The Curve-Fitting Problem, Akaike-type Model Selection, and the Error Statistical Approach

Aris Spanos[*]

Department of Economics,
Virginia Tech,
<aris@vt.edu>

October 17, 2006

Abstract

The curve-fitting problem is often viewed as an exemplar which encapsulates the multitude of dimensions and issues associated with *inductive inference*, including *underdetermination* and the *reliability* of inference. The prevailing view is that the 'fittest' curve is one which provides the optimal trade-off between goodness-of-fit and simplicity, with the Akaike Information Criterion (AIC) the preferred method. The paper argues that the AIC-type procedures do not provide an adequate solution to the curve fitting problem because (a) they have no criterion to assess when a curve *captures the regularities in the data* inadequately, and (b) they are prone to unreliable inferences. The thesis advocated is that for more satisfactory answers one needs to view the curve-fitting problem in the context of error-statistical approach where (i) *statistical adequacy* provides a criterion for selecting the fittest curve and (ii) the error probabilities can be used to calibrate the reliability of inductive inference. This thesis is illustrated by comparing the Kepler and Ptolemaic models in terms of statistical adequacy, showing that the latter does not 'save the phenomena' as often claimed. This calls into question the view concerning the pervasiveness of the problem of underdetermination; statistically adequate 'fittest' curves are rare, not common.

1

1 Introduction

The curve-fitting problem has a long history in both statistics and philosophy of science, and it's often viewed as an exemplar which encapsulates the many dimensions and issues associated with *inductive inference*, including the *underdetermination* conundrum and the *reliability of inference* problem. Glymour (1981) highlighted the importance of curve-fitting and argued that the problem was not well understood either in its philosophical or its mathematical dimensions. A decade later, however, the prevailing view in philosophy of science is that the use of model selection procedures associated with the Akaike Information Criterion (AIC) could address the problem in a satisfactory way by trading goodness-of-fit against simplicity (see Forster and Sober (1994), Kukla (1995), Kieseppa (1997), Mulaik (2001), inter alia). This is despite the fact that Glymour (1981) argued persuasively that *simplicity cannot* provide an adequate solution.

The question posed in this paper is the extent to which the AIC and related model selection procedures provide a satisfactory solution to the curve-fitting problem. The main thesis is that these procedures are inadequate for the task because they do not provide a satisfactory way to assess the circumstances that a particular curve can be attested to *capture the 'regularities' in the data* adequately or not. As a result, they have no way to ensure the reliability of inductive inferences associated with the 'fittest' curve, and give rise to misleading impressions as to the pervasiveness of the underdetermination problem. It is argued that for more satisfactory answers one needs to view the curve-fitting problem in the context of **error-statistical approach** (see Mayo, 1996), where (i) *statistical adequacy* provides the missing criterion of when a curve *captures the 'regularities' in the data* adequately, and (ii) the associated *error probabilities* can be used to calibrate the reliability of inductive inference.

In section 2 the curve-fitting problem is summarized as a prelude to section 3 which brings out the mathematical approximation perspective dominating the current understanding of the problem, and provides the motivation for the AIC procedures. It is argued that this perspective is inadequate if reliable inductive inference is the primary objective. Using the early history of curve-fitting as the backdrop, it is argued that Gauss (1809) provided the first attempt to place the curve-fitting problem in a broader modeling framework. The inadequacy of the mathematical approximation perspective, as providing a foundation for inductive inference, is brought out by comparing Legendre's (1805) use of least-squares as a curve-fitting method with Gauss's empirical modeling perspective. Section 4 summarizes the basic tenets of the error-statistical approach as a prelude to section 5 where this perspective is used to bring out the unsatisfactoriness of the prevailing view in philosophy of science concerning curve-fitting. The error-statistical perspective is illustrated by comparing the Kepler and Ptolemaic models on statistical adequacy grounds, demonstrating that, despite the excellent fit of both models, statistical adequacy declares the former a clear winner. Indeed, the statistical inadequacy of the Ptolemaic model is symptomatic of

2

the curve-fitting perspective as a mathematical approximation procedure. This calls into question the widely held view concerning the pervasiveness of the problem of underdetermination; statistically adequate curves are rare, not common. In section 6 the AIC type procedures are reconsidered in view of the error-statistical perspective. It is argued that trading goodness-of-fit against simplicity provides no sound basis for addressing the curve-fitting problem because these procedures ignore both statistical adequacy as well as the reliability of inference issue.

2   Summarizing the curve-fitting problem

The curve-fitting problem assumes that there exists a *true* relationship between two variables, say $y = h(x)$, and curve-fitting amounts to finding a curve, say $y = g_m(x)$, that fits the existing *data:*

$$\mathbf{z}_0 := \{(x_k, y_k),\ k = 0, 1, \ldots, n\}, \tag{1}$$

'best', and approximates $h(x)$ well enough to ensure its predictive accuracy beyond the data in hand. An important aspect of inductive inference exemplified by the curve-fitting problem is that of *underdetermination*, considered to be pervasive, because it is often claimed that more than one fitted curve *captures the regularities in the data equally well.*

The problem of curve-fitting, as currently understood, is thought to comprise two stages.

Stage 1. The choice of a family of curves, say:

$$g_m(x; \boldsymbol{\alpha}) = \sum_{i=0}^{m} \alpha_i \phi_i(x), \tag{2}$$

where $\boldsymbol{\alpha} := (\alpha_1, \alpha_2, ..., \alpha_m)$, $\{\phi_i(x),\ i = 1, 2, ...m\}$ are known functions, e.g. ordinary polynomials:

$$\phi_0(x) = 1,\, \phi_1(x) = x,\, \phi_2(x) = x^2,\, ...,\, \phi_m(x) = x^m,$$

or even better, *orthogonal polynomials*; see Hildebrand (1974).

Stage 2. The selection of the 'best' fitting curve (within the chosen family) using a certain goodness-of-fit criterion. Least squares is the preferred method for choosing $g_m(x_k; \boldsymbol{\alpha})$ that minimizes the sum of squares of the errors $\varepsilon(x_k, m) = y_k - g_m(x_k; \boldsymbol{\alpha})$:

$$\ell(\boldsymbol{\alpha}) = \sum_{k=1}^{n} \left(y_k - \sum_{i=0}^{m} \alpha_i \phi_i(x_k) \right)^2. \tag{3}$$

That is, $\ell(\boldsymbol{\alpha})$ is minimized over $\boldsymbol{\alpha} \in \mathbb{R}_A^m$:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^m} \ell(\boldsymbol{\alpha}) = \sum_{k=1}^{n} \left(y_k - g_m(x_k; \widehat{\boldsymbol{\alpha}}_{LS}) \right)^2,$$

giving rise to $\widehat{\boldsymbol{\alpha}}_{LS}$, the least squares estimator of $\boldsymbol{\alpha}$; see Hildebrand (1974).

The basic problem with this perspective is that goodness-of-fit cannot be the sole criterion for 'best' because $\ell(\boldsymbol{\alpha})$ can be made arbitrarily small by choosing a large

enough m, giving rise to overfitting. Indeed, one can render the errors of approximation zero by choosing $m = n-1$; see Hildebrand (1974).

Glymour (1981) reviewed several attempts to justify curve-fitting by using other criteria in addition to goodness-of-fit, and concluded:

> "The only moral I propose to draw is that there is no satisfactory rationale for curve-fitting available to use yet." (ibid., p. 340)

Among the various attempts that Glymour considered inadequate were Bayesian procedures based on choosing the curve that is most probable given the evidence, as well as attempts to supplement goodness-of-fit with pragmatic criteria such as simplicity:

> "Attempts to explain curve-fitting without resort to probability have focused on simplicity. But it explains very little to claim merely that the fewer the parameters the simpler the parametric family, and, further that we prefer the simplest family consistent with the data." (ibid., p. 324-5)

Despite that denunciation, discussions in philosophy of science in the mid 1990s seem to suggest that using the *Akaike Information Criterion* (AIC) for model selection, which trades overfitting against simplicity (parsimony), does, after all, provide a satisfactory solution to the curve-fitting problem; see Forster and Sober (1994), Kukla (1995), Kieseppa (1997), Mulaik (2001), inter alia. More recently, Hitchcock and Sober (2004) have used the AIC to shed light on the distinction between *prediction* and *accommodation* as they relate to overfitting and novelty. The main argument for AIC-type model selection can be summarized by the following claims:

> (I) the selection of the best fitting curve in (3) is well understood in the sense that least-squares provides *the* standard solution, and

> (II) simplicity can be justified on prediction grounds because simpler curves enjoy better predictive accuracy.

Both claims are called into question in the discussion that follows.

3   Curve-fitting or statistical modeling?

The main thesis of the paper is that AIC-type procedures do not provide a satisfactory solution to the curve-fitting problem primarily because they do not address the crucial *adequacy* issue of 'when a fitted curve captures the regularities in the data'. Rather, it is implicitly assumed that curves with the same high goodness-of-fit capture the regularities in the data equally well, but simpler curves have a clear advantage on prediction grounds. It is argued that a curve is 'fittest' not because of the magnitude of its residuals but their non-systematic (in a probabilistic sense) nature; fitted curves with excellent fit and predictive accuracy can be shown to be statistically inadequate - they do not account for the regularities in the data.

