experimental data.

3.3 Experimental economics and theory testing

The notion of experimental knowledge and the idea of an experiment having 'a life of its own' have recently been introduced into experimental economics in an attempt to provide sound philosophical underpinning to endeavors to test economic theories, by designing and conducting experiments pertaining to economic behavior under laboratory conditions; see Guala (2005). Some notable pioneers in experimental economics, like Vernon Smith (2007, ch. 13), came to appreciate that the error statistical perspective provides a promising framework wherein a lot of the issues experimental economists have been grabbling with can find a natural home and a number of systematic ways to address them. After declaring that: "Experimental knowledge drives the experimental method" (pp. 305), he argued that the problem of understanding the phenomenon that came to be known as 'Brownian motion':

"... was finally solved by drawing on the extensive bag of experimental tricks, tools and past mistakes that constitute "a log of the extant experimental knowledge of the phenomena in question." (Mayo, 1996, p. 240)." (ibid., p. 308)

Similarly, Sugden (2008) acknowledged that: "the account of scientific method that comes closest to what I have in mind is that of Deborah Mayo (1996)."

Viewed in the context of the error statistical perspective discussed above, experimental economics attempts to bridge the theory-data gap by bringing the data closer to the theory; see figure 1. Instead of using observational data, which are the result of a highly complicated on-going actual DGM which involves numerous influencing factors, one generates experimental data under 'controlled' conditions which aim to correspond closely to the conditions envisaged by economic theory. This data can then be used to evaluate theoretical predictions concerning economic behavior. In this sense, the structural (estimable) model – in view of the data – is often much closer to the theory model than in the case of observational data. Despite this important difference, the statistical modeling of experimental data in economics – with a view to learn about economic phenomena – raises the same issues and problems of reliability as observational data with only differences in emphasis.

To be more specific, the same sequence of interconnecting models (theory, structural, statistical and empirical –see figure 1) are equally elucidating when modeling experimental data, raising very similar issues as they pertain to probing for errors at different levels of modeling and the resulting reliability of inductive inferences. In particular, the various 'canonical' errors (I-IV), mentioned above, are no less applicable to experimental data, but there are differences in emphasis. For example, (I) inaccurate data is easier to address directly in experimental economics because the experimenter, the data gatherer and modeler are usually one and the same person (or group). Moreover, hypotheses of interest in experimental economics are usually rendered testable by embedding them into a statistical model (implicitly or explicitly), raising the problem of (II) statistical misspecification. The fact that the data are the result of a 'controlled' experiment makes it easier in practice to ensure that the data can be realistically viewed as a realization of a random sample (IID), but that, in itself, does not suffice to secure the statistical adequacy of the estimated model, and thus the statistical reliability of inference. Experimental design techniques, such as randomization and blocking, are applied in particular situations, but statistical adequacy – secured by thorough M-S testing—provides the only way to establish the absence of systematic effects in a particular data. Similarly, the problem of (III) incongruous measurement is an issue that is easier to deal with in the context of generating one's own experimental data.

On the other hand, the problem of (IV) Substantive inadequacy is often more difficult to address when modeling experimental data because one needs to demonstrate that the inferences drawn from the experimental data (generated in lab) are germane to, and shed light on, the phenomenon of interest. This problem is particularly critical in experimental economics where it has unfolded in a variety of guises; Guala (2005) provides an illuminating discussion of these issues under the banner of 'external validity': legitimately generalize inferences based on experimental data – generated in a lab's artificial set up – to the real-world phenomenon in question. This issue lies at the heart of the problem of theory testing in economics because in the absence of external validity it's not obvious how one would assess the ability of the theory in question to explain (or shed light on) the phenomenon of interest.

A particularly interesting methodological issue that has arisen in experimental economics concerns several 'exhibits' that have been accumulated over the last decade or so, considered as 'anomalies' for mainstream economic theory; see Sugden (2005, 2008). These 'exhibits' are essentially viewed as inductively established 'replicable empirical effects' that, in the eyes of experimental economists, constitute 'experimental knowledge'. As argued by Mayo (1996, 2008), establishing experimental knowledge requires a thorough probing of all the different ways such claims can be in error at all levels of modeling. In particular, replicability, by itself, does not suffice in such a context because, for instance, all the experiments conducted might ignore the same confounding factor giving rise to the detected effect. In addition, statistical adequacy constitutes a necessary condition in order to eliminate the possibility that the 'exhibit' in question is not a statistical artifact. Moreover, going from statistical significance to substantive significance raises crucial methodological issues (the fallacies mentioned above) which need to be addressed before such 'exhibits' can qualify as experimental knowledge.

4 Theory-data confrontation before the mid 20th century

4.1 Theory-data confrontation in Classical Economics

Since the founding of economics (originally called political economy) as a separate scientific discipline by Adam Smith (1776), the notion of a 'theory' was understood as a system of propositions deductively derived from certain premises defined in terms of particular postulates (principles). One of the earliest examples of initial postulates was given by Senior (1836) who claimed that political economy is ultimately based only on four general postulates:

> "(1) every man desires to maximize wealth with as little sacrifice as possible, (2) population is limited by the available resources, (3) capital enhances the productivity of labor, and (4) agriculture exhibits diminishing returns." (p. 26)

An early example of economic theory propositions was given by Ricardo (1817) and concerned the 'long-run' equilibrium tendencies in key variables:

> (a) the price of corn would tend to increase, (b) the share of rent in the national income would tend to go up, (c) the rate of profit on capital would tend to fall and (d) the real wages would tend to stay constant.

Methodological differences among political economists during the 19th century were mainly focused on the relationship between the initial postulates and the deductively derived propositions on one side, and real-world data on the other. On one side Malthus (1836) argued in favor of anchoring the initial postulates on observation and then confronting the derived propositions with real-world data. On the other side, Ricardo (1817) considered the initial postulates as self-evident truths, not subject to empirical scrutiny, and viewed the relationship between theory and data as based on one's 'impressions' concerning the broad agreement between the deductively derived propositions and 'tendencies' in historical data. It is important to stress that economic data began to be collected from the mid 17th century; see Porter (1836).

Political economy, from its very beginnings, was entangled in philosophical/methodological discussions concerning the nature and structure of its theories and methods. Methodological discussions concerning induction vs. deduction, the logical structure of economic theories, the status of economic postulates and deductively inferred propositions, verification and testing of such postulates and propositions, were considered an integral part of the broader discourse concerning 'learning about economic phenomena of interest'; see Redman (1997). Outstanding examples of such discussions during the 19th century include: Malthus (1936), Senior (1836), McCulloch (1864), Mill (1844; 1871), Cairnes (1888) and Keynes (1891). Political economists who are equally well-known as philosophers include Jevons (1873) and Mill (1884).

During the first half of the 19th century the idea of 'observation' as a collection of 'facts' was very unclear and fuzzy because real-world data usually come in the form of historical data series which included numerous observations on several different variables without any obvious way to summarize them into a form that can be used for theory appraisal. The descriptive statistics tradition did not form a coherent body of knowledge until the end of the 19th century, and even then it did not provide a way to deal effectively with links (B)-(C). Viewed retrospectively, the theory-data confrontation during the 18th and 19th centuries amounted to nothing more than pointing out that one's 'impressions' concerning 'tendencies' in historical data do not seem to contradict the deductively derived propositions in question.

This was clearly a highly unsatisfactory state of affairs, rendering theory testing against data an unavailing exercise with very little credibility. Indeed, apparent empirical falsification did not seem to have any negative effect on the credibility of certain economic theories. As argued by Blaug (1958), when Ricardo's propositions (a)-(d) are appraised in terms of these observational implications they constitute a dismal failure; none of them were borne out by the historical data up to the 1850s. Despite their empirical inadequacy, Ricardo's theories are viewed as a great success story by the subsequent economics literature to this day:

> "(1) As a theory and method, the Ricardian paradigm provided the basis for further advances in logical deduction, and (2) it also had the practical value of providing a basis for economic policy." (see Ekelund and Hebert, 1975, p. 110)

The clarity and economy of the logical deduction from a few initial postulates, to the neglect of the role of the data, captured the imagination of the overwhelming majority of political economists during the second half of the 19th century.

