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

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

March 2006

Abstract

The curve-fitting problem is often viewed as an exemplar which encapsulates many of the problems associated with inductive inference, including underdetermination and the reliability of inference. The current 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 primary objective of this paper is twofold. *First*, to argue that the AIC-type procedures do not provide an adequate solution to the curve fitting problem because they have no criterion to assess when a curve *captures the regularities in the data* inadequately. As a result, these procedures cannot ensure the reliability of inductive inference. *Second,* to argue that for more satisfactory answers one needs to view the curve fitting problem in the context of error-statistical approach where *statistical adequacy* provides such a criterion and the associated error probabilities can be used to calibrate the trustworthiness of inductive inference.

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 problem of *underdetermination* and the *reliability of inference* issue. Glymour (1981) highlighted the importance of curve fitting and argued that the problem was not well understood in both its philosophical as well as mathematical dimensions. By the mid 1990s, however, the dominating view 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); despite the fact that Glymour (1981) argued persuasively that simplicity does not provide an adequate solution.

The question posed in this paper is the extent to which the AIC and related procedures provide a satisfactory solution to the original 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 when a curve *captures the 'regularities' in the data* inadequately. As a result, these procedures have no way to ensure the reliability of inductive inferences associated with the 'fittest' curve, and provide 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 statistical adequacy provides such a criterion and the associated error probabilities can be used to calibrate the trustworthiness 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 that dominates the current understanding of the problem, and underlies 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 problem in an error-statistical 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 inadequacy of choosing the fittest curve by trading goodness-of-fit against simplicity and the misleading impression concerning the pervasiveness of the problem of underdetermination. Section 6 discusses specific weaknesses of the AIC type procedures that give rise to unreliable inferences.

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 $y = g(x)$, that fits the existing *data:*

$$\{(x_k, y_k),\ k = 1, \ldots, n\} \tag{1}$$

'best', and approximates $h(x)$ well enough to be used to predict adequately 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, in

this context, because more than one curve *captures the regularities in the data equally well*.

Looked at from this (approximation) perspective the problem of curve-fitting is thought to comprise two stages:

(a) 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 $\{\phi_i(x),\ i=1,2,...\}$ are known functions, e.g. ordinary polynomials:

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

(b) the selection of the 'best' fitting curve (within the chosen family) using certain criteria, e.g. minimizing the sum of squares of the errors:

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

The basic problem with this perspective is that goodness of fit cannot be the sole criterion for 'best' because it 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 two 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. Their main argument is that (i) the selection of the best fitting curve in (b) 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. 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.

3 Curve fitting or statistical modeling?

This paper argues that AIC-type procedures do not provide a satisfactory solution to the curve fitting problem primarily because they do not address the crucial *adequacy* problem of 'when a fitted curve captures the regularities in the data'. Viewing the problem in terms of choices (a) and (b) 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. transform the curve-fitting into a choice of *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. 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

In general, the least-square approximation errors $\varepsilon_k(x_k,m)=y_k-\sum_{i=0}^{m}\alpha_i\phi_i(x_k)$, depend on both x_k and m in systematic ways. The result that the least-squares least-squares approximation gives rise to is that, under certain restrictions on the true function $y=h(x)$, the sum of squares of the errors goes to zero as m, the number of terms in the approximation, goes to infinity, i.e.

$$\lim_{m\to\infty}\sum_{k=1}^{n}(y_k-\sum_{i=0}^{m}\alpha_i\phi_i(x_k))^2=0. \tag{4}$$

That is, for every $\epsilon>0$ there exists a large enough integer $N(\epsilon)$ such that for $m>N(\epsilon)$,

$$|\varepsilon_k(x_k,m)|<\epsilon, \tag{5}$$

providing an upper bound on the magnitude of the error. For particular values of (n,m), however, the error term is a systematic function of both x_k and m; see Hildebrand (1974). Can one use the asymptotic result in (4) as a basis for inference? The answer is no because the results in (4) and (5) provide no way to assess the reliability of any inductive inference for given values of (n,m).

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 upper bound for the error in (5). For that one needs to impose additional probabilistic structure on the error term. Indeed, any attempt to strengthen the above *mean convergence* in (5) to, say, *uniform convergence*, i.e.

$$|\varepsilon_k(x_k,m)|<\epsilon,\quad \text{for all } x_k,\ k=1,\ldots \tag{6}$$

it would still not be good enough to use as a basis for inference because the mathematical approximation errors remain 'systematic' functions of x_k and m; see Hildebrand (1974). Neither type of convergence provide one with a way to evaluate the error probabilities associated with any inference, and thus the reliability of inference cannot be assessed for given values of (n, m), even approximately. This is because the mathematical approximation convergence cannot be used to evaluate error probabilities associated with inductive inferences to assess their reliability; note that the convergence is based on the degree of the polynomial going to infinity, i.e. $m \to \infty$.