Viewing the curve-fitting problem in terms of the two stages described above, (a) selecting the family of curves, and (b) picking out the fittest curve within this family, is largely the result of imposing a *mathematical approximation* perspective on the

problem. As argued below, however, this mathematical framework provides an inadequate basis for *inference purposes*. For that one needs to go beyond this framework and introduce a statistical modeling perspective which recasts the approximation problem into one of modeling the 'systematic information' in the data, i.e. embed the curve-fitting into *a statistical model* specified in terms of probabilistic assumptions. This enables one to address *statistical adequacy*: the probabilistic assumptions constituting the statistical model are valid for the data in question. Equivalently, the residuals from a fitted curve are *non-systematic* (as opposed to small) in a specific probabilistic sense. To shed further light on why the mathematical approximation perspective is inadequate as a basis for inductive inference, we need to untangle the two perspectives.

3.1   Legendre's mathematical approximation perspective

The problem of curve-fitting as specified by (1)-(3) fits perfectly into Legendre's least-square approximation perspective. Developments in mathematical approximation theory since Legendre (1805), ensure that under certain smoothness conditions on the true function $h(.)$, the approximating function $g_m(x; \boldsymbol{\alpha}) = \sum_{i=0}^{m} \alpha_i \phi_i(x)$, specified over a net of points $\mathbb{G}_n(\mathbf{x}) := \{x_k, \ k = 1, \ldots, n\}$, where $y_k = h(x_k)$, provides a solution to the problem, in the sense that the error of approximation $\varepsilon(x_k, m) = h(x_k) - g_m(x_k; \boldsymbol{\alpha})$, converges to zero on $\mathbb{G}_n(\mathbf{x})$ as $m \to \infty$:

$$\lim_{m \to \infty} \sum_{k=1}^{n} |\varepsilon(x_k, m)|^2 = 0. \tag{4}$$

The convergence in (4), known as *convergence in the mean*, implies that for every $\epsilon > 0$ there exists a large enough integer $N(\epsilon)$ such that:

$$\sum_{k=1}^{n} |\varepsilon(x_k, m)|^2 < \epsilon, \text{ for } m > N(\epsilon). \tag{5}$$

One can achieve a stronger form of convergence, known as *uniform convergence*, by imposing some additional smoothness restrictions on $h(x)$, say, continuous second derivatives. This ensures that, for every $\epsilon > 0$ there exists a large enough integer $N(\epsilon)$ such that:

$$\sum_{k=1}^{n} |\varepsilon(x_k, m)| < \epsilon, \text{ for } m > N(\epsilon) \text{ and all } x_k \in \mathbb{G}_n(\mathbf{x}). \tag{6}$$

These are well-known results in mathematics; see Hilderbrand (1974), Isaacson and Keller (1994). What is less well-known, or appreciated, is that there is nothing in the above approximation results which prevents the residuals

$$\widehat{\varepsilon}(x_k, m) = (y_k - g_m(x_k; \widehat{\boldsymbol{\alpha}})), \ k = 1, 2, ..., n,$$

where $\widehat{\boldsymbol{\alpha}}$ denotes the least-squares estimator of $\boldsymbol{\alpha}$, from varying systematically with k, x_k and m. These approximation results provide at best an upper bound for the

magnitude of $\widehat{\varepsilon}(x_k, m)$, and under some additional restrictions on the smoothness of $h(x)$, they might provide an explicit expression for $\widehat{\varepsilon}(x_k, m)$ in terms of x_k and m. However, a closer look at these theorems reveals that the residuals usually do vary *systematically* with k, x_k and m, rendering the use of the estimated curve for inference purposes highly problematic; it contradicts the claim that the fitted curve accounts for the regularities in the data. A typical result in this literature is the following theorem.

Theorem 1. Let $h(x)$ be a continuous function on $[a, b]$ and $g_m(x; \boldsymbol{\alpha}) = \sum_{i=0}^{m} \alpha_i \phi_i(x)$, a polynomial approximation to $h(x)$ on $[a, b]$. $g_m(x; \boldsymbol{\alpha})$ will provide a best (and unique) approximation if and only if (iff), the error of approximation $\varepsilon(x, m) = h(x) - g_m(x; \boldsymbol{\alpha})$, $x \in [a, b]$, takes values $\max_{x \in [a,b]} |\varepsilon(x, m)|$, with *alternating changes in sign* at least $m + 2$ times over the interval $[a, b]$; see Isaacson and Keller (1994).

The *iff* condition of 'alternating changes in sign' for the residuals indicates the presence of systematic information in a probabilistic sense. This is because the presence of cycles in the residuals $\{\widehat{\varepsilon}(x_k, m), \ k = 1, 2, ..., n\}$, indicates the presence of dependence over $k = 1, 2, ..., n$.

To get some idea as to how the approximation error term varies systematically with (x, m), consider the case of Largrange interpolation. Let $h(x)$ be a continuous function on $[a, b]$. The *Lagrange interpolation polynomial* on a net of points $\mathbb{G}_n(\mathbf{x}) := \{x_k, \ k = 1, \ldots, n\}$, $n \geq m$, spanning the interval $[a, b]$, takes the form:

$$g_m(x; \boldsymbol{\alpha}) = \sum_{i=0}^{m} y_i \prod_{j=0, j \neq i}^{m} \left(\frac{x - x_j^*}{x_i^* - x_j^*} \right), \quad x \in [a, b], \tag{7}$$

where the interpolation points $(x_0^*, x_1^*, \cdots, x_m^*, x) \in [a, b]$ are chosen to be distinct. For a smooth enough function $h(x)$ (derivatives up to order $m + 1$ exist), the error is a *systematic function* of both x and m since:

$$\varepsilon(x, m) = \frac{h^{(m+1)}(\xi)}{(m+1)!} \prod_{j=0}^{m} \left(x - x_j^* \right), \quad \xi \in [a, b], \tag{8}$$

where $h^{(m+1)}(x) = \frac{d^{m+1} h(x)}{d^{m+1} x}$. (8) suggests that the error curve $\varepsilon(x, m)$ behaves like a polynomial in x over $[a, b]$, with $m + 1$ roots $(x_0^*, x_1^*, \cdots, x_m^*)$:

$$\varepsilon(x, m) = a(x - x_0^*)(x - x_1^*) \cdots (x - x_m^*) = ax^{m+1} + b_m x^m + \cdots + b_1 x + b_0.$$

Such an *oscillating curve* is also typical for error term arising from the least squares approximation; see Dahliquist and Bjorck (1974).

Hence, the results in (4)-(6) provide no basis to assess the reliability of any inductive inference because (i) the approximation error varies systematically with k, x and m, and (ii) provides no means to assess the reliability of any inference based on a fitted curve. Moreover, both convergence results are based on the degree of the polynomial going to infinity, i.e. $m \to \infty$ guarantees the convergence to the true function $h(x)$. For a given m and sample size n, these results provide no basis to evaluate the reliability or the precision of inference.

To complicates matters even more, the convergence results in (4)-(6) are very different form of convergence results needed to establish asymptotic properties, such as *consistency*, for the least squares estimator $\widehat{\boldsymbol{\alpha}}_{LS}$; the latter is associated with the sample size $n \to \infty$. Consistency, however, presupposes that the sample size is large enough to ensure that m achieves an adequate approximation to begin with. Often this means that n must be much larger than m; a rule of thumb is $2\sqrt{n} > m$ — see Dahliquist and Bjorck (1974).

In classical (frequentist) statistics one assesses the reliability of inference using the *error probabilities* associated with the particular inference method or procedure. No such error probabilities can be evaluated on the basis of the above approximation results in (4)-(6). What is missing is some form of *probabilistic structure* that would state the circumstances under which the errors do *not* vary systematically with k, x and m. Indeed, one needs such a probabilistic structure, however rudimentary, to be able to talk about $\widehat{\boldsymbol{\alpha}}_{LS}$ being *consistent* for $\boldsymbol{\alpha}$; see Spanos (2006a). It is argued below that one needs to assume a probabilistic structure rich enough to give rise to error probabilities which are necessary to assess the reliability of inference.

3.1.1   The bottom line

In summary, the mathematical approximation method, using least squares will take one up to the estimation of a curve, say (see section 6):

$$\widehat{y}_t \;\; = 167.115 + 1.907x_t, \;\;\; s = 1.7714, \; T = 35, \tag{9}$$

but it provides inadequate information for reliable error bounds on these point estimates. No reliable inference can be drawn on the basis of (9) without probabilistic assumptions to render possible a reliability assessment; see Spanos (1999, 2006b).

3.2   Gauss's statistical modeling perspective

Gauss's major contribution was to embed the mathematical approximation formulation into a statistical model. He achieved that by transforming the approximation error term:

$$\varepsilon_k(x_k, m) = y_k - \sum_{i=0}^{m} \alpha_i \phi_i(x_k), \tag{10}$$

into a generic statistical error (free of x and m) :

$$\varepsilon_k(x_k, m) = \varepsilon_k \curvearrowright \mathsf{NIID}(0, \sigma^2), \;\; k = 1, 2, \ldots, n, ..., \tag{11}$$

where $\mathsf{NIID}(0, \sigma^2)$ stands for 'Normal, Independent and Identically Distributed with mean 0 and variance σ^{2}'. The error in (11) is *non-systematic,* in a probabilistic sense. This he achieved by imposing *additional probabilistic structure* on the error freeing it from its dependence on (x_k, m). Gauss (1809) effectively recast the original

mathematical approximation into a statistical modeling problem based on what we nowadays call the *Gauss Linear model* (see table 1):

$$y_k = \sum_{i=0}^{m} \alpha_i \phi_i(x_k) + \varepsilon_k, \quad \varepsilon_k \backsim \mathsf{NIID}(0, \sigma^2), \quad k = 1, 2, \ldots, n, \ldots \quad (12)$$

What makes his contribution all-important is that the statistical model in (12) provides the premises for assessing the reliability of inductive inference; see Farebrother (1999) for an excellent summary of the historical background.