Mill (1844; 1871) was instrumental in shaping a methodological framework for Ricardo's deductive turn, which justified the emphasis on deductively derived propositions and the subordinate role attributed to data. He argued that causal mechanisms underlying economic phenomena are too complicated – they involve too many contributing factors – to be disentangled using observational data. This is in contrast to physical phenomena whose underlying causal mechanisms are not as complicated – they involve only a few dominating factors – and the use of experimental data can help to untangle them by 'controlling' the 'disturbing' factors. Hence, economic theories can only establish general tendencies and not precise enough implications whose validity can be assessed using observational data. These tendencies are framed in terms of the primary causal contributing factors with the rest of the numerous (possible) disturbing factors relegated to ceteris paribus clauses whose appropriateness cannot, in general, be assessed using observational data. This means that empirical evidence contrary to the implications of a theory can always be explained away as due to counteracting disturbing factors. As a result, Mill (1944) rendered the theory-data gap unavoidable, and attributed to the data the auxiliary role of investigating the ceteris paribus clauses in order to shed light on the unaccounted by the theory disturbing factors which prevent the establishment of the tendencies predicted by the theory in question.

The subordinate role of data in theory appraisal was relegated even further by Cairnes (1875/1888) by pronouncing data, more or less, irrelevant for appraising the truth of deductively established economic theories. He turned the weaknesses acknowledged by Mill on their head and claimed that the focus on the deductive component of economic modeling, as well as the gap between theory and data, were in fact strengths not weaknesses: given the 'self-evident truth' of the initial postulates, and the deductive validity of the propositions which follow, the question of verification using data does not even arise because the truth of the premises ensures the truth of the conclusions! As a result, so the argument goes, the deductive nature of economic theories bestows upon them a superior status that even physical theories do not enjoy. This is because the premises of Newtonian mechanics are not 'self-evident truths', as in economics – established introspectively via direct access to the ultimate causes of economic phenomena – but mere inductive generalizations that need to rely on experimentation and inductive inferences which are known to be fallible; see Cairnes (1888), pp. 72-94.

Keynes (1890) made an attempt to summarize the methodological discussions, as well as synthesize the various views, that dominated political economy during the 19th century. His synthesis appears to be closer to Mill's position with an added emphasis on the potential role statistics in theory assessment. Despite these pronouncements, his discussion sounded more like leap-service and did nothing to promote the use of statistics in theory appraisal vis-a-vis data.

4.2 Theory-data confrontation in Neoclassical Economics

The marginalist revolution in economics, which began in the 1870s, brought with it a mathematical turn in economic theorizing which gave rise to economic models (Bowley, 1924) based on optimization at the level of the individual agent (consumer or producer); see Backhouse (2002). Marshall (1891), in a most influential textbook, retained Mill's methodological stance concerning the pre-eminence of theory over data in economic theorizing, and demonstrated how neoclassical economics can be erected using marginal analysis based on calculus.

During the first part of the 20th century the majority of economists, lead by Robbins (1935), reverted to Cairnes' extreme methodological position which pronounced data, more or less, irrelevant for appraising the truth of deductively established propositions. Robbins was aware of the development of statistical techniques since the 1870s, but dismissed their application to theory appraisal in economics on the basis of the argument that such techniques are only applicable to data which can be considered as 'random samples' from a particular population. Since there were no experimental data in economics which could potentially qualify as such a 'random sample', statistical analysis of economic data had no role to play in theory assessment. This argument stems from ignorance concerning the applicability and relevance of modern statistical methods, but unfortunately the ignorance lingers on to this day (Mirowski, 1994).

These apparently extreme positions reflect the methodological/ philosophical presuppositions thought to underlie and justify the mainstream economic theorizing and modeling of the period. Actual or perceived characteristics (and limitations) of theory and of data in economics are often used as the basis for erecting a conception of the goals of economic methodology that may be satisfied by a science with those features.

Despite several dissenting voices like Hutchison (1938), who used ideas from logical positivism and Karl Popper to restate Malthus' thesis that both the initial postulates and the deductively derived propositions should be subjected to empirical scrutiny, the prevailing view in economics during the first half of the 20th century adopted the pre-eminence of theory viewpoint and gave data the subordinate role of availing the 'quantifying of theoretical relationships'. The problem, as it was seen at the time, was that economic theory models, derived on the basis of self-evident postulates, are necessarily valid, if only one can verify them by bestowing 'empirical content' onto them using the 'right' data, and the appropriate statistical techniques.

In the 1940s methodological discussions in economics focused primarily on the realism of the initial postulates. An example of this was whether the assumption that firms 'maximize profits' is realistic or not; see Machlup (1963). Friedman (1953) settled that issue for economists by arguing that the realism of a theory's initial postulates is irrelevant, and the success of a theory should be solely assessed in terms of its empirical predictive ability. Its appeal among practicing economists can be explained by the fact that Friedman's basic argument justified what most economists were doing, including the pre-eminence of theory over data: "we cannot perceive "facts" without a theory" (ibid., p. 34). The paper generated a huge literature, both in the methodo-logy of economics (Caldwell, 1984) and philosophy of science; Nagel (1963), Musgrave (1981).

5 Early econometrics and theory testing

5.1 Early pioneers in econometrics

By the early 20th century the descriptive statistics tradition (see Mills, 1924), together with the development of new techniques such as regression, correlation and related curve fitting methods based on least-squares associated with Galton, Karl Pearson and Yule, offered economists more powerful tools, as well as a statistical framework, for theory appraisal in economics.

It is not surprising that the first attempts at theory appraisal vis-à-vis real-world data focused primarily on the demand/supply model which was popularized by Marshall's (1891) influential textbook, and was seen as the cornerstone of the newly established neoclassical tradition. These early empirical studies of demand/supply (see Hendry and Morgan, 1995, Morgan, 1990), when viewed retrospectively from the error statistical perspective, seem inexorably naïve. They seem to under-appreciate the enormity of the gap between the theory (intentions to buy/sell corresponding to hypothetical prices), and the available data on quantities transacted and the corresponding prices; see Spanos (1995). Moreover, they appear to be unaware of the numerous possible errors, inaccurate data, statistical misspecification, incongruous measurement and substantive inadequacy, raised in section 3.2; see Spanos (2006a). As argued below, these weaknesses can be explained by the fact that (a) these empirical studies were dominated by the pre-eminence of theory viewpoint, which considered theory appraisal as simply the 'quantification of theoretical relationships' using data, and (b) the statistical framework associated with the descriptive statistics tradition (see Bowley, 1926, Mills, 1924) was inadequate for the task.

5.2 Ragnar Frisch and the Econometric Society

Notwithstanding the declarations of the founding statement of the Econometric Society in 1930 (Frisch, 1933, p. 106), the newly established econometric community viewed the theory-data confrontation as the 'quantification of theoretical relationships' in economics: theory provides the 'structural' relationships and data avail their quantification. This viewpoint went as far as to question the value of the newly established model-based approach to statistical induction associated the Fisher-Neyman-Pearson (F-N-P) sampling theory methods because it was thought to be inextricably bound up with agricultural experimentation. It was generally believed that these methods are relevant only for analyzing 'random samples' from experimental data; see Frisch (1934), p. 6. The prevailing viewpoint was that economic phenomena:

(i) are not amenable to the 'experimental method',
(ii) are influenced by numerous potential factors (hence the *ceteris paribus* clause),
(iii) are intrinsically heterogeneous (spatial and temporal variability), and
(iv) economic data are often vitiated by errors of measurement.

The intention was to develop a statistical framework which asserts the pre-eminence of theory perspective and could account for the perceived features (i)-(iii). This attitude led to an eclectic approach to statistical modeling and inference which was based on the pre-Fisher descriptive statistics paradigm but supplemented with Quetelet's (1842) scheme. This scheme viewed data exhibiting randomness (chance regularity) as comprising two different components:

(Q1) a *systematic (deterministic) component* (constant causes), determined by substantive information, and

(Q2) a random part which represents the *non-systematic error (accidental causes) component* (see Desrosières, 1998).