In summary, the mathematical approximation method, using least squares will take one up to the estimation of such a curve, say:

$$\widehat{y}_t \quad = 167.115 + 1.907x_t, \quad s = 1.7714, \; T = 35, \tag{7}$$

(see section 6) but no inference can be constructed on (7) without probabilistic assumptions to provide one with inference procedures whose reliability can be assessed.

3.2 Gauss's statistical modeling perspective

Gauss's major contribution was to transform the approximation error term:

$$\varepsilon_k(x_k, m) = y_k - \textstyle\sum_{i=0}^{m} \alpha_i \phi_i(x_k), \tag{8}$$

into a generic statistical error:

$$\varepsilon_k(x_k, m) = \varepsilon_k \backsim \mathsf{NIID}(0, \sigma^2), \; k = 1, 2, \ldots, n, ..., \tag{9}$$

where $\mathsf{NIID}(0, \sigma^2)$ stands for 'Normal, Independent and Identically Distributed with mean 0 and variance σ^2'. The error in (9) 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):

$$y_k = \textstyle\sum_{i=0}^{m} \alpha_i \phi_i(x_k) + \varepsilon_k, \quad \varepsilon_k \backsim \mathsf{NIID}(0, \sigma^2), \; k = 1, 2, \ldots, n, ... \tag{10}$$

What makes his contribution all-important is that the statistical model in (10) provides the premises for assessing the reliability of inductive inference.

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 (a)-(b) in 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 $g_m(x; \boldsymbol{\alpha})$ (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. 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*; see Mayo (1996). As argued above, the error statistical approach can be traced back to Gauss (1809). In the next section we summarize some of the basic tenets of this approach.

4 A summary of the Error-Statistical framework

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) for further discussion.

4.1 Theory and 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 captured 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 via:

$$y_k = h^*(\mathbf{x}_k, \boldsymbol{\xi}_k), \quad k \in \mathbb{N}. \tag{11}$$

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 (11) into a *structural model* of the form:

$$y_k = h(\mathbf{x}_k; \boldsymbol{\phi}) + \epsilon(\mathbf{x}_k, \boldsymbol{\xi}_k), \quad k \in \mathbb{N}, \tag{12}$$

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

$$\{\epsilon(\mathbf{x}_k, \boldsymbol{\xi}_k) = y_k - h(\mathbf{x}_k; \boldsymbol{\phi}),\ k \in \mathbb{N}\}, \tag{13}$$

is the structural error term, viewed as a function of both $\mathbf{x}_k$ and $\boldsymbol{\xi}_k$, representing all unmodeled influences. For (13) 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:

$$\left.\begin{array}{ll} \text{(i)} & E(\epsilon(\mathbf{x}_k,\boldsymbol{\xi}_k))=0, \\ \text{(ii)} & E(\epsilon(\mathbf{x}_k,\boldsymbol{\xi}_k)\cdot\epsilon(\mathbf{x}_\ell,\boldsymbol{\xi}_\ell))=\left\{\begin{array}{ll} \sigma^2, & k=\ell \\ 0, & k\neq\ell \end{array}\right.,\ k,\ell\in\mathbb{N}, \\ \text{(iii)} & E(\epsilon(\mathbf{x}_k,\boldsymbol{\xi}_k)\cdot h(\mathbf{x}_k;\boldsymbol{\phi}))=0, \end{array}\right\}\ \forall(\mathbf{x}_k,\boldsymbol{\xi}_k)\in\mathbb{R}_{\mathbf{x}}\times\mathbb{R}_{\boldsymbol{\xi}}. \quad (14)$$

The above perspective on theory and structural models provides a much broader framework than that of mathematical approximation dominating the current discussions of curve fitting. In common with the mathematical approximation perspective, however, the error term is statistically non-opearational. Assumptions (i)-(iii) are empirically non-testable because their assessment involves 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. Not surprisingly, the nature of 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.