To summarize the argument which will unfold in the next three sections, when the curve-fitting problem is viewed from the statistical modeling perspective, the choices of (a) the family of curves, and (b) the fittest curve within this family (section 2), need to be reconsidered. The 'fittest curve' is no longer the one achieving the optimal trade-off between the smallest sum of squared residuals and the number of parameters, but the one that *captures the 'regularities' in the data'*, irrespective of the number of parameters needed to achieve that. 'Capturing the regularities' needs to be operationalized in the form of a curve $g_m(x_k; \boldsymbol{\alpha})$ (a statistical model) whose residuals $\{[y_k - g_m(x_k; \widehat{\boldsymbol{\alpha}})], \ k = 1, 2, ..., n\}$ are non-systematic in a probabilistic sense. The Gauss modeling framework provides a way to assess whether the residuals contain systematic information; this needs to be established vis-a-vis the particular data $\mathbf{z}_0$.

A statistically adequate model provides a reliable basis for inductive inference using the error probabilities associated with the different inference procedures to calibrate their trustworthiness. A *statistically inadequate* model does not capture the regularities in the data and, as a result, the reliability of the associated inference is called into question. This modeling perspective is known as the *error statistical approach*, the origins of which can be traced back to Gauss (1809). Before we discuss the error statistical approach, however, it is important to discuss the role of least squares in the context of statistical modeling. As argued by Hald (1998): "It is important to distinguish between Legendre's algebraic method and Gauss's probabilistic method of least squares, which is uniquely defined by the statistical model." (p. 383)

3.2.1 What does Least Squares provide *the* standard solution of?

Least squares, as a mathematical approximation method, can be justified in that context by how well the resulting curves approximate the unknown function. However, least squares as a statistical estimation method needs to be justified on probabilistic grounds. The argument made above is that what justifies the choice of the fittest curve using (3) is statistical adequacy, not the least-squares method. This is contrary to the prevailing claim that the selection of the best fitting curve is well understood in the sense that least-squares provides *the* standard solution - whatever that means.

The best probabilistic justification for least-squares was first proposed by Gauss (1809) in the form of the Normal distribution for the errors:

$$f(\varepsilon_k) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{\varepsilon_k^2}{\sigma^2}\right), \quad \varepsilon_k \in \mathbb{R}, \quad k = 1, 2, \ldots, n, \ldots$$

This assumption ensures that the Least-Squares estimators $\widehat{\boldsymbol{\alpha}}_{LS}$ of $\boldsymbol{\alpha} := (\alpha_1, \alpha_2, ..., \alpha_m)$ coincide with the Maximum Likelihood Estimators (MLE) $\widehat{\boldsymbol{\alpha}}_{ML}$, whose justification stems from their optimal properties. MLEs, under certain regularity conditions, enjoy a number of optimal properties such as *invariance, unbiasedness-full efficiency, sufficiency, consistency, asymptotic efficient and Normality*, given the premises; see Spanos (1999), ch. 13. Of particular value is the property of *invariance* which asserts that if $\widehat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta} := (\alpha_1, \alpha_2, ..., \alpha_m, \sigma^2)$, then for $\boldsymbol{\phi} = \mathbf{H}(\boldsymbol{\theta})$, for any (Borel) function $\mathbf{H}(.)$, $\widehat{\boldsymbol{\phi}} = \mathbf{H}(\widehat{\boldsymbol{\theta}})$ is the MLE of $\boldsymbol{\phi}$.

When the distribution of the error term is assumed to be, say, the *Laplace* distribution:

$$f(\varepsilon_k) = \tfrac{1}{2\sigma} \exp\left(-\tfrac{|\varepsilon_k|}{\sigma}\right), \ \ \varepsilon_k \in \mathbb{R}, \ \ k = 1, 2, \ldots, n, \ldots,$$

the Least-Squares estimators of $\boldsymbol{\alpha}$ are not longer optimal in the sense defined above. The appropriate objective function to minimize from the probabilistic perspective is no longer the sum of squared errors (3), but the sum of the *Least Absolute Deviations* (LAD):

$$\sum_{k=1}^{n} \left| y_k - \sum_{i=0}^{m} \alpha_i \phi_i(x_k) \right|, \tag{13}$$

giving rise to the LAD estimator, which is optimal, in the sense that it coincides with the MLE.

Similarly, when the error distribution is assumed to be *Uniform* over $[-\sigma, \sigma]$:

$$f(\varepsilon_k) = \tfrac{1}{2\sigma}, \ \ \varepsilon_k \in [-\sigma, \sigma], \ \ k = 1, 2, \ldots, n, \ldots,$$

the appropriate objective function is:

$$\sup_{\varepsilon_k \in [-\sigma, \sigma]} \left| y_k - \sum_{i=0}^{m} \alpha_i \phi_i(x_k) \right|, \tag{14}$$

and the resulting estimator coincides with the MLE, ensuring its optimality.

It is important to bring out the fact that in the context of *mathematical approximation theory*, the norms underlying (3), (13) and (14) are special cases of the L_p norm:

$$L_p(\epsilon) = \left[\int_{x \in \mathbb{R}_x} |\varepsilon(x, m)|^p \, dx \right]^{\frac{1}{p}}, \quad p \geq 1,$$

for $p = 2$, $p = 1$ and $p = \infty$, respectively. However, from the statistical modeling perspective the choice of any one of these norms is completely arbitrary unless it is justified on probabilistic grounds via the optimality of the estimators they give rise to, as argued above.

3.2.2   The Gauss-Markov theorem

Ironically, Gauss's embedding of the mathematical approximation problem into a statistical model is rarely appreciated as the major contribution that it is; see Spanos

(1986). Instead, what Gauss is widely credited for is a theorem that is often misconstrued as providing a formal justification for least squares via the optimality of the estimators it gives rise to. This is the celebrated *Gauss-Markov theorem*, which relaxes the Normality assumption in (12), i.e.

$$y_k = \sum_{i=0}^{m} \alpha_i \phi_i(x_k) + \varepsilon_k, \ \ k = 1, 2, \ldots, n, \ldots$$

$$E(\varepsilon_k) = 0, \ E(\varepsilon_k^2) = \sigma^2, \ E(\varepsilon_k \varepsilon_j) = 0, \ \text{for } k \neq j.$$

(15)

In the context of the statistical model in (15), Gauss (1823a,b) proved that the least squares estimator $\widehat{\boldsymbol{\alpha}}_{LS}$ has minimum variance within the class of *linear* and *unbiased* estimators of $\boldsymbol{\alpha}$.

As argued in Spanos (1986, 1999), the property of minimum variance, within a very restrictive class of estimators (linear and unbiased), is nothing to write home about. Indeed, this theorem cannot be used as a basis of *reliable* inference, because it does not furnish a sampling distribution for $\widehat{\boldsymbol{\alpha}}_{LS}$, in terms of which the relevant error probabilities can be evaluated, and thus the reliability of inference can be assessed. Having only the first two moments of the sampling distribution of $\widehat{\boldsymbol{\alpha}}_{LS}$, provides a poor basis for reliable inference. This is because without a sampling distribution for $\widehat{\boldsymbol{\alpha}}_{LS}$ one is forced to use probabilistic inequalities (in conjunction with the first two moments) to evaluate the relevant error probabilities. These error bounds are usually very crude, giving rise to highly imprecise inferences; see Spanos (1999), ch. 10. Moreover, invoking asymptotic arguments (as $n \to \infty$) to approximate the sampling distribution of $\widehat{\boldsymbol{\alpha}}_{LS}$ by a Normal distribution, defeats the whole purpose of not assuming Normality at the outset; there is no way one can establish the reliability of any inference based on the asymptotic distribution unless one tests the Normality assumption; see Spanos (2002).

4   A summary of the Error-Statistical framework

The term *error-statistical* approach was coined by Mayo (1996) to denote a modification/extension of the framework for frequentist inductive inference, usually associated with Fisher, Neyman and Pearson. The modification/extension comes primarily in the form of delineating the central role of error probabilities in delimiting the reliability of inference, and supplementing the original framework with a post-data assessment of inference in the form of severity evaluations.

An important feature of the error-statistical approach is the distinction between different types of models that will enable one to bridge the gap between the phenomenon of interest and the data, the primary objective being to learn from the data about the phenomenon of interest. In direct analogy to the series of models proposed by Mayo (1996), we distinguish between a theory (primary) model, a structural (experimental) model and a statistical (data) model; see Spanos (2006a,b) for further discussion.