This viewpoint is clearly at odds with the Fisher-Neyman-Pearson perspective which considers the stochastic nature of data to be inherent; see section 2.3. However, for economists the descriptive statistics framework in conjunction with the Quetelet scheme had an alluring appeal because: (a) it preserves the pre-eminence of theory in the form of deterministic models, and (b) offers a general way to use modern statistical inference techniques by attaching IID errors to structural models. Indeed, by tacking IID error terms to structural models (c) renders the distinction between a statistical and a structural model virtually redundant. The prevailing view in the 1930s was that the 'quantification of theoretical relationships' using data was unproblematic because the cogency of the theory would ensure statistical validity. Frisch (1934) was the first to implement explicitly the error-tacking strategy in econometrics. He argued that observable variables $(X_{kt}, k{=}1, 2, \ldots, m)$ in economics can be decomposed into a latent systematic (deterministic) component and a random white-noise error:

$$X_{kt} = \mu_{kt} + \varepsilon_{kt},\;\; k = 1, 2, \ldots, m, \tag{1}$$

known as the *errors-in-variables* (or errors-of-measurement) scheme, where:

$$\text{[i] } E(\varepsilon_{kt}){=}0,\; \text{[ii] } E(\varepsilon_{kt}^2){=}\sigma^2,\; \text{[iii] } E(\varepsilon_{kt}\varepsilon_{js}){=}0,\; k \neq j,\; k, j{=}1, ..., m,\; t \neq s,\; t, s{=}1, ..., n, \tag{2}$$

Economic theory determines the deterministic relationships among the (latent) systematic components in the form of the system of m potential equations:

$$\alpha_{1k}\mu_{1t} + \alpha_{2k}\mu_{2t} + \cdots \alpha_{mk}\mu_{mt} = 0, \quad k = 1, 2, ..., m. \tag{3}$$

Combining (1) and (3) gives rise to the formulation:

$$\alpha_{1k}X_{1t} + \alpha_{2k}X_{2t} + \cdots \alpha_{mk}X_{mt} = \varepsilon_t, \quad t = 1, 2, ..., n. \tag{4}$$

Although this can be written as a regression, such a perspective is incorrect because it ignores the particular structure of the error term: is $\varepsilon_t = \alpha_{1k}\varepsilon_{1t} + \alpha_{2k}\varepsilon_{2t} + \cdots \alpha_{mk}\varepsilon_{mt}$. Hence, Frisch proposed a novel solution to the estimation problem in the form of 'curve fitting' (he called confluence analysis) viewed as a problem in vector space geometry. The idea was to determine r, the number of (exact) theoretical relationships among m latent variables:

$$\alpha_{1k}\mu_{1t} + \alpha_{2k}\mu_{2t} + \cdots \alpha_{mk}\mu_{mt} = 0, \quad k = 1, 2, ..., r, \tag{5}$$

where ($r \leq m < n$). Since the matrix $\mathbf{M} := [\mu_{kt}]_{k=1,...,m}^{t=1,...,n}$ is not observable, r needs to be determined by the corank of the data matrix $\mathbf{X} := [x_{kt}]_{k=1,...,m}^{t=1,...,n}$, which includes the errors; see Kalman (1982).

Frisch's scheme (1)-(3) gave rise to a statistical formulation which is clearly different from traditional regression. The idea was that such models distinguished empirical research in economics from other fields, including mainstream statistics, in order to custom tailor statistical inference and modeling for economic phenomena and the available data.

Viewed retrospectively from the error statistical perspective of section 3, the Frisch scheme had little chance to succeed as a general modeling strategy that would give rise to learning from the data about economic phenomena because:

(a) it embraced the pre-eminence of theory over data perspective and focused almost exclusively on the 'quantification of theoretical relationships',
(b) the bridging of the theory-data gap using the error-tacking strategy was clearly much too simplistic to provide adequate answers to this crucial problem,
(c) with exception of measurement error, it ignored the other possible errors, including incongruous measurement, statistical misspecification, and substantive inadequacy, and
(d) Its highly restrictive probabilistic perspective (linear/homoskedastic/static systems), and its statistical framework based on descriptive statistics and the Quetelet scheme, were totally inadequate for dealing with the links (A)-(C).

5.3 Trygve Haavelmo and frequentist inference

The first systematic attempt to introduce the Fisher-Neyman-Pearson (F-N-P) approach into econometrics was made by Haavelmo (1944). In his classic monograph he argued fervently in favor of adopting the new statistical inference, and addressed the concern expressed by Robbins (1935) and Frisch (1934), that such methods were only applicable to cases where the data can be viewed as 'random samples' from static populations. The development of the theory of stochastic processes – a sequence of dated random variables, say – by Kolmogorov and Khintchin in the early 1930s (see Doob, 1953) was both crucial and timely because it extended significantly the intended scope of the F-N-P approach beyond the original IID frame-up.

The Haavelmo (1944) monograph constitutes the best example of viewing the confrontation between theory and data in the context of bridging of the gap between theory and data, where both the theory and the data are accorded 'a life of their own'. As argued in Spanos (1989) it contains a wealth of methodological insights, which, unfortunately, had no impact on the subsequent developments of econometrics.

5.4 The Cowles Commission modeling perspective

The part of Haavelmo's monograph that had the biggest impact on the subsequent literature was his proposed technical 'solution' to the simultaneity problem. This was considered a major issue in economics because the dominating theory – general equilibrium – gives rise to multi-equation systems, known as the Simultaneous Equations Model (SEM).

To bring out the essence of the simultaneity problem consider a theory-based structural model concerning the behavior of two endogenous variables y_{1t} and y_{2t} :

$$\begin{aligned} y_{1t} &= \alpha_{10} + \alpha_{11} y_{2t} + \alpha_{12} x_{1t} + \alpha_{13} x_{2t} + \alpha_{14} x_{3t} + \varepsilon_{1t} \\ y_{2t} &= \alpha_{20} + \alpha_{21} y_{1t} + \alpha_{22} x_{1t} + \alpha_{23} x_{4t} + \alpha_{24} x_{5t} + \varepsilon_{2t} \end{aligned} \tag{6}$$

where the structural errors are assumed to have the following probabilistic structure:

$$\begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix} \backsim \mathsf{NIID}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \omega_{11} & \omega_{12} \\ \omega_{12} & \omega_{22} \end{pmatrix} \right), \ t=1,2,...,n. \tag{7}$$

The perspective underlying the specification of (6)-(7) constitutes a variant of Quetelet's scheme (Q1)-(Q2) (deterministic component plus random error), but replacing Frisch's errors-in-variables (1) with errors-in-equations (6). The problem, as perceived by the Cowles group at the time, was simply a technical one concerning how one can use the data $\mathbf{Z}_0 := (\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n)$, where $\mathbf{Z}_t = (y_{1t}, y_{2t}, x_{1t}, x_{2t}, \ldots, x_{5t})$ to avail the quantification of the structural parameters $\boldsymbol{\alpha} = (\alpha_{ij}, i{=}1,2,\ j{=}1,..,4)$, in the form of consistent estimators. It was known at the time that using least-squares to estimate the structural parameters $\boldsymbol{\alpha}$, would yield biased and inconsistent estimators because of the simultaneity (co-determination) problem; y_{1t} causes y_{2t} and vice versa.

Haavelmo (1943, 1944) proposed a way to construct 'good' estimators for $\boldsymbol{\alpha}$ using data $\mathbf{Z}_0$. One way to understand his structural estimation method is to view it in conjunction with the so-called reduced form model, which arises by 'solving' ((6)) for (y_{1t}, y_{2t}) :

$$\begin{aligned} y_{1t} &= \beta_{11} x_{1t} + \beta_{12} x_{2t} + \beta_{13} x_{3t} + \beta_{14} x_{4t} + \beta_{15} x_{5t} + u_{1t} \\ y_{2t} &= \beta_{21} x_{1t} + \beta_{22} x_{2t} + \beta_{23} x_{3t} + \beta_{24} x_{4t} + \beta_{25} x_{5t} + u_{2t} \end{aligned} \tag{8}$$

where the reduced form errors are assumed to have the following probabilistic structure:

$$\begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix} \backsim \mathsf{NIID}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} \right), \ t=1,2,...,n. \tag{9}$$

The 'solving' itself gives rise to an implicit relationship between the structural and statistical parameters: $\boldsymbol{\alpha}$ and $\boldsymbol{\beta} = \left(\beta_{ij}, i{=}1,2,\ j{=}1,..,5\right)$, known as *identification* restrictions: $\mathbf{G}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathbf{0}$.