4.1.1 Experimental data

In the case where one can perform experiments, 'experimental design' techniques allow one to ensure that the *error term* is no longer a function of $(\mathbf{x}_k,\boldsymbol{\xi}_k)$, but, instead:

$$\epsilon(\mathbf{x}_k,\boldsymbol{\xi}_k)=\varepsilon_k\backsim\mathsf{IID}(0,\sigma^2),\quad \text{for all values } (\mathbf{x}_k,\boldsymbol{\xi}_k)\in\mathbb{R}_{\mathbf{x}}\times\mathbb{R}_{\boldsymbol{\xi}}\quad k=1,...,n. \quad (15)$$

For instance, *randomization* and *blocking* are often used to 'neutralize' and 'isolated' the phenomenon from the potential effects of $\boldsymbol{\xi}_k$ by ensuring that the uncontrolled factors cancel each other out; see Fisher (1935). As a direct result of the experimental 'control' via (15) the structural model (12) is essentially transformed into a *statistical model*:

$$y_k=h(\mathbf{x}_k;\boldsymbol{\theta})+\varepsilon_k,\quad \varepsilon_k\backsim\mathsf{IID}(0,\sigma^2),\ k=1,2,...,n, \quad (16)$$

where the statistical error term ε_k in (16) is qualitatively very different from the structural error term $\epsilon(\mathbf{x}_k,\boldsymbol{\xi}_k)$ in (12) because ε_k is no longer a function of $(\mathbf{x}_k,\boldsymbol{\xi}_k)$ and its assumptions are rendered empirically testable. For more precise inferences one needs to be more particular about the probabilistic assumptions defining the statistical model, including the functional form $h(.)$. This is because the more finical the probabilistic assumptions (the more constricting the statistical premises), the more precise the inferences.

A widely used special case of (16) is the Gauss Linear model, a simple form of which is given in table 1.

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.

4.1.2 Observational data

This is the case where the observed data on $(y_k, \mathbf{x}_k)$ are the result of an ongoing actual data generating process, undisturbed by any experimental control or intervention. In this case the experimental controls and interventions are no longer available and one needs a different way to secure the generic non-systematic nature of the error. It turns out that *conditioning* supplies the primary tool in dealing with modeling observational data.

As shown in Spanos (1986, 1999, 2006a-b), sequential conditioning provides a general way to transform an arbitrary stochastic process $\{\mathbf{Z}_t,\ t \in \mathbb{T}\}$ into a *martingale difference process*; a modern form of a white-noise process. This replaces the controls and interventions in experimental situations with the choice of the *relevant conditioning information set* $\mathfrak{D}_t$ that would render the error term non-systematic. The technical details of how one can specify a statistical model with observational data are beyond the scope of this paper (see Spanos, 2006a), but one can get some idea of what is involved by focusing on a particularly important statistical model.

The Normal/Linear Regression model results from a probabilistic reduction based on assuming that $\{\mathbf{Z}_t,\ t \in \mathbb{T}\}$ is a NIID vector process. These assumptions ensure that the error process $\{u_t,\ t \in \mathbb{T}\}$, defined by:

$$u_t = y_t - E(y_t|\mathfrak{D}_t), \tag{17}$$

where $\mathfrak{D}_t$=$\{\mathbf{X}_t = \mathbf{x}_t\}$, constitutes a non-systematic (martingale difference) process. The model itself can be viewed as a reduction from $D(\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_T; \boldsymbol{\phi})$ via the sequential conditioning:

$$D(\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_T; \boldsymbol{\phi}) \overset{\mathsf{I}}{=} \prod_{t=1}^{T} D_t(\mathbf{Z}_t, ; \boldsymbol{\psi}) \overset{\mathsf{IID}}{=} \prod_{t=1}^{T} D(y_t|\ \mathbf{X}_t; \boldsymbol{\psi}_1) \cdot D(\mathbf{X}_t; \boldsymbol{\psi}_2). \tag{18}$$

This reduction ensures that statistical error term u_t takes the form:

$$(u_t|\ \mathbf{X}_t = \mathbf{x}_t) \backsim \mathsf{NIID}(0, \sigma^2),\ k = 1, 2, ..., n. \tag{19}$$

This is analogous to (15) in the case of experimental data, but now the the error term has been operationalized by a judicious choice of the conditioning information set $\mathfrak{D}_t$. The complete specification of the Linear Regression model is given in table 2 where assumptions [**1**]-[**5**] assumptions pertain to the structure of the observable process $\{(y_t|\ \mathbf{X}_t{=}\ \mathbf{x}_t),\ t{\in}\mathbb{T}\}$.