4.1 Structural vs. Statistical Models

In postulating a *theory model* to explain the behavior of an observable variable, say y_k, one demarcates the segment of reality to be modeled by selecting the primary influencing factors $\mathbf{x}_k$, well aware that there might be numerous other potentially relevant factors $\boldsymbol{\xi}_k$ (observable and unobservable) influencing the behavior of y_k. This reasoning gives rise to a generic *theory model:*

$$y_k = h^*(\mathbf{x}_k, \boldsymbol{\xi}_k), \quad k \in \mathbb{N}. \tag{16}$$

Indeed, the potential presence of a large number of contributing factors explains the invocation of *ceteris paribus* clauses. The guiding principle in selecting the variables in $\mathbf{x}_k$ is to ensure that they collectively account for the *systematic* behavior of y_k, and the omitted factors $\boldsymbol{\xi}_k$ represent non-essential disturbing influences which have only a non-systematic effect on y_k. This line of reasoning transforms the theory model (16) into a *structural model* of the form:

$$y_k = h(\mathbf{x}_k; \boldsymbol{\phi}) + \epsilon(\mathbf{x}_k, \boldsymbol{\xi}_k), \quad k \in \mathbb{N}, \tag{17}$$

where $h(.)$ denotes the postulated functional form, $\boldsymbol{\phi}$ stands for the structural parameters of interest, and:

$$\left\{ \epsilon(\mathbf{x}_k, \boldsymbol{\xi}_k) = y_k - h(\mathbf{x}_k; \boldsymbol{\phi}), \, k \in \mathbb{N} \right\}, \tag{18}$$

is the structural error term, viewed as a function of both $\mathbf{x}_k$ and $\boldsymbol{\xi}_k$, representing all unmodeled influences. For (18) to provide a meaningful model for y_k the error term needs to be non-systematic: a *white-noise* (non-systematic) stochastic process $\{\epsilon(\mathbf{x}_k, \boldsymbol{\xi}_k), \, k \in \mathbb{N}\}$ satisfying the properties:

$$\left. \begin{array}{ll} \text{[i]} & E(\epsilon(\mathbf{x}_k, \boldsymbol{\xi}_k)) = 0, \\ \text{[ii]} & E(\epsilon^2(\mathbf{x}_k, \boldsymbol{\xi}_k) \cdot \epsilon(\mathbf{x}_\ell, \boldsymbol{\xi}_\ell)) = \sigma^2, \\ \text{[iii]} & E(\epsilon(\mathbf{x}_k, \boldsymbol{\xi}_k) \cdot \epsilon(\mathbf{x}_\ell, \boldsymbol{\xi}_\ell)) = 0, \, k \neq \ell, \, k, \ell \in \mathbb{N}, \end{array} \right\} \quad \forall (\mathbf{x}_k, \boldsymbol{\xi}_k) \in \mathbb{R}_{\mathbf{x}} \times \mathbb{R}_{\boldsymbol{\xi}}. \tag{19}$$

In addition to [i]-[iii], one needs to ensure that the generating mechanism (17) is 'nearly isolated' in the sense that the unmodeled component (the error $\epsilon(\mathbf{x}_k, \boldsymbol{\xi}_k)$) is *uncorrelated* with the modeled influences (systematic component $h(\mathbf{x}_k; \boldsymbol{\phi})$):

$$\text{[iv]} \quad E(\epsilon(\mathbf{x}_k, \boldsymbol{\xi}_k) \cdot h(\mathbf{x}_k; \boldsymbol{\phi})) = 0, \, \forall (\mathbf{x}_k, \boldsymbol{\xi}_k) \in \mathbb{R}_{\mathbf{x}} \times \mathbb{R}_{\boldsymbol{\xi}};$$

see Spanos (1986,1995).

The above perspective on theory and structural models provides a much broader framework than that of mathematical approximation. In common with the mathematical approximation perspective, however, the error term is statistically non-opearational. Assumptions [i]-[iv] are empirically non-verifiable because their assessment would involve *all possible values* of both $\mathbf{x}_k$ and $\boldsymbol{\xi}_k$. To render them testable

one needs to embed this structural into a statistical model; a crucial move that often goes unnoticed.

The embedding itself depends crucially on whether the data $\mathbf{Z} := (\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n)$ are the result of an experiment or they are non-experimental (observational) in nature, but the aim in both cases is to find a way to transform the structural error $\epsilon(\mathbf{x}_k, \boldsymbol{\xi}_k)$, $(\mathbf{x}_k, \boldsymbol{\xi}_k) \in \mathbb{R}_\mathbf{x} \times \mathbb{R}_{\boldsymbol{\xi}}$ into a generic white noise error process.

In the case where one can perform experiments, 'experimental design' techniques such as *randomization* and *blocking* are often used to 'neutralize' and 'isolate' the phenomenon from the potential effects of $\boldsymbol{\xi}_k$ by ensuring that the uncontrolled factors cancel each other out; see Fisher (1935). The idea is to transform the structural error into a generic white noise process:

$$\epsilon(\mathbf{x}_k, \boldsymbol{\xi}_k) = \varepsilon_k \backsim \mathsf{IID}(0, \sigma^2), \qquad k = 1, ..., n. \tag{20}$$

This embeds the structural model (17) into a *statistical model* of the form:

$$y_k = h(\mathbf{x}_k; \boldsymbol{\theta}) + \varepsilon_k, \quad \varepsilon_k \backsim \mathsf{IID}(0, \sigma^2), \ k = 1, 2, ..., n, \tag{21}$$

where the statistical error term ε_k in (21) is qualitatively very different from the structural error term $\epsilon(\mathbf{x}_k, \boldsymbol{\xi}_k)$ in (17) because ε_k is no longer a function of $(\mathbf{x}_k, \boldsymbol{\xi}_k)$ and its assumptions are rendered empirically testable; see Spanos (2006b-c) for further details. A widely used special case of (21) is the *Gauss Linear model*, a simple form of which is given in table 1. The name was chosen because Gauss was the first to embed a structural model into a statistical one; see section 3.2 above.

Table 1 - The Gauss Linear Model

Statistical GM:	$y_t = \sum_{i=0}^m \beta_i \phi_i(x_k) + \varepsilon_t, \ t \in \mathbb{T},$
[1] Normality:	$y_t \backsim \mathsf{N}(.,.),$
[2] Linearity:	$E(y_t) = \sum_{i=0}^m \beta_i \phi_i(x_k)$, linear in $\boldsymbol{\beta}$,
[3] Homoskedasticity:	$Var(y_t) = \sigma^2$, not changing with x_t,
[4] Independence:	$\{y_t, \ t \in \mathbb{T}\}$ is independent process,
[5] t-invariance:	$(\boldsymbol{\beta}, \sigma^2)$ do not change with t.

This is the case where the observed data on $(y_k, \mathbf{x}_k)$ are the result of an ongoing actual data generating process, the experimental control and intervention are replaced by judicious conditioning on an appropriate conditioning information set $\mathfrak{D}_t$ to transform the structural error into a generic white-noise statistical error:

$$(u_t | \mathfrak{D}_t) \backsim \mathsf{IID}(0, \sigma^2), \ k = 1, 2, ..., n. \tag{22}$$

Spanos (1986, 1999) demonstrates how sequential conditioning provides a general way to decompose orthogonally a stochastic process $\{\mathbf{Z}_t, \ t \in \mathbb{T}\}$ into a systematic

component μ_t and a *martingale difference process* u_t relative to a conditioning information set $\mathfrak{D}_t$; a modern form of a white-noise process – see Spanos (2006b) for further details.

A widely used special case of (22) is the *Normal/Linear Regression model* given in table 2, where assumptions [**1**]-[**5**] assumptions pertain to the structure of the observable process $\{(y_t| \mathbf{X}_t = \mathbf{x}_t),\ t \in \mathbb{T}\}$, $\mathbf{Z}_t := (y_t : \mathbf{X}_t^\top)^\top$.

Table 2 - The Normal/Linear Regression Model

Statistical GM:	$y_t = \beta_0 + \boldsymbol{\beta}_1^\top \mathbf{x}_t + u_t,\ t \in \mathbb{T},$

[**1**] **Normality:**	$(y_t \mid \mathbf{X}_t = \mathbf{x}_t) \backsim \mathsf{N}(.,.),$
[**2**] **Linearity:**	$E(y_t \mid \mathbf{X}_t = \mathbf{x}_t) = \beta_0 + \boldsymbol{\beta}_1^\top \mathbf{x}_t,$ linear in $\mathbf{x}_t$,
[**3**] **Homoskedasticity:**	$Var(y_t \mid \mathbf{X}_t = \mathbf{x}_t) = \sigma^2,$ free of $\mathbf{x}_t$,
[**4**] **Independence:**	$\{(y_t \mid \mathbf{X}_t = \mathbf{x}_t),\ t \in \mathbb{T}\}$ is an independent process,
[**5**] **t-invariance:**	$\boldsymbol{\theta} := (\beta_0, \boldsymbol{\beta}_1, \sigma^2)$ do not change with t.

At this point it is important to emphasize that the premises of statistical induction, comes in the form of a pair $(y_0, \mathcal{M}_\theta(\mathbf{y}))$, where $\mathbf{y}_0 := (y_1, y_2, ..., y_n)$ denotes the observed data and $\mathcal{M}_\theta(\mathbf{y})$ the statistical model of interest:

$$\mathcal{M}_\theta(\mathbf{y}) = \{f(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta\},\ \mathbf{y} \in \mathcal{Y} := \mathbb{R}_Y^n,$$

where $\mathbf{Y} := (Y_1, Y_2, ..., Y_n)$ denotes the sample, Θ the parameter space and $\mathcal{Y}$ the sample space. The connection between $\mathcal{M}_\theta(\mathbf{y})$ and data $\mathbf{y}_0$ is that the latter is viewed as a 'truly typical realization' of the stochastic mechanism described by the former. Whether a particular data $\mathbf{y}_0$ constitute a truly typical realization of $\mathbf{Y}$ can be assessed using Mis-Specification (M-S) testing: testing the probabilistic assumptions underlying $\mathcal{M}_\theta(\mathbf{y})$; see Spanos (1999). Statistical inference comes in the form of claims concerning the true value of $\boldsymbol{\theta} \in \Theta$. The reliability of such inferences is evaluated using the relevant error probabilities; see Mayo (1996), Spanos (2006a).