Since the model (8)-(9) is just a multivariate linear regression model, the least-squares estimators are both unbiased and consistent for $\boldsymbol{\beta}$, and thus solving $\mathbf{G}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathbf{0}$ for $\boldsymbol{\alpha}$ will give rise to consistent estimators. In the above example, the identification restrictions take the form:

$$\beta_{11}{=}\beta_{21}\alpha_{11}{+}\alpha_{12}, \quad \beta_{12}{=}\alpha_{11}\beta_{22}{+}\alpha_{13} \quad \beta_{13}{=}\beta_{23}\alpha_{11}{+}\alpha_{14} \quad \beta_{14}{=}\beta_{24}\alpha_{11}, \quad \beta_{15}{=}\beta_{25}\alpha_{11}$$

$$\beta_{21}{=}\beta_{11}\alpha_{21}{+}\alpha_{22} \quad \beta_{24}{=}\beta_{14}\alpha_{21}{+}\alpha_{23} \quad \beta_{25}{=}\beta_{15}\alpha_{21}{+}\alpha_{24} \quad \beta_{23}{=}\beta_{13}\alpha_{21} \quad \beta_{22}{=}\beta_{12}\alpha_{21} \tag{10}$$

Viewed retrospectively from the error statistical perspective of section 3, the above Simultaneous Equations Model (SEM) had little chance to succeed as a way to learn from the data about economic phenomena for the same reasons (a)-(d) as the Frisch scheme. In particular, when the SEM is viewed from the error statistical perspective of figure 1, the reduced form model is the (implicit) statistical model in the context of which the structural model is embedded; it constitutes a reparameterization/restriction. This suggests that when the former is statistically misspecified, any inference based on it, including the estimation of the structural parameters, is likely to be unreliable. Moreover, the adequacy of the structural model requires, in addition to the statistical adequacy of the reduced form, the appraisal of the *overidentifying* restrictions:

$$H_0\text{: } \mathbf{G}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathbf{0}, \text{ vs. } H_1\text{: } \mathbf{G}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \neq \mathbf{0};$$

the restrictions over and above the ones needed for identifying $\boldsymbol{\alpha}$. In the above example, there are 2 such restrictions because there are 10 statistical parameters in (8)-(9) but only 8 structural parameters in (6)-(7); see Spanos (1986, 1990).

The agenda of the Cowles Commission (see Koopmans, 1950) held the promise to address the inference problems associated with the Simultaneous Equations Model (SEM). However, when their proposed inference tools were applied to economic data in the early 1950s the results were very disappointing in terms of accounting for the regularities in the data and yielding accurate predictions and/or policy evaluations; see Klein (1950), Christ (1951). Partly as a result of these disappointing results, some of the protagonists, including Frisch, Haavelmo and Koopmans, moved away from econometrics into other areas of economics; see Epstein (1987).

6 Theory-data confrontation during the second half of the 20th century

6.1 Textbook Econometrics and it problems

The prevailing view of the theory-data confrontation was dominated by the pre-eminence of theory perspective that led to the conception that econometrics is concerned with the 'quantifi-cation of theoretical relationships', or equivalently "to give empirical content to a priori reasoning in economics" (see Klein, 1962, p. 1). In view of this, econometricians of that period embraced the descriptive statistics perspective in conjunction with the error- tacking strategy to statistical modeling. Hence, the textbook approach to econometrics, as formulated by Johnston (1963) and Goldberger (1964), can be seen as a continuation/modification of the Cowles Commission agenda. The modifications came primarily in the form of (i) less emphasis on simultaneity, and (ii) being less rigid about statistical modeling by allowing non-IID error terms.

The textbook approach to econometrics revolves around the linear regression model:

$$y_k = \beta_0 + \boldsymbol{\beta}_1^{\top}\mathbf{x}_k + u_k,\ \ k = 1, 2, ..., n, \tag{11}$$

$$\text{[i] } E(u_t){=}0,\ \ \text{[ii] } E(u_t^2){=}\sigma^2,\ \ \text{[iii] } E(u_t u_s){=}0,\ \ t \neq s, \qquad t, s{=}1, 2, ..., n. \tag{12}$$

This statistical model is viewed from the pre-eminence of theory perspective as based on a theoretical relationship , with a tacked white noise error term ut representing random causes, such as disturbing factors in the ceteris paribus clause, approximation errors, measurement errors, etc. The underlying statistical theory revolves around the Gauss-Markov theorem, which is totally inadequate for reliable and precise inference; see Spanos (2006a).

Anemic confirmation. Theory appraisal takes the form of anemic confirmation in the following sense. One would estimate (11) by least-squares yielding , and use goodness-of-fit measure like the R^2, together with the sign, magnitude and statistical significance of the coefficients, as the primary criteria for assessing the broad agreement between theory and data. The influence of the F-N-P statistical paradigm comes into this analysis only via the various techniques (estimation, testing and prediction methods) for carrying out the anemic confirmation exercise: instantiating a theory assumed to be true a priori.

When viewed from the error statistical perspective of figure 1, it becomes apparent that anemic confirmation, as a theory appraisal procedure, suffers from the same (a)-(d) weaknesses as the Frisch scheme and the Cowles Commission approach, and as a modeling strategy it hoards some additional flaws. First, the criteria used are wanting unless the estimated model is statistically adequate ([i]-[iii] are valid). Otherwise, the t-statistics do not have the assumed error probabilities and the R^2 is an inconsistent estimator of the underlying goodness-of-fit parameter; see Spanos (2006a). Second, the probabilistic assumptions [i]-[iii] provide an incomplete specification for the linear regression model; see Spanos (2006a). Third, any departures from the error assumptions [i]-[iii] in (12) are not treated as indicating that some systematic statistical information is not accounted for by the original model, where respecification of the model is called for, but as a nuisance to be 'fixed' by using different estimation methods, such as Generalized Least-Squares (GLS), Generalized Method of Moments (GMM) and Instrumental Variables (IV), to replace the original Ordinary Least-Squares (OLS); see Greene (2003).

These textbook error-fixing strategies, however, are questionable on several grounds:

(i) By focusing exclusively on the error term the textbook perspective is geared toward 'saving the theory' and overlooks the ways in which the systematic component may be in error. That is, it ignores alternative theories which might fit the same data equally well or even better. As a result, it (ab)uses the data in ways that 'appear' to provide empirical (inductive) support for the theory in question, when in fact the inferences are unwarranted.

(ii) The 'error-fixing' strategies make extensive (ab)use of the fallacies of acceptance/ rejection; see Mayo and Spanos (2004).

(iii) The error-fixing strategies do not distinguish between different sources of error, mentioned in section 3.2. A well-known example of this is the conflating of statistical with substantive inadequacy; see Spanos (2006b).

In addition to these problems the anemic confirmation procedure is open to considerable abuse in the sense that one could try numerous combinations of variables in xt, as well as several estimation methods, in order to fabricate the 'correct' signs, magnitudes and significance levels for the estimated coefficients, and just report only the last part of the 'fishing' expedition; this practice was branded 'cookbook econometrics' by Leamer (1978). Indeed, the criticisms leveled at the 'cookbook econometrics' made practitioners unduly sensitive to accusations of 'double-use of data' to such an extent that even looking at data plots, or testing the model assumptions, was considered unwarranted; see Kennedy (2003).