Table 2 - The Normal/Linear Regression Model

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

[**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.

5 Curve fitting and the error-statistical perspective

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 because the primary criterion for 'best' becomes *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 choices (a) and (b) in section 2. Indeed, the probabilistic assumptions constituting the statistical model operationalize what one means by the 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:

$$\widehat{\varepsilon}_k = y_k - \textstyle\sum_{i=0}^m \widehat{\alpha}_i\phi_i(x_k), \quad k = 1, 2, \ldots, n, \tag{20}$$

are non-systematic. If there is systematic information in the residuals one respecifies the model, by allowing for specific departures such as non-Normality, nonlinearity, heteroskedasticity, etc., until the new model is statistically adequate for

data $\{(x_k, y_k),\ k{=}1, ..., n\}$. 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.

Empirical example. Kepler's empirical regularity in the form of the elliptical motion of Mars turned out to be a real regularity which persisted over time, not because an ellipse is simpler than a circle, but because fitting an ellipse renders the resulting statistical model statistically adequate. To see this, consider the structural model:

$$y_t = \alpha_0 + \alpha_1 x_t + \epsilon(x_k, \boldsymbol{\xi}_k),\ t \in \mathbb{T}, \tag{21}$$

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.* Embedding (21) into the Normal/Linear Regression model (table 2), and estimating it using **Kepler's original data** ($n = 28$) yield:

$$y_t = \underset{(.000002)}{0.662062} + \underset{(.000003)}{.061333} x_t + \widehat{u}_t,\ R^2 = .999,\ s = .0000111479. \tag{22}$$

The plot of the residuals in figure 1, as well as formal 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:	$F(2, 22) = 2.034[.155]$

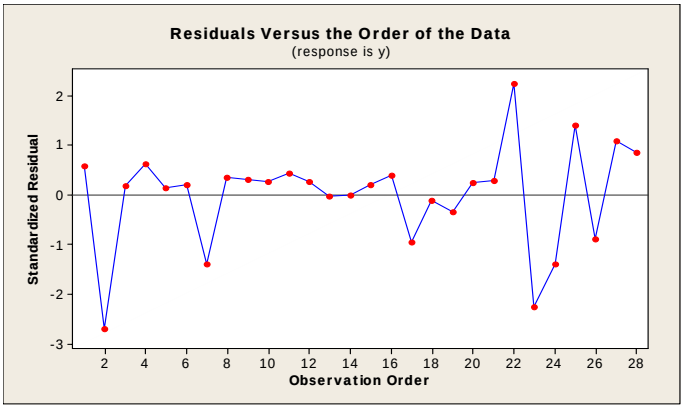

Fig. 1: Residuals from the Kepler regression

On the basis of the statistically adequate model in (22), we can test the substantive hypothesis that the motion is a circle 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 strong evidence against H_0.

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 extremely 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* (see Hilerbrand, 1974):

$$g_m(x;\boldsymbol{\alpha}) = \textstyle\sum_{i=0}^{m} y_i(\prod_{j=0,j\neq i} \frac{x-x_j}{x_i-x_j}), \tag{23}$$

ignores the fact that such a polynomial is useless 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 $(\widehat{\alpha}_0, \widehat{\alpha}_1, ..., \widehat{\alpha}_m)$.

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 $g_m(x;\widehat{\boldsymbol{\alpha}})$ 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 $g_m(x;\widehat{\boldsymbol{\alpha}})$ 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).

6 Problems with Akaike-type procedures

Where does this leave 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, and

(ii) asymptotically (as $n \to \infty$) the chosen m is not a consistent estimator of the true m^*.

On the basis of this criticism the AIC has been modified; see Schwartz (1978), Hannan and Quinn (1979). It is argued, however, that even the modified procedures are inadequate for addressing the curve fitting problem for a variety of reasons which will be summarized below.