5 Curve-fitting vs. the error-statistical approach

When viewed from the error statistical perspective summarized in the previous section, curve-fitting becomes an empirical modeling problem, and this perspective sheds a very different light on the problems raised by the current discussions of curve-fitting.

In particular, the choice of a family of curves in (a) and the choice of the 'best' in (b) are inextricably bound up when the primary criterion for 'best' is *statistical adequacy*: the probabilistic assumptions constituting the premises of inference (see

table 1) are valid for the data in question, i.e. statistical adequacy becomes a *necessary* criterion for adjudicating both (a) choosing the family of curves, and (b) choosing the fittest curve within this family, in section 2. Indeed, the probabilistic assumptions constituting the statistical model provide the framework in the context of which one can operationalize the notion that 'a curve *captures the 'regularities' in the data*'; see Spanos (1999).

Another important implication of viewing the curve-fitting problem from the error-statistical perspective is that the *trade-off* between goodness-of-fit and parsimony becomes largely irrelevant. The approximating function $g_m(x_k; \boldsymbol{\alpha}) = \sum_{i=0}^{m} \alpha_i \phi_i(x_k)$ is chosen to be *as elaborate as necessary* to ensure that it captures all the systematic information in the data, *but no more elaborate*. This guards effectively against overfitting by ensuring that the residuals:

$$\widehat{\varepsilon}_k = y_k - \sum_{i=0}^{m} \widehat{\alpha}_i \phi_i(x_k), \quad k = 1, 2, \ldots, n, \tag{23}$$

are non-systematic. Overfitting, such as the inclusion of higher degree polynomials, or more lags, than needed, is likely to give rise to systematic residuals; see Spanos (1986), p. 479. One can guard against such overfitting by careful and thorough M-S testing. If these tests detect systematic information in the residuals one respecifies the model, in an attempt to account for the systematic information indicated by the specific departure(s) such as non-Normality, non-linearity, heteroskedasticity, etc., until a new, statistically adequate model, is ascertained. Hence, statistical adequacy is the only criterion to be used to determine the 'fittest' curve, a criterion which involves much more than the choice of the optimal value of m.

Statistical adequacy is both necessary and sufficient for the statistical reliability of the any inference based on the estimated fittest curve, but it does not guarantee substantive adequacy; see Spanos (2006b).

5.1   Revisiting underdetermination

Returning to the curve-fitting problem viewed in the context of the error-statistical approach reveals that the problem of *underdetermination* is not as pervasive as commonly assumed. This is because '*capturing the regularities in the data*' is not just a matter of goodness-of-fit and/or parsimony, but it involves ensuring the statistical adequacy of the estimated model. Indeed, finding a single statistically adequate model for the data in question is a daunting task; finding more than one is rather rare; see Spanos (1999). Moreover, the simplistic view that one can accommodate the data by choosing a polynomial of degree $m = n - 1$, giving rise to the *Lagrange interpolation polynomial* in (7) ignores the fact that such a polynomial is unusable for inference purposes because it amounts to reshuffling the original data; it does not constitute a statistical model, or even a restriction on the data. In this case one trades the $m+1$ data points $\{(x_i, y_i), \ i=0, 1, ..., m\}$ with the $m+1$ estimated coefficients $(\widehat{\alpha}_0, \widehat{\alpha}_1, ..., \widehat{\alpha}_m)$ using a one-to-one mapping.

The *predictive accuracy* of a fitted curve (statistical model) is no longer just a matter of 'small' prediction errors, but *non-systematic* prediction errors. A statistically adequate curve $g_m(x; \hat{\alpha})$ captures all the systematic information in the data and unless the the invariance structure of the underlying data generating process has changed, it will give rise to non-systematic prediction errors. Any fitted curve $g_m(x; \hat{\alpha})$ that is not statistically adequate is likely to systematically over-predict or under-predict the values $\{y_k, \ k=n+1, n+2, ...\}$, and is rendered weak on predictive grounds. This is contrary to the claims by Hitchcock and Sober (2004) that "predictive accuracy provides evidence that a hypothesis has appropriately balanced simplicity against goodness-of-fit." (see ibid., p. 20). Viewing predictive accuracy in terms of 'small' (however that is defined) prediction error is nothing more than goodness-of-fit projected beyond the observation period. As such it suffers from the same problems as any goodness-of-fit criterion. Even astrologers get it right occasionally, but that's no basis to deduce that their theories have substantive explanatory power.

To substantiate the above arguments let us revisit two widely discussed examples of curve-fitting, Kepler's and Ptolemy's models of planetary motion.

5.2   Kepler's model of planetary motion

Kepler's model for the elliptical motion of Mars turned out to be a real regularity because the statistical model, in the context of which his structural model was embedded, can be shown to be statistically adequate. To see this, consider Kepler's structural model:

$$y_t = \alpha_0 + \alpha_1 x_t + \epsilon(x_k, \boldsymbol{\xi}_k), \ t \in \mathbb{T}, \tag{24}$$

where $y := (1/r)$ and $x := \cos \vartheta$, r - *the distance of the planet from the sun, ϑ - the angle between the line joining the sun and the planet and the principal axis of the ellipse*; see fig. 1.
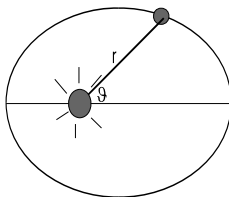

Fig. 1: Elliptical motion of planets

It is important to emphasize that historically this law was originally proposed by Kepler not as a structural relationship, but as an *empirical regularity* that he 'deduced' from Brahe's data; when he proposed this law in 1609 he had no way to interpret the parameters (α_0, α_1) and no explanation for the elliptic motion.

The structural interpretation of Kepler's first law stems from the fact that the parameters (α_0, α_1) enjoy a clear *theoretical interpretation* bestowed upon them by

Newton's (1687) law of universal gravitation $F = \frac{G(m \cdot M)}{r^2}$, where F is the force of attraction between two bodies of *mass m* (planet) and M (sun); G is a *constant of gravitational attraction*. Hence, the *structural interpretation*:

$\alpha_0 = \frac{MG}{4\kappa^2}$, where κ denotes Kepler's constant,

$\alpha_1 = (\frac{1}{d} - \alpha_0)$, where d is the shortest distance between the planet and the sun; see Hahn (1998) for further details. Moreover, the error term $\epsilon(x_k, \boldsymbol{\xi}_k)$ also enjoys a structural interpretation in the form of 'deviations' from the elliptic motion due to potential *measurement errors* as well as other *unmodeled effects*. Hence, the white-noise error assumptions [i]-[iii] in (19) are *inappropriate* in cases where: (i) the data suffer from 'systematic' observation errors, (ii) *the third body problem effect* is significant, (iii) the *general relativity* terms (see Lawden, 2002) are significant.

Embedding (24) into the Normal/Linear Regression model (table 2), and estimating it using **Kepler's original data** ($n = 28$) yield:

$$y_t = \underset{(.000002)}{0.662062} + \underset{(.000003)}{.061333} x_t + \widehat{u}_t, \ \ R^2 = .999, \ \ s = .0000111479. \tag{25}$$

The misspecification tests (see Spanos and McGuirk, 2001), reported in table 3, indicate most clearly that the estimated model is statistically adequate; the p-values are given in square brackets.

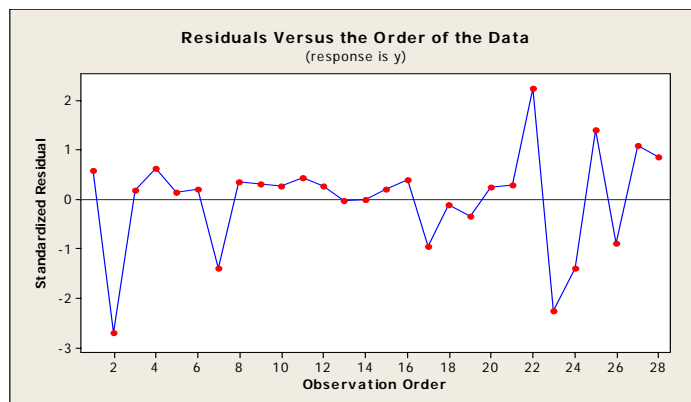| Table 3 - Misspecification tests	
Non-Normality:	$D'AP = 5.816[.106]$
Non-linearity:	$F(1, 25) = 0.077[.783]$
Heteroskedasticity:	$F(2, 23) = 2.012[.156]$
Autocorrelation:	$F(2, 22) = 2.034[.155]$

Fig. 2: Residuals from the Kepler regression

A less formal but more intuitive confirmation of the statistical adequacy of (25) is given by the residual plot in fig.2, which exhibits no obvious departures from a realization of a Normal, white-noise process.

On the basis of the statistically adequate model in (25), we can proceed to test the substantive hypothesis that the motion is circular, as assumed by Copernicus, in

the form of the hypotheses:
$$H_0 : \alpha_1 = 0, \ H_1 : \alpha_1 > 0.$$

The t-test yields: $\tau(\mathbf{y}) = \frac{.061333}{.000003} = 20444.3[.000000]$, providing very strong evidence against H_0; see Spanos (2006c) for further details.

5.3 Ptolemy's model of planetary motion[1]

The prevailing view in philosophy of science is that Ptolemy's geocentric model, despite being false, can 'save the phenomena' and yield highly accurate predictions. Indeed, the argument goes, one would have a hard time making a case in favor of Kepler's model and against the Ptolemaic model solely on *empirical grounds* because of the latter's excellent fit to planetary motion data; see Laudan (1977). Hence, the way to make such a case would be on the basis of other internal and external virtues of the two theories, including simplicity; see Sober (2000).