Leamer's (1978) suggestion to deal with 'cookbook econometrics' was to formalize these ad hoc procedures using informal Bayesian procedures driven by the modeler's degrees of belief. The presumption was that these strategies need to be formalized in order to account for the data-mining and double-use of data activities. The problem is that when statistical adequacy is recognized as a pre-condition for any reliable searching – to ensure the fidelity of the nominal error probabilities – these 'cookbook' econometrics strategies are guided by unreliable procedures, and formalizing them will achieve nothing of any inferential value; applying a 5% significance level test when the actual type I error is 95% will lead inference results astray; see Spanos and McGuirk (2001). Hence, statistical misspecification and other sources of error are a considerably more worrisome than 'fragility' as understood in the context of Leamer's extreme bounds analysis (see Leamer and Leonard, 1983). The latter results are vacuous when the underlying model is misspecified because one cannot address the misspecification of a statistical model by modifying the priors. Instead, what is needed is a guiding principle which can be used to distinguish between legitimate and illegitimate modeling strategies including, double-use of data, preliminary data analysis, misspecification testing, etc. The severity principle provides a more appropriate guiding rationale; see Mayo (1996), Spanos (2000).

6.2 The discontent with empirical modeling in economics

The initial optimism associated by the promise of the new statistical methods of the Cowles Commission to significantly improve empirical modeling in economics turned into pessimism by the late 1960s. After two decades of laborious efforts to build large theory-based macro-econometric models, and millions of dollars spent by Central Banks and other government institutions, the results were more than a little disappointing. The combination of the Cowles Commission and the newly established textbook approach to econometrics did very little to allay the doubts created in the 1950s that empirical modeling in economics was not an effective tool in learning from data about economic phenomena of interest, nor was it useful for appraising different theories or forecasting and policy decision purposes; see Epstein (1987).

The increasing discontent with empirical analysis in economics reached a crescendo in the early 1970s with leading economists like, Leontief (1971), lambasting both economic theorists and econometricians for the prevailing state of affairs. He was especially disparaging against deliberate attempts to enshroud the lack of substance under a veil of sophisticated mathematical formulations, both in economic theory and econometrics. More specifically, he diagnosed a major imbalance between abstract theorizing and its empirical foundation and blamed the 'indifferent results' in empirical applications primarily on the unreliability of empirical evidence arising from non-testable probabilistic assumptions concerning errors:

> "... the validity of these statistical tools depends itself on the acceptance of certain convenient assumptions pertaining to stochastic properties of the phenomena which the particular models are intended to explain; *assumptions that can be seldom verified*. In no other field of empirical inquiry has so massive and sophisticated a statistical machinery been used with such indifferent results."
> (p. 3) [emphasis added]

Leontief's presumption that probabilistic assumptions about error terms, such as [i]-[iii] in (13), cannot be assessed is clearly false, but the connection between such assumptions and the corresponding assumptions in terms of the observable process was not clear at the time; see Mayo and Spanos (2004).

Similar discontent with the state of scientific knowledge in economics was voiced by other well-known economists, including Boulding (1970), Brown (1972) and Worswick (1972). Leontief's discontent with the 'indifferent results' of econometric modeling was acknowledged more generally, and did not leave econometricians indifferent.

6.3 Data-driven modeling: the beginnings of the 'third way'

The first credible challenge for the Cowles Commission pre-eminence of theory perspective came in the form of a persistently inferior predictive performance (see Cooper, 1972) of its multi-equation structural models when compared with the single equation (theory free) data-oriented formulation known as the Autoregressive-Integrated-

Moving Average [ARIMA(p,d,q)] model:

$$y_t = \alpha_0 + \sum_{k=1}^{p} \alpha_k y_{t-k} + \sum_{\ell=1}^{q} \beta_\ell \varepsilon_{t-\ell} + \varepsilon_t, \quad \varepsilon_t \backsim \mathsf{NIID}(0, \sigma^2), \ t \in \mathbb{N}, \qquad (13)$$

popularized by Box and Jenkins (B-J) (1970). The ARIMA model in (14) aims to capture the temporal correlation of a single data series y_t by regression on its own past $(y_{t-1}, y_{t-2}, \ldots, y_{t-p})$ (autoregression), and past errors $(\varepsilon_t, \varepsilon_{t-1}, \ldots, \varepsilon_{t-q})$ (moving average), where any heterogeneity in the original data (y_t^*) is eliminated by 'differencing' $y_t := \Delta^d y_t^*$. Despite their obvious weakness that ARIMA models ignore all substantive information stemming from economic theory, they persistently outperformed the multi-equation (macroeconometric) structural models on prediction grounds. The question that naturally arose at the time was:

"why did these large macroeconometric models performed so poorly
on prediction grounds?"

The answer was rather simple and grew out of the 'third way' mentioned in the introduction, but has not been fully appreciated in econometrics to this day: empirical models which do not account for the statistical regularities in the data are likely to give rise to untrustworthy empirical evidence and very poor predictive performance; see Granger and Newbold (1986).

The primary problem with the Cowles Commission structural macroeconometric models was that economic theory gave rise to static multi-equation equilibrium models, which were totally ignoring the temporal (over time) statistical information in the data.

The success of ARIMA modeling on predictive grounds encouraged several econometricians to challenge the then dominating pre-eminence of theory over data perspective and call for greater reliance on the statistical regularities in the data using data-driven models and less reliance on substantive subject matter information.

Sims (1980) argued that substantive information in macroeconomics is often "incredible", and "cannot be taken seriously" (pp. 1-2). Indeed, the only such information needed for empirical modeling is some low level theory as to which variables might be involved in explaining a certain phenomenon of interest, say $\mathbf{Z}_t := (Z_{1t}, \ldots, Z_{mt})$, and the modeling should focus on the statistical information. He proposed the Vector Auto-Regressive [VAR(p)] model:

$$\mathbf{Z}_t = \mathbf{a}_0 + \sum_{k=1}^{p} \mathbf{A}_k \mathbf{Z}_{t-k} + \mathbf{E}_t, \quad \mathbf{E}_t \backsim \mathsf{NIID}(\mathbf{0}, \mathbf{\Omega}), \ t \in \mathbb{N}, \qquad (14)$$

which captures the joint temporal dependence and heterogeneity contained in the data. Note that the VAR(p) model is simply a multi-equation extension of the AR(p) model, a variation on the ARIMA formulation, relying on the Normality, Markov and stationarity assumptions.

The LSE tradition (Sargan, 1964, Hendry, 2000) also called for more reliance on statistical regularities in the data, proposing the Autoregressive Distributed Lag [AD(p,q)] model:

$$y_t = \alpha_0 + \sum_{k=1}^{p} \alpha_k y_{t-k} + \sum_{\ell=1}^{q} \boldsymbol{\beta}_\ell^\top \mathbf{x}_{t-\ell} + \varepsilon_k, \quad \mathbf{x}_t : k \times 1, \ \varepsilon_k \backsim \mathsf{NIID}(0, \sigma^2), \ t \in \mathbb{N}, \qquad (15)$$

which is based on the same probabilistic assumptions as (14), but retained an indirect link to the theory via the long-run equilibrium solution of (15); see Hendry (1995).

Viewed in the context of the error statistical perspective, the formulations in (13)-(15) constitute proper statistical models in the error statistical context, being parameterizations of generic stochastic processes, assumed to be Normal, Markov and stationary, whose statistical adequacy vis-à-vis data $\mathbf{y}_0:=(y_1, \ldots, y_n)$ could, in principle, be assessed using Mis-Specification (M-S) testing. Historically, the Box and Jenkins (1970) 'diagnostic checks' provided the initial impetus for M-S testing; this was then strongly encouraged by the LSE tradition.

What is particularly interesting about the Box-Jenkins, Sims and LSE approaches is that all three: (a) share a data-oriented objective to allow the data 'a voice of its own', stemming from 'accounting for the statistical regularities in the data', (b) rely on statistical models in the context of a frequentist statistical framework, and (b) emphasize the use of error probabilities in statistical induction; all three being crucial features of the error-statistical perspective discussed in section 2.