In general, the Akaike Information criterion is defined by (see Akaike, 1973):

$$\mathsf{AIC} = \text{-}2\ln(\text{estimated likelihood}) + 2(\text{number of parameters}). \tag{24}$$

In the above case of the Gauss Linear model, the log-likelihood is:

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

giving rise to:

$$\mathsf{AIC}(m) \quad = \text{const.} + n\ln(\widehat{\sigma}^2) + 2m, \quad \widehat{\sigma}^2 = \tfrac{1}{n}\sum_{k=1}^{n}\left(y_k - \sum_{i=0}^{m}\widehat{\alpha}_i\phi_i(x_k)\right)^2. \tag{25}$$

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 $n\ln(\widehat{\sigma}^2)$ with parsimony $2m$, taking the likelihood function as given. This, however, means that the AIC procedure ignores the statistical adequacy issue because the likelihood function presupposes 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:

$$M_1: \quad y_t \underset{(.610)}{=} 167.115 + \underset{(.024)}{1.907}x_t + \widehat{u}_t, \quad R^2 = .995, \quad s = 1.7714, \; T = 35, \tag{26}$$

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:

$$M(m): \quad y_k = \sum_{i=0}^{m}\alpha_i\phi_i(x_k) + \varepsilon_k, \tag{27}$$

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 $m = 4$. Note 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 (27)

Model	$\mathsf{AIC(m)} = n\ln(\widehat{\sigma}^2) + 2m,$	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
$\mathsf{AIC}(4) =$	$(35)\ln(1.8658) + 2(6) = 33.829$	1
$\mathsf{AIC}(5) =$	$(35)\ln(1.8018) + 2(7) = 34.608$	2

(28)

As shown in Mayo and Spanos (2004), the original model (26), as well as the models in (27), are statistically misspecified. However, the AIC procedure ignores statistical adequacy and (fallaciously) chooses model $M(4)$ within a family of completely 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:

$$M(k,\ell): \quad y_t = \beta_0 + \beta_1 x_t + \sum_{i=1}^{k} \delta_i t + \sum_{i=1}^{\ell} [a_i y_{t-i} + \gamma_i x_{t-i}] + \varepsilon_t, \tag{29}$$

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 'correct' 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. Let us explore this question using the above data.

Empirical example - continued. Applying AIC criterion to select a model from the $M(k,\ell)$ family in (29) 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*.

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 $M(3,2)$, 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.

Table 5 - Akaike model selection from (29)

Model	$\mathsf{AIC(m)} = n\ln(\widehat{\sigma}^2) + 2m,$	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
$M(3,2):$	$(35)\ln(.029651) + 2(10) = -103.14$	1
$M(3,3):$	$(35)\ln(.026709) + 2(12) = -102.80$	2

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:

$$\begin{array}{ll} M_3: & y_t=\beta_0+\beta_1x_t+\beta_2x_t^2+\beta_3x_t^3+\varepsilon_t, \quad m_3=4, \\ M_2: & y_t=\alpha_0+\alpha_1x_t+\alpha_2x_t^2+u_t, \qquad\qquad\quad m_2=3. \end{array}$$

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:

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

One can easily relate the AIC decision in favor of M_3 to rejecting the null in of a test of the hypotheses:

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

It can be shown that the t-test for this hypothesis takes the form (see Spanos, 1986, p. 426):

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

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 suggests that the AIC procedure amounts to rejecting H_0 when:

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

In cases where n is large relative to m_3 :

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

That is, the AIC choice of M_3 over M_2 amounts to applying a N-P test with a critical value of $c_\alpha=1.414$ which corresponds to a significance level of $\alpha=.16$; this is much higher than the traditional choices of a type I error and it can easily give rise to unreliable inferences if one does not know the actual error probability.

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) and (b) 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 effectively preforms N-P testing with unknown error probabilities.

7 Conclusions

The current perspective dominating the discussions on curve fitting is one of mathematical approximation which provides an inadequate basis for reliable inductive inference. The mathematical convergence results provide no satisfactory basis for ascertainable error probabilities to calibrate the trustworthiness of inference procedures. The Akaike-type model selection procedures provide an extension of the mathematical approximation perspective which does not address the reliability of inference problem. Indeed, it is shown to give rise to misleading inferences even in the best case scenario where the true model belongs to the chosen family of curves.

It is argued that a more satisfactory framework for inductive inference is provided by viewing curve fitting as an empirical modeling problem in the context of the error-statistical approach. This approach embeds the true relationship $y = h(x)$ into a statistical model, and chooses the 'fittest' curve to be a statistically adequate model: one which accounts for all the systematic information 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. The learning from data about the phenomenon of interest is achieved by using trustworthy methods to draw inductive inferences; see Mayo (1996).

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

[3] 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.

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

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

[6] 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.

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

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

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

[10] 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.

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

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

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

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

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

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

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

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

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

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

[21] 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.

[22] Spanos, A. (2006b) "Revisiting the Omitted Variables Argument: Substantive vs. Statistical Reliability of Inference," forthcoming in the *Journal of Economic Methodology*.

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