This prevailing view is called into question by demonstrating that the Kepler model does account for the regularities in the data, as shown above, but the Ptolemaic model does not, despite its excellent fit; the former is statistically adequate but the latter is not. Indeed, the Ptolemaic model is a quintessential example of curve-fitting which yields excellent goodness-of-fit but does not account for the regularities in the data.

The Ptolemaic model of the motion of an outer planet based on a single *epicycle*, with radius a, rolling on the circumference of a deferent of radius A, and an equant of distance c, can be parameterized in polar coordinates by the following model:

$$d_t^2 = \alpha_0 + \alpha_1 \cos(\varphi_t) + \alpha_2 \cos(\delta\varphi_t) + \alpha_3 \cos((\delta-1)\varphi_t) + \alpha_4 \sin(\varphi_t) + \alpha_5 \sin(3\varphi_t) + u_t,$$

where d_t denotes the distance of the planet from the earth, φ_t the angular distance measured eastward along the celestial equator from the equinox, and $\delta = \frac{A}{a}$. Ptolemy's model does not have a structural interpretation, but one can interpret the coefficients $(\alpha_0, ..., \alpha_5)$ in terms of the underlying geometry of the motion; see Spanos (2006c) for further details.

The data chosen to estimate this model are daily geocentric observations taken from Washington DC, referring to the *motion of Mars*, with a sample size $T = 687$, chosen to ensure a full cycle for Mars. Estimating this model by embedding it into the Normal/Linear Regression model (table 2) yields:

$$\begin{aligned}
d_t^2 &= \underset{(.047)}{2.77} - \underset{(.053)}{1.524}\cos(\varphi_t) - \underset{(.069)}{1.984}\cos(1.3\varphi_t) + \underset{(.106)}{2.284}\cos(.3\varphi_t) - \\
&\quad - \underset{(.087)}{2.929}\sin(\varphi_t) - \underset{(.014)}{.260}\sin(3\varphi_t) + \widehat{u}_t
\end{aligned} \tag{26}$$

$$R^2 = .992, \ s = 0.21998, \ T = 687.$$

---

[1] Estimating the Ptolemaic model was suggested to me by Clark Glymour.

The value of the scaling factor δ, chosen on goodness-of-fit grounds, is $\delta = 1.3$, which is not very different from the value for $\frac{A}{a} = 1.51$, assumed by Ptolemy.

A cursory look at the standardized residuals (see fig. 3) confirms the excellent goodness-of-fit $(R^2 = .992)$, in the sense that the residuals are small in magnitude since none of them is outside a band of 2.5 standard deviations from the average value zero. Despite being relatively small, a closer look reveals that the residuals exhibit crucial *departures from the white-noise assumptions* in the form of systematic statistical information. The cycles exhibited by the (standardized) residuals plot (fig. 3) reflect a departure from the independence assumption (they suggest the presence of Markov dependence - see Spanos, 1999, ch. 5), and the barely discernible (upward) mean trending indicates the presence of some heterogeneity. These patterns can be more clearly recognized when one compares the residual plot in fig. 3 with a t-plot of a typical white-noise realization given in fig. 4; analogical reasoning could easily bring out the differences between the two plots.
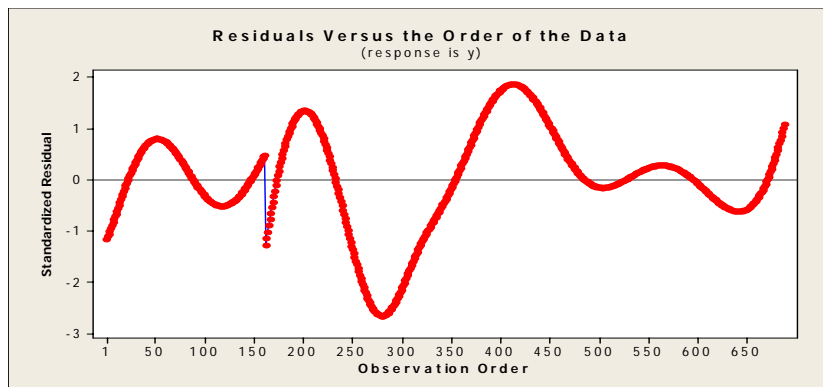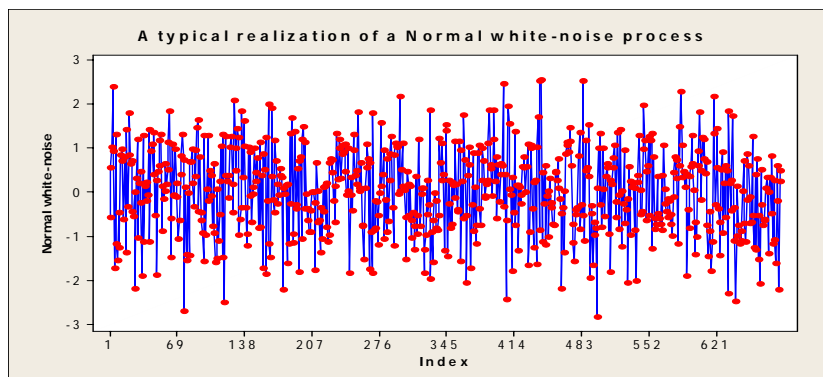

Fig. 3: t-plot of the residuals from (26)

Fig. 4: t-plot of a Normal white-noise realization

These highly significant departures are confirmed by proper *misspecification tests* shown in table 1 below; see Spanos and McGuirk (2001). The tiny p-values (zero to the sixth decimal) in square brackets indicate strong departures from all the statistical assumptions.

Table 4 - Misspecification tests	
Non-Normality:	$D\prime AP = 39.899[.00000]$
Non-linearity:	$F(2, 679) = 21.558[.00000]$
Heteroskedasticity:	$F(3, 677) = 77.853[.00000]$
Autocorrelation:	$F(2, 677) = 60993.323[.00000]$
Mean heterogeneity:	$F(1, 678) = 18.923[.00000]$

It is interesting to note that Babb (1977) gives a plot of the residuals for the Ptolemy model estimated for Mars, but ordered according to the angle of observation $(0°-360°)$, and that plot exhibits the same forms of departures from white-noise as fig. 3.

The Ptolemaic model has been widely praised as yielding highly accurate predictions, and that in turn was interpreted as an indication of the empirical validity of the model. The estimated model in (26) was used to predict the next 13 observations (688-700), and on the basis of Theil's measure:

$$U = \sqrt{\frac{\sum_{t=1}^{13}(y_t-\widehat{y}_t)^2}{\sum_{t=1}^{13}y_t{}^2+\sum_{t=1}^{13}\widehat{y}_t{}^2}} = .030, \quad y_t- \text{ actual}, \quad \widehat{y}_t- \text{ predicted},$$

the predictive accuracy seems excellent; $0 \leq U \leq 1$, the closer to zero the better - see Spanos (1986), p. 405. However, a plot of the actual and fitted values in fig. 5 reveals a very different picture: the predictive accuracy of the Ptolemaic model is very weak since it *underpredicts systematically*; a symptom of statistical inadequacy.
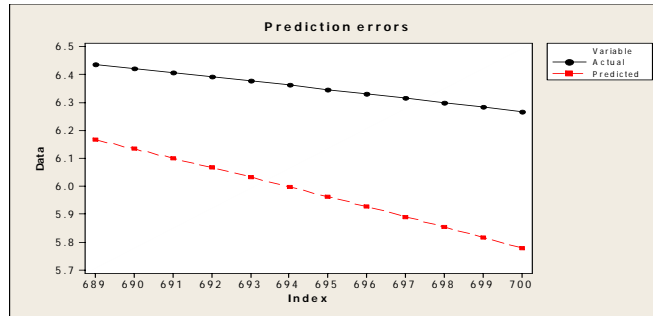

Fig.5: Actual vs. predicted

The discussion in section 3 explains the above empirical results associated with the Ptolemaic model as a classic example of how curve-fitting, as a mathematical approximation method, would usually give rise to systematic residuals (and prediction errors), irrespective of the goodness-of-fit. In this case the use of epicycles, is tantamount to approximating a periodic function $h(x)$ using orthogonal trigonometric polynomials $\{1, \; \cos x, \; \sin x, \; \cos 2x, \; \sin 2x, ..., \cos mx, \; \sin mx\}$. The approximating function takes the form:

$$g_m(x;\boldsymbol{\theta}) \;\; = \tfrac{1}{2}a_0 + \sum_{k=1}^{m}(a_k \cos kx + b_k \sin kx), \;\; \text{for } m \geq 1, \; x\in[-\pi, \pi].$$

19

It is obvious that every additional epicycle increases the degree of the fitted polynomial by one degree; this argument was first suggested by Bohr (1949)[2]. The claim that the Ptolemaic model can approximate the motions of the planets very well is an example of the following best approximation theorem for periodic functions using Fourier series.

Theorem 2. $g_m(x)$ provides the best approximation for a periodic function $h(x)$ *iff* the error of approximation $\varepsilon(x, m) = h(x) - g_m(x)$, $x \in [-\pi, \pi]$, takes values $\max_{x \in [-\pi, \pi]} |\varepsilon(x, m)|$, with *alternating changes in sign* at least $2m + 2$ over the interval $[-\pi, \pi]$. This approximation is unique; see Isaacson and Keller (1994).