In addition, these approaches put forward a number of right-headed innovations which are designed to enhance the reliability of the 'voice' of the data in empirical modeling. In particular, the Box-Jenkins approach views statistical modeling as an iterative process that involves several stages, identification, estimation, diagnostic checking, and prediction. In addition, graphical techniques and exploratory data analysis were rendered legitimate tools in choosing a statistical model and ensuring its adequacy by assessing whether it accounts for the regularities in the data.

The LSE approach rejected the pre-eminence of theory perspective, and proposed a number of procedures for securing data-oriented models which are congruent with data before relating them to existing theories. A model congruent with data is reached using sound frequentist statistical procedures such as the 'general-to-specific':

> "Model selection starts from the most general feasible model and arrives at a parsimonious model by imposing acceptable restrictions." (Campos et al, 2005, p. xxxvii)

The primary motivation for following this procedure stems from maintaining the reliability of the ensuing inferences by ensuring the 'optimality' of the testing procedure as well as keeping track of the relevant error probabilities. What distinguishes the LSE approach from other more data-oriented traditions was its persistent emphasis on justifying the methodological foundations of its procedures using the scientific credentials of frequentist inference, and its intention to ultimately relate the estimated models congruent with data to economic theory; see Hendry (2000).

6.4 The pre-eminence of theory perspective reaffirming itself

The pre-eminence of theory tradition is so deeply entrenched in economics that it was able to brush aside the 'crises' of the 1970s and 1980s by mounting a counter-attack on the Box-Jenkins, Sims and LSE approaches, scorning them as indulging in 'data mining', and reaffirm itself by explaining away the predictive failure of the theory-dominated structural models as primarily due to the inadequacies of the prevailing theories; an anomaly that can be accounted for by better theories, and not by paying due attention to the regularities in the data.

Lucas (1976) and Lucas and Sargent (1981) countered the econometricians call for more reliance on statistical information, by arguing, instead, for enhanced new Classical Macro-economic theories based on: (a) modeling expectations and rendering structural models dynamic to account for the temporal structure of economic time series, and (b) constructing structural models which are invariant to policy interventions for prediction and policy evaluations, so that their predictive performance improves.

Viewed from the error statistical perspective as an attempt to bridge the theory-data gap, the Lucas-Sargent call for improved dynamic structural macro-models is a move in the right direction, but if the enhancement of the theory is at the expense of ignoring the statistical information, it's an empty gesture because it simply re-affirms the pre-eminence of theory over data. One has no way of distinguishing between 'better' and 'worse' theories vis-à-vis the data, i.e. theories which account for the regularities in the data and those that don't.

More regretful was the call from Kydland and Prescott (1991) to, just about, ignore the data information almost entirely. Instead of using data to estimate the parameters of structural models applying proper statistical procedures, one should use data to 'calibrate' them using 'informal' procedures. That is, use data in conjunction with 'ad hoc' methods to 'assign' numerical values to the unknown structural parameters in such a way so as to ensure that, when simulated, these models yield artificial data that tend to 'mimic' the behavior of the actual data in very broad terms. Given that 'calibration' purposefully forsakes error probabilities and provides no way to assess the reliability of inference, how does one assess the adequacy of the calibrated model?

> "The issue of how confident we are in the econometric answer is a subtle one which cannot be resolved by computing some measure of how well the model economy mimics historical data. The degree of confidence in the answer depends on the confidence that is placed in the economic theory being used."
> (see Kydland and Prescott, 1991, p. 171)

Indeed, they proceed to re-affirm the pre-eminence of theory by casting aside the data altogether:

> "The model economy which better fits the data is not the one used.
> Rather currently established theory dictates which one is used."
> (see Kydland and Prescott, 1991, p. 174)

Sargent (1976), p. 233, also argues for a highly tenuous accord between theory and data:

> "All in all, the empirical results provide some evidence that the causal structure imposed on the data by the classical model of Section I *is not obscenely at*

variance with the data." [emphasis added]

This is clearly a highly unsatisfactory way to appraise the empirical adequacy of a structural model and makes a mockery of the theory-data confrontation; see Hoover (2001b, 2006) for a thorough discussion of new classical macroeconomics and their disregard for data information.

6.5 The prevailing view concerning the theory vs. data debates

The data-oriented approaches, including the 'third way', because of their emphasis on 'accounting for the regularities' in the data, have been under an incessant attack from the prevailing pre-eminence of theory tradition, being charged with a litany of unscientific practices with catchy names like: 'measurement without theory', 'data mining', 'pre-test bias', 'ignoring selection effects', 'repeated testing', 'hunting for statistical significance', 'lack of proper identification', 'double-use of data', etc., etc.; see Hendry (2000), p. 469. Indeed, the mainstream critics act as if their own philosophical foundations are firmly secure because of the ostensible mathematical 'rigor' of their models and the longevity of their tradition going back to Ricardo's clarity of logical deduction, and Mill's subordination of data to theory. Estimating, a structural model presumed 'true', using some seemingly sophisticated method (e.g. GMM), and quoting some fit statistics and a few 'significant' coefficients confirming the theory, meets their primary objective; never-mind the endemic statistical misspecifications and other ways the inductive inferences might be in error, or the inability of these structural models to account for the regularities in the data, their pervasive predictive failures, and their unsoundness as tools for policy formulation.

Unfortunately for economics, the end result of the methodological discussions on the theory-data confrontation of the last 30 years or so is that the practitioners of the 'third way' found themsleves constantly on the defensive. Its proponents are being required to justify their every procedure which does not conform to the prevailing pre-eminence of theory practices. Frustrated by the lack of a coherent philosophical foundation that could be used to defend their methods and procedures from these incessant attacks, 'third way' proponents like Hendry find themselves acting as reluctant philosophers without much help from philosophy of science itself. Indeed, the latter's prevailing perspectives seem to suggest that those who have not given up on 'objective' methods are somehow deceiving themselves (, drowning out the error statistical discourse that by controlling the reliability of inference through frequentist error probabilities one can address inveterate classic philosophical problems, including 'how we learn from data'.

7 Conclusions

The main thesis of this paper is that the error statistical perspective can provide a coherent framework wherein the 'third way' can foster its twin goals of developing an adequate methodology and a corresponding philosophical foundations. It can help its practitioners attain the goal of objectively criticizing theories by giving data 'a voice of its own', furnishing a natural home and a common language for their methods and procedures.

The error statistical perspective provides a methodological framework that can be used, not only to defend frequentist procedures against unwarranted charges like many of the ones mentioned above, but also to mount a counter-attack exposing the major weaknesses in the foundations of the pre-eminence of theory perspective and their unscientific practices. The currently practiced instantiation (anemic confirmation) of theories, which "is not obscenely at variance with the data", has obviated any form of constructive dialogue between theory and data, and thus undermined the scientific credentials of economics as an empirical science. Instead, what is needed is a methodology of error inquiry that encourages detecting and identifying the different ways an inductive inference could be in error by applying effective procedures which would detect such errors when present with very high probability; Mayo (1996).

An important dimension of the error statistical framework is its broader technical modeling perspective (section 3) where the Normal/linear/homoskedastic models relating to the ARIMA, VAR and ADL formulations can be systematically extended to greatly enrich the family of potential statistical models which can be specified by changing the probabilistic assumptions of Normality, Markov and stationarity; see Spanos (1989, 2006a, 2008).

Most importantly, the error statistical perspective can provide an trenchant framework for the theory-data confrontation, wherein the sequence of models (theory, structural, statistical and empirical), discussed in sections 2-3, can be used to bring out as well as deal adequately with the three interrelated modeling stages: (A) from theory to testable hypotheses, (B) from raw data to reliable evidence, and (C) confronting hypotheses with evidence. The basic idea is to use objective frequentist procedures to localize errors in different models and apply severe testing in a piece-meal way at a level which enables one to probe for errors one at a time in an exhaustive fashion, and then pieced together to provide an overall testing procedure. As argued by Mayo (1996), the sequence of interconnected models performs double duty: (a) it provides a way to foreground (and thus help to bridge) the gap between theory and data, and (b) organizes the key tasks of 'learning from data' about observable phenomena of interest (pp. 129-130). The notion of 'statistical knowledge' is crucial for theory testing because by bestowing on data 'a voice of its own', separate from the one ideated by the theory in question, it succeeds in securing the reliability and scientific objectivity of the theory-data confrontation.

Moreover, statistical knowledge (like experimental knowledge) is independent of high-level theories, and in fields like economics relying on observational data, the accumulation of statistically adequate models is of paramount importance because it can help to guide the challenging search for substantive explanations. It can provide a guiding constrain for new theories aspiring to explain the particular phenomenon of interest in so far as they need to account for this statistical knowledge; see Spanos (2006b).

References

[1] Abadir, K. and G. Talmain, (2002), "Aggregation, Persistence and Volatility in a Macro Model," *Review of Economic Studies*, 69: 749-779.

[2] Backhouse, R. E. (ed) (1994), *New Directions in Economic Methodology*, Routeledge, London.

[3] Backhouse, R. E. (2002), *The Penguin History of Economics*, Penguin Books, London.

[4] Blaug, M. (1958), *Ricardian Economics: a Historical Study*, Yale University Press, CT.

[5] Boulding, K. (1970), *Economics as a Science*, McGraw-Hill, NY.

[6] Bowley, A. L. (1926), *Elements of Statistics*, 5th ed. London: Staples Press.

[7] Bowley, A. L. (1924), *The Mathematical Groundwork of Economics*, Reprints of Economic Classics, 1960, Augustus, M. Kelley, NY.

[8] Box, G. E. P. and Jenkins, G. M. (1970), *Time Series Analysis: Forecasting and Control*, San Francisco: Holden-Day.

[9] Brown, E. H. P. (1972), "The Underdevelopment of Economics," *Economic Journal*, 82: 1-10.

[10] Cairnes, J. E. (1888), *The Character and Logical Method of Political Economy*, Reprints of Economic Classics, 1965, Augustus, M. Kelley, NY.

[11] Caldwell, B. (1984), Appraisal and Criticism in Economics: a book of readings, Allen & Unwin, London.

[12] Campos, J., N. R. Ericsson, and D. F. Hendry (2005) (eds.), *General-to-Specific Modeling*, Vol. 1-2, Edward Elgar, Northampton, MA.

[13] Christ, C. F. (1951), "A Test of an Econometric Model for the United States 1921-1947," in NBER Conference on *Business Cycles*, NBER, NY.

[14] Cooper, R. L. (1972), "The predictive performance of quarterly econometric models of the United States," in *Econometric Models of Cyclical Behavior*, ed. by Hickman, B. G., Columbia University Press, NY.

[15] Cox, D. R. and D. V. Hinkley (1974), *Theoretical Statistics*, Chapman & Hall, London.

[16] Desrosières, A. (1998), *The Politics of Large Numbers: A History of Statistical Reasoning*, Harvard University Press, Cambridge, MA.

[17] Doob, J. L. (1953), *Stochastic Processes*, Wiley, NY.

[18] Ekelund, R. B. and R. F. (1975), *A History of Economic Theory and Method*, McGraw-Hill, NY.

[19] Epstein, R. J. (1987), A History of Econometrics, North Holland, Amsterdam.

[20] Fisher, R. A. (1921), "On the 'probable error' of a coefficient deduced from a small sample," *Metron*, 1: 2–32.

[21] Fisher, R. A. (1922), "On the mathematical foundations of theoretical statistics," *Philosophical Transactions of the Royal Society*, A 222: 309–68.

[22] Fisher, R. A. (1925), *Statistical Methods for Research Workers*, Oliver & Boyd, Edinburgh.

[23] Fisher, R. A. (1935), The Design of Experiments, Oliver & Boyd, Edinburgh.

[24] Friedman, M. (1953), "The Methodology of Positive Economics," pp. 3-43 in *Essays in Positive Economics*, Chicago University Press, Chicago.

[25] Frisch, R. (1933), "Editorial," *Econometrica*, 1: 1–4.

[26] Frisch, R. (1934), *Statistical Confluence Analysis by Means of Complete Regression Schemes*, Universitetets Okonomiske Institutt, Oslo.

[27] Goldberger, A.S. (1964), *Econometric Theory*, Wiley, NY.

[28] Granger, C. W. J. (1990), (ed.) *Modelling Economic Series*, Clarendon Press, Oxford.

[29] Granger, C. W. J. and P. Newbold (1986), *Forecasting Economic Time Series*, 2nd, ed., Academic Press, London.

[30] Greene, W. H. (2003), *Econometric Analysis*, 5th ed., Prentice Hall, NJ.

[31] Guala, F. (2005), *The Methodology of Experimental Economics*, Cambridge University Press, Cambridge.

[32] Haavelmo (1943), "The Statistical Implications of a System of Simultaneous Equations," *Econometrica*, 11, 1-12.

[33] Haavelmo, T. (1944), "The probability approach to econometrics," *Econometrica*, 12 (Suppl), 1–115.

[34] Hendry, D. F. (1980), "Econometrics – Alchemy or Science?", *Economica*, 47: 387-406; reprinted in Hendry (2000).

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

[36] Hendry, D. F. (2000), *Econometrics: Alchemy or Science?* Blackwell, Oxford.

[37] Hendry, D. F., E. E. Leamer and D. J. Poirier (1990), "The ET dialogue: a conversation on econometric methodology," *Econometric Theory*, 6: 171-261.

[38] Hendry, D. F., M. S. Morgan (eds.) (1995), *The Foundations of Econometric Analysis*, Cambridge University Press, Cambridge.

[39] Hood, W. C. and Koopmans, T. C. (1953), (eds.), *Studies in Econometric Method*, Cowles Commission Monograph, No. 14, Wiley, NY.

[40] Hoover, K. D. (2001a), Causality in Macroeconomics, Cambridge University Press, Cambridge.

[41] Hoover, K. D. (2001b), The Methodology of Empirical Macroeconomics, Cambridge University Press, Cambridge.

[42] Hoover, K. D. (2006), "The Methodology of Econometrics," pp. 61-87, in New Palgrave Handbook of Econometrics, vol. 1, ed. T. C. Mills and K. Patterson, Macmillan, London.

[43] Hutchison, T. W. (1938), The Significance and Basic Postulates of Economic Theory, Reprints of Economic Classics, 1965, Augustus, M. Kelley, NY.

[44] Jevons, W. S. (1871/1911), The Theory of Political Economy, 4th ed. 1911, Reprints of Economic Classics, 1965, Augustus, M. Kelley, NY.

[45] Johnston, J. (1963), Econometric Methods, McGraw-Hill, NY.

[46] Kalman, R. E. (1982), "System Identification from Noisy Data," in Dynamic Systems II, eds. A. R. Bednarek and L. Cesari, Academic Press, London.

[47] Kennedy, P. (2003), A Guide to Econometrics, 5th edition, MIT Press, Cambridge.

[48] Keynes, J. N. (1890/1917), The Scope and Method of Political Economy, 2nd ed. 1917, Reprints of Economic Classics, 1986, Augustus, M. Kelley, NJ.

[49] Klein, L. R. (1950), Economic Fluctuations in the United States 1921–1941. Cowles Commission for Research in Economics, Monograph No. 11, Wiley, NY.

[50] Klein, L. R. (1962), An Introduction to Econometrics, Prentice-Hall, NJ.

[51] Koopmans, T. C. (1950), Statistical Inference in Dynamic Economic Models, Cowles Commission Monograph, No. 10, Wiley, NY.

[52] Kydland, F. and P. Prescott (1991), "The Econometrics of the General Equilibrium Approach to Business Cycles," Scandinavian Journal of Economics, 93: 161-178.

[53] Leamer, E. E. (1978), Specification Searches: Ad Hoc Inference with Nonexperimental Data, Wiley, NY.

[54] Leamer, E. E. and H. B. Leonard (1983), "Reporting the fragility of regression estimates," Review of Economics and Statistics, 65, 306-317.

[55] Lehmann, E. L. (1986), Testing statistical hypotheses, 2nd edition, Wiley, New York.

[56] Leontief, W. W. (1971), "Theoretical Assumptions and Nonobserved Facts," American Economic Review, 1-7.