As argued in section 3, the *iff* condition gives rise to cycles (see fig. 3) which indicate that the residuals are likely to contain systematic statistical information.

6   Problems with Akaike-type procedures

The Akaike Information Criterion (AIC) procedure is viewed as minimizing a penalize likelihood function which trades goodness-of-fit against parsimony – measured in terms of the number of unknown model parameters. In its general form the AIC is defined by (see Akaike, 1973):

$$\mathsf{AIC} = \text{-}2 \ln (\text{estimated likelihood}) + 2(\text{number of parameters}). \qquad (27)$$

For example, in the above case of the Gauss Linear model, the log-likelihood is:

$$\ln L(\boldsymbol{\theta}; \mathbf{z}) = -\frac{n}{2} \ln \left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} \sum_{k=1}^{n} \left(y_k - \sum_{i=0}^{m} \alpha_i \phi_i(x_k)\right)^2,$$

giving rise to:

$$\mathsf{AIC}(m) = \text{const.} + n \ln(\widehat{\sigma}^2) + 2m, \qquad (28)$$

where $\widehat{\sigma}^2 = \frac{1}{n} \sum_{k=1}^{n} \left(y_k - \sum_{i=0}^{m} \widehat{\alpha}_i \phi_i(x_k)\right)^2$.

What does the above discussion suggest for the Akaike model selection procedure as a way to address the curve-fitting problem?

To begin with the AIC procedure suffers from some well-documented weaknesses:
(i) in small samples leads to overfitting, the chosen $m > m^* -$ the true value, (see Hurvich and Tsai, 1989), and
(ii) asymptotically (as $n \to \infty$) the chosen m is not a consistent estimator of the true m^*; see Schwartz (1978), Hannan and Quinn (1979).

It is argued that in addition to these weaknesses, the AIC-type procedures suffer from major flaws that renders them inadequate for addressing the curve-fitting problem. The most important of these defects stem from the fact that these procedures:
(iii) ignore the problem of statistical adequacy and
(iv) allow for no 'control' of the error probabilities in drawing inferences.

---

[2]This reference was pointed out to me by Clark Glymour.

6.1 Searching within the 'wrong' family of models

The most crucial weakness of the Akaike model selection procedure is that it *ignores statistical adequacy*. The AIC criterion trades goodness-of-fit $n \ln(\widehat{\sigma}^2)$ against parsimony $2m$, but it takes the likelihood function as given, ignoring the validity of the statistical model that the likelihood function is supposed to provide a summary of. Statistical adequacy is the only way one can ensure that the likelihood function is valid. Hence, all proofs of consistency for such procedures assume that the likelihood function adopted is valid. However, that is exactly what is at stake. As seen above, the derivation of the likelihood function pre-supposes that assumptions [1]-[5] are valid. As argued by Lehmann (1990), when choosing the original family of models in (a), the AIC procedure assumes the problem of statistical adequacy away. Hence, when the choice in (a) is inappropriate (the statistical model is misspecified), guarding against overfitting by trading-off goodness-of-fit with parsimony makes little sense; it will lead to unreliable inferences with probability one.

Empirical example. An economist proposes a model to predict changes in the U.S.A. population:

$$M_1: \quad y_t = \underset{(.610)}{167.115} + \underset{(.024)}{1.907} x_t + \widehat{u}_t, \quad R^2 = .995, \quad s = 1.7714, \; T = 35, \qquad (29)$$

where y_t denotes the population (in millions) of the USA during the period 1955-1989 and x_t denotes a secret variable; the numbers in brackets underneath the estimates denote standard errors. On goodness-of-fit grounds this estimated relationship is excellent, but is the choice of m optimal? Let us consider applying the Akaike criterion by nesting it within the broader family of curves:

$$M(m): \quad y_k = \sum_{i=0}^{m} \alpha_i \phi_i(x_k) + \varepsilon_k, \qquad (30)$$

where $\phi_i(x_k)$ are *orthogonal* Chebyshev polynomials (see Hildebrand, 1974). The evaluation of the AIC criterion for model selection gives rise to the results in table 4 which suggest that the 'optimal' model is $m = 4$. It is important to confirm that this result does not change when one uses the small sample 'corrected' AIC:

$$\mathsf{AIC}_c(m) = n \ln(\widehat{\sigma}^2) + 2m + \left(\frac{2m(m-1)}{n-m-1} \right)$$

proposed by Hurvich and Tsai (1989).

Table 4 - Akaike model selection from (30)		
Model	$\mathsf{AIC}(\mathsf{m}) = n \ln(\widehat{\sigma}^2) + 2m,$	rank
$\mathsf{AIC}(1) =$	$(35) \ln(2.9586) + 2(3) = 43.965$	5
$\mathsf{AIC}(2) =$	$(35) \ln(2.5862) + 2(4) = 41.257$	3
$\mathsf{AIC}(3) =$	$(35) \ln(2.5862) + 2(5) = 43.257$	4
$\mathrm{AIC}(4) =$	$\mathbf{(35) \ln (1.8658) + 2(6) = 33.829}$	**1**
$\mathsf{AIC}(5) =$	$(35) \ln(1.8018) + 2(7) = 34.608$	2

(31)

As shown in Mayo and Spanos (2004), the original model (29), as well as the models in (30), are statistically misspecified; assumptions [1]-[5] (table 2) are all invalid. However, the AIC procedure ignores statistical adequacy and (misleadingly) selects model $M(4)$ (table 4) within a family of misspecified models.

It turns out that when one uses statistical adequacy as a guide to model selection, one is led to an entirely different family of statistical models:

$$M(k,\ell): \quad y_t = \beta_0 + \beta_1 x_t + \sum_{i=1}^{k} \delta_i t + \sum_{i=1}^{\ell}[a_i y_{t-i} + \gamma_i x_{t-i}] + \varepsilon_t, \qquad (32)$$

where t denotes a time trend and the lags $(x_{t-i}, y_{t-i}, \ i=1,2,...,\ell)$, k refers to the degree of the time polynomial and ℓ to the highest lag included in the model; see Mayo and Spanos (2004).

6.2   Searching within the 'right' family of models

The question that arises is whether, if one were to search within a family known to contain the true model, the AIC-preferred model will coincide with the one chosen on statistical adequacy grounds. As shown below, the answer is no. Let us demonstrate this result using the above data.

Empirical example - continued. Applying AIC criterion to select a model from the $M(k,\ell)$ family in (32) yields the results in table 5. The model selected by the AIC criterion is $M(3,2)$, which is different from the one selected on statistical adequacy grounds, $M(1,2)$; see Mayo and Spanos (2004). Hence, even in this case the AIC is likely to lead one astray. It is argued that the Akaike procedures is often unreliable because it constitutes a form of Neyman-Pearson (N-P) hypothesis testing with *unknown error probabilities*.

Table 5 - Akaike model selection from (32)		
Model	$\mathsf{AIC}(\mathsf{m}) = n\ln(\widehat{\sigma}^2) + 2m,$	rank
$M(1,1):$	$(35)\ln(.057555) + 2(6) = -87.925$	9
$M(1,2):$	$(35)\ln(.034617) + 2(8) = -101.72$	3
$M(1,3):$	$(35)\ln(.033294) + 2(10) = -99.083$	5
$M(2,1):$	$(35)\ln(.040383) + 2(7) = -98.327$	6
$M(2,2):$	$(35)\ln(.033366) + 2(9) = -101.01$	4
$M(2,3):$	$(35)\ln(.032607) + 2(11) = -97.813$	7
$M(3,1):$	$(35)\ln(.042497) + 2(8) = -94.541$	8
$\mathbf{M(3,2):}$	$\mathbf{(35)\ln(.029651) + 2(10) = -103.14}$	**1**
$M(3,3):$	$(35)\ln(.026709) + 2(12) = -102.80$	2

To see this, we note that the choice of the model $M(1,2)$ on statistical adequacy grounds involved testing the statistical significance of the coefficients:

$$H_0: \delta_2 = \delta_3 = \alpha_3 = 0, \text{ vs. } H_1: \delta_2 \neq 0, \text{ or } \delta_3 \neq 0, \text{ or } \alpha_3 \neq 0,$$

and not rejecting the null. In contrast, the Akaike procedure, by choosing $M(3,2)$ (table 5) inferred (indirectly) that the H_0 is false. What contributed to these different inferences? As shown below, the implicit type I error is often unusually high, giving rise to more frequent rejections of true null hypotheses than assumed.

6.3   N-P testing with unknown error probabilities

To simplify the derivations consider a special case of the above Gauss Linear model, where one needs to decide on the optimal m^* using ordinary polynomials, say, one compares the following two curves:

$$
\begin{aligned}
M_3: \quad & y_t = \beta_0 + \beta_1 x_t + \beta_2 x_t^2 + \beta_3 x_t^3 + \varepsilon_t, \quad m_3 = 4, \\
M_2: \quad & y_t = \alpha_0 + \alpha_1 x_t + \alpha_2 x_t^2 + u_t, \quad\quad\quad m_2 = 3.
\end{aligned}
$$

Let us assume that on the basis of the AIC procedure model M_3 was chosen, i.e. $m^* = 4$. That is, $AIC(m_2) > AIC(m_3)$, i.e.

$$
[n \ln(\widehat{\sigma}_1^2) + 2m_2] > [n \ln(\widehat{\sigma}_2^2) + 2m_3].
$$

This, in turn implies that:

$$
\frac{\widehat{\sigma}_1^2}{\widehat{\sigma}_2^2} > \exp\left(\tfrac{2}{n}(m_3 - m_2)\right) \Rightarrow \left(\frac{\widehat{\sigma}_1^2 - \widehat{\sigma}_2^2}{\widehat{\sigma}_2^2}\right) > \exp\left(\tfrac{2}{n}(m_3 - m_2)\right) - 1.
$$

One can relate the AIC decision in favor of M_3 to a Neyman-Pearson rejection of the null in testing the hypotheses:

$$
H_0: \beta_3 = 0, \quad vs. \quad H_1: \beta_3 \neq 0.
$$

It can be shown that the t-test for this hypothesis takes the form:

$$
\tau(\mathbf{y}) = \frac{(\widehat{\beta}_3 - 0)}{\sqrt{\widehat{\mathbf{Var}(\widehat{\beta}_3)}}} = \sqrt{\frac{(n - m_3)(\widehat{\sigma}_1^2 - \widehat{\sigma}_2^2)}{\widehat{\sigma}_2^2}} \overset{H_0}{\sim} \mathsf{St}(n - m_3), \quad C_1 = \{|\tau(\mathbf{y})| > c_\alpha\},
$$

(see Spanos, 1986, p. 426) where C_1 denotes the rejection region and c_α the critical value associated with $\mathsf{St}(n - m_3)$; the Student's t distribution with $n - m_3$ degrees of freedom. This indicates that the AIC procedure amounts to rejecting H_0 when:

$$
\tau(\mathbf{y}) > c_\alpha = \sqrt{(n - m_3)\left(\exp\left(\tfrac{2}{n}\right) - 1\right)}.
$$

That is, the implicit critical value is c_α, with an unknown implicit type I error α. Depending on the values of (n, m_3) different critical values c_α and α are (implicitly) chosen, but the AIC procedure is oblivious to this implicit choice. For instance, in the above case where $n = 35$, $m_3 = 4$, the implicit $c_\alpha = 1.35$, and thus $\alpha = .187$. That is, the AIC choice of M_3 over M_2 amounts to applying a N-P test with an implicit type I error much higher than the traditional choices; this can easily give rise to unreliable

inferences if one is unaware of the actual error probabilities. More generally, if n is large relative to m_3 :

$$\sqrt{(n - m_3)\left(\exp\left(\tfrac{2}{n}\right) - 1\right)} \simeq \sqrt{2} = 1.414 \Rightarrow \alpha = .16$$

provides a reasonable approximation, with the implicit α being rather high.

The above discussion suggests that the AIC model selection procedure is ineffective in addressing the curve-fitting problem. *First*, when the choice of the original family of models in (a) is inappropriate (the associated statistical model is misspecified), the AIC procedure will lead to unreliable inferences with probability one. The only way to ensure the reliability of inference is to choose the statistical model which captures the regularities in the data by securing statistical adequacy. The process of ensuring statistical adequacy, however, solves both problems (a) choosing the family of curves, and (b) choosing the fittest within this family, in section 2, rendering the application of the AIC procedure (i) superfluous and (ii) potentially misleading. *Second*, when the choice of the original family of models in (a) is appropriate (the associated statistical model is adequate), the AIC procedure is still unreliable because it implicitly performs N-P testing with unknown error probabilities.

7   Conclusions

The current perspective dominating discussions on curve-fitting is that of mathematical approximation which, as argued above, provides an inadequate basis for reliable inductive inference. The mathematical convergence results provide no basis for ascertainable error probabilities to calibrate the reliability of inference. The Akaike-type model selection procedures provide an extension of the mathematical approximation perspective, by trading goodness-of-fit with parsimony, but they ignore the reliability of inference problem. Indeed, it is shown that these procedures give rise to misleading inferences even in the best case scenario where the 'true' model belongs to the pre-selected family of curves.

It is argued that a more satisfactory framework is provided by viewing curve-fitting as an empirical modeling problem in the context of the error-statistical approach. This approach embeds the approximation problem into a statistical model, and selects the 'fittest' curve to be a *statistically adequate* model: one which accounts for *the statistical 'regularities' in the data*. This ensures the reliability of inference associated with such a model because the nominal error probabilities are approximately equal to the actual error probabilities. Moreover, fittest curves, in the sense of being statistically adequate, turn out to be rare, contrary to the conventional wisdom concerning underdetermination. This argument is illustrated by comparing Kepler's law of motion of the planets with that of Ptolemy's on statistical adequacy grounds; Ptolemy's model is shown to be statistically inadequate.

References

[1] Akaike, H. (1973) "Information theory and an extension of the maximum likelihood principle," pp. 267-281 in B. N. Petrov and F. Csaki (ed.), *2nd International Symposium on Information Theory*, Akademia Kiado, Budapest.

[2] Babb, S. E. (1977) "Accuracy of Planetary Theories, Particularly for Mars," *Isis*, **68**: 426-434.

[3] Bohr, H. (1949) "On Almost periodic functions and the theory of groups," *The American Mathematical Monthly*, **56**: 595-609.

[4] Farebrother, R. W. (1999) *Fitting Linear Relationships: A History of the Calculus of Observations 1750-1900,* Springer-Verlag, New York.

[5] Fisher, R. A. (1935) *The Design of Experiments*, Oliver and Boyd, Edinburgh.

[6] Forster, M. and E. Sober (1994) "How to Tell when Simpler, More Unified, or Less Ad Hoc Theories will Provide More Accurate Predictions," *British Journal for the Philosophy of Science*, **45**: 1-35.

[7] Gauss, C. F.. (1809) *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*, F. Perthes and I. H. Besser, Humburg.

[8] Gauss, C. F.. (1823) "Theoria conbinationis observationum erroribus minimis obnoxiae, Pas prior, et Pars posterior," *Comm. Soc. Reg. Gottingensis Rec.* **5**, 33-62.

[9] Glymour, C. (1981) *Theory and Evidence*, Princeton University Press, NJ.

[10] Hahn, A. J. (1998), *Basic Calculus: From Archimedes to Newton to its Role in Science*, Springer, New York.

[11] Hald, A. (1998) *A History of Mathematical Statistics: From 1750 to 1930*, Wiley, N. Y.

[12] Hannan, E. J. and Quinn, B. G. (1979) "The determination of the order of an autoregression," *Journal of the Royal Statistical Society*, B, **41**, 190-195.

[13] Hildebrand, F. B. (1974) *Introduction to Numerical Analysis*, McGraw-Hill, NY.

[14] Hitchcock, C. and E. Sober (2004) "Prediction Versus Accommodation and Risk of Overfitting,' *British Journal for the Philosophy of Science*, **55**: 1-34.

[15] Hurvich, C. M. and C. L. Tsai (1989) "Regression and Time Series Model Selection in Small Samples," *Biometrika*, **76**: 297-307.

[16] Isaacson, E. and H. B. Keller (1994) *Analysis of Numerical Methods*, Dover, NY.

[17] Kieseppa, I. A. (1997) "Akaike Information Criterion, Curve-fitting, and the Philosophical Problem of simplicity," *British Journal for the Philosophy of Science*, **48**: 21-48.

[18] Kukla, A. (1995) "Forster and Sober on the Curve-Fitting Problem," *British Journal for the Philosophy of Science*, **46**: 248-252.

[19] Laudan, L. (1977), *Progress and Its Problems: Towards a Theory of Scientific Growth*, University of California Press, Berkeley, CA.

[20] Legendre, A. M. (1805) *Nouvelle Methodes pour la Determination des Orbites des Cometes*, Mme, Courcier, Paris.

[21] Lehmann, E. L. (1990) "Model specification: the views of Fisher and Neyman, and later developments", *Statistical Science*, **5**: 160-168.

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

[23] Mayo, D. G. and A. Spanos (2004) "Methodology in Practice: Statistical Misspecification Testing", *Philosophy of Science*, **71**: 1007-1025.

[24] Mayo, D. G. and A. Spanos (2006) "Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction," *The British Journal for the Philosophy of Science,* 57: 323-357, 2006.

[25] Mulaik, S. A. (2001) "The Curve-Fitting Problem: An Objectivist View," *Philosophy of Science*, **68**: 218-241.

[26] Schwarz, G. (1978) "Estimating the dimension of a model," *Annals of Statistics*, **6**: 461-464.

[27] Sober, E. (2000) "Simplicity" pp. 433-441 in *A companion to the Philosophy of Science*, Blackwell, Oxford, edited by of Newton-Smith, W. H.

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

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

[30] Spanos, A. (1999) *Probability Theory and Statistical Inference: econometric modeling with observational data*, Cambridge University Press, Cambridge.

[31] Spanos, A. (2002) "Parametric versus Non-parametric Inference: Statistical Models and Simplicity," ch. 11, pp. 181-206 in Simplicity, Inference and Modelling: Keeping it Sophisticatedly Simple, edited by A. Zellner, H. A. Keuzenkamp and M. McAleer, Cambridge University Press.

[32] Spanos, A. (2006a) "Econometrics in Retrospect and Prospect," pp. 3-58 in Mills, T.C. and K. Patterson, *New Palgrave Handbook of Econometrics*, vol. 1, MacMillan, London.

[33] Spanos, A. (2006b) "Revisiting the Omitted Variables Argument: Substantive vs. Statistical Adequacy," *Journal of Economic Methodology*, **13**: 179-218.

[34] Spanos, A. (2006c) "Curve-fitting vs. Empirical Modeling: Ptolemy vs. Kepler," Virginia Tech working paper.

[35] Spanos, A. and A. McGuirk (2001) "The Model Specification Problem from a Probabilistic Reduction Perspective," *Journal of the American Agricultural Association*, **83**: 1168-1176.