[57] Lucas, R. E. (1976), "Econometric Policy Evaluation: a Critique," pp. 19-46 of The Phillips Curve and Labour Markets, ed. by K. Brunner and A. M. Metzer, Carnegie-Rochester Conference on Public Policy, I, North-Holland, Amsterdam.

[58] Lucas, R. E. and T. J. Sargent (1981), Rational Expectations and Econometric Practice, George Allen & Unwin, London.

[59] Machlup, F. (1963), Essays in Economic Semantic, Prentice-Hall, NJ.

[60] Malthus, T. R. (1936), Principles of Political Economy, Reprints of Economic Classics, 1986, Augustus, M. Kelley, NJ.

[61] Marshall, A. (1891/1920), Principles of Economics, 8th ed., 1920, McMillan, London.

[62] Mayo, D. G. (1991), "Novel evidence and severe tests," Philosophy of Science, 58: 523–52.

[63] Mayo, D. G. (1996), Error and the Growth of Experimental Knowledge, The University of Chicago Press, Chicago.

[64] Mayo, D. G. (2005), "Philosophy of statistics," In Philosophy of Science: An Encyclopedia, S. ed. S. Sarkar and J. Pfeifer, Routledge, London.

[65] Mayo, D. G. and D. R. Cox (2006), "Frequentist statistics as a theory of inductive inference," pp. 96-123 in The Second Erich L. Lehmann Symposium: Optimality, Lecture Notes-Monograph Series, Volume 49, Institute of Mathematical Statistics.

[66] Mayo, D. G. and Spanos, A. (2004), "Methodology in Practice: Statistical Misspecification Testing," Philosophy of Science, 71: 1007–25.

[67] Mayo, D. G. and Spanos, A. (2006), "Severe testing as a basic concept in a Neyman–Pearson philosophy of induction," British Journal for the Philosophy of Science, 57, 323–57.

[68] Mayo, D. G. and Spanos, A. (2008), "Error Statistics," forthcoming in Philosophy of Statistics, Handbook of Philosophy of Science, Elsevier, (editors) D. Gabbay, P. Thagard, and J. Woods.

[69] McCulloch, J. R. (1864), The Principles of Political Economy, Reprints of Economic Classics, 1965, Augustus, M. Kelley, NY.

[70] Mill, J. S. (1844), Essays on some Unsettled Questions of Political Economy, Reprints of Economic Classics, 1974, Augustus, M. Kelley, NJ.

[71] Mill, J. S. (1871), Principles of Political Economy, Reprints of Economic Classics, 1976, Augustus, M. Kelley, NJ.

[72] Mill, J. S. (1884), A System of Logic, Ratiocinative and Inductive, 8th ed., Harper & Brothers, NY.

[73] Mills, F. C. (1924/1938), Statistical Methods, Henry Holt, NY.

[74] Mirowski, P. (1994), "What are the questions?" pp. 50- 74 in Backhouse (1994).

[75] Morgan, M. S. (1990), The History of Econometric Ideas, Cambridge University Press, Cambridge.

[76] Morgenstern, O. (1950/1963), On the accuracy of economic observations, 2nd edition, Princeton University Press, New Jersey.

[77] Musgrave, A. (1981), "Unreal Assumptions in Economic Theory: The F-Twist Untwisted," Kyklos, 34: 377-387.

[78] Nagel, E. (1963), "Assumptions in Economic Theory," American Economic Review, 53, 211-219.

[79] Neyman, J. (1950), First Course in Probability and Statistics, Henry Holt, NY.

[80] Neyman, J. (1977), "Frequentist probability and frequentist statistics," Synthèse, 36: 97–131.

[81] Neyman, J. and Pearson, E. S. (1928), "On the use and interpretation of certain test criteria for the purposes of statistical inference," Parts I and II, Biometrika, 20A, 175-240, 263-294.

[82] Neyman, J. and Pearson, E. S. (1933), "On the problem of the most efficient tests of statistical hypotheses," Philosophical Transactions of the Royal Society of London, Series A, 231: 289–337.

[83] Pearson, E. S. (1966), "The Neyman–Pearson story: 1926–34," In Research Papers in Statistics: Festschrift for J. Neyman, ed. F. N. David, Wiley, NY.

[84] Porter, G. R. (1836/1912), The Progress of a Nation, 2nd ed., 1912, Reprints of Economic Classics, 1970, Augustus, M. Kelley, NY.

[85] Quetelet, A. (1942), A Treatise on Man and the Development of his Faculties, Chambers, Edinburgh.

[86] Ricardo, D. (1817), Principles of Political Economy and Taxation, vol. 1 of The Collected Works of Davie Ricardo, ed. P. Sraffa and M. Dobb, Cambridge University Press, Cambridge.

[87] Robbins, L. (1932/1935), *An Essay on the Nature and Significance of Economic Science*, 2nd ed., 1935, McMillan, London.

[88] Sargan, J. D. (1964) "Wages and Prices in the United Kingdon: A study in econometric methodology," in P.E. Hart, G. Mills and J.K. Whitaker, eds., *Econometric Analysis for Economic Planning*, Butterworths, London.

[89] Sargent, T. (1976), "A Classical Macroeconomic Model for the United States," *Journal of Political Economy*, 84: 207-238.

[90] Senior, N. W. (1836), *An Outline of the Science of Political Economy*, Reprints of Economic Classics, 1965, Augustus, M. Kelley, NY.

[91] Sims, C. A. (1980), "Macroeconomics and Reality," *Econometrica*, 48: 1-48.

[92] Smith, A. (1776), *An Inquiry into the Nature and Causes of the Wealth of Nations*, ed. R. H. Campell, A. S. Skinner and W. B. Todd, Clarendon Press, Oxford.

[93] Smith, V. L. (2007), *Rationality in Economics: Constructivist and Ecological Forms*, Cambridge University Press, Cambridge.

[94] Spanos, A. (1986), *Statistical Foundations of Econometric Modelling*," Cambridge University Press, Cambridge.

[95] Spanos, A. (1989), "On re-reading Haavelmo: a Retrospective view of Econometric Modeling," Econometric Theory, 5: 405–429.

[96] Spanos, A. (1990), "The Simultaneous Equations Model Revisited: Statistical Adequacy and Identification," *Journal of Econometrics*, 44: 87–108.

[97] Spanos, A. (1995), "On theory testing in econometrics: modeling with nonexperimental data," *Journal of Econometrics*, 67: 189–226.

[98] Spanos, A. (1999), *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*," Cambridge University Press, Cambridge.

[99] Spanos, A. (2000), "Revisiting Data Mining: 'hunting' with or without a license," T**he Journal of Economic Methodology**, 7: 231-264.

[100] Spanos, A. (2006a), "Econometrics in retrospect and prospect," In *New Palgrave Handbook of Econometrics*, vol. 1, ed. T. C. Mills and K. Patterson, Macmillan, London.

[101] Spanos, A. (2006b), "Revisiting the omitted variables argument: substantive vs. statistical adequacy," *Journal of Economic Methodology*, 13: 179–218.

[102] Spanos, A. (2006c), "Where do statistical models come from? Revisiting the problem of specification," in *Optimality: The Second Erich L. Lehmann Symposium*, ed. J. Rojo, Lecture Notes-Monograph Series, vol. 49, Institute of Mathematical Statistics, Beachwood, OH.

[103] Spanos, A. (2008), "Philosophy of Econometrics," forthcoming in *Philosophy of Economics*, the *Handbook of Philosophy of Science*, Elsevier (editors) D. Gabbay, P. Thagard, and J. Woods.

[104] Spanos, A. and McGuirk, A. (2001), "The model specification problem from a probabilistic reduction perspective," *Journal of the American Agricultural Association*, 83: 1168–76.

[105] Sugden, R. (2005), "Experiments as exhibits and experiments as tests," *Journal of Economic Methodology*, 12: 291-302.

[106] Sugden, R. (2008), "The changing relationship between theory and experiment in economics," forthcoming *Philosophy of Science*, PSA 06.

[107] Worswick, G. D. N. (1972), "Is Progress in Economic Science Possible?" *Economic Journal*, 82: 73-86.