Akaike-type Criteria and the Reliability of Inference: Model Selection vs. Statistical Model Specification

Aris Spanos*

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

December 2008

Abstract

Over the last two decades or so, the Akaike Information Criterion (AIC) and its various modifications/extensions have found wide applicability in econometrics as objective procedures which can be used to select parsimonious statistical models. The aim of this paper is to argue that these model selection procedures invariably give rise to unreliable inferences primarily because their selection within a prespecified family of models, (a) assumes away the problem of model validation, and (b) ignores the relevant error probabilities.

The paper argues for a return to the original *statistical model specification* problem, as envisaged by Fisher (1922), where the task is understood as one of selecting a statistical model in such a way so as to render the particular data a *truly typical realization* of the stochastic process underlying the model in question. This problem can be addressed by evaluating a statistical model in terms of its *statistical adequacy,* i.e. whether it accounts for the chance regularities in the data, as opposed to trading goodness-of-fit against parsimony.

*Thanks are due to Steven Durlauf, Andros Kourtellos, Nikitas Pittis and two anonymous referees for their constructive comments and suggestions that helped to improve the paper substantially.

1 Introduction

R. A. Fisher (1922) pioneered modern frequentist statistics as a model-based approach to statistical induction anchored on the notion of a *statistical model*, formalized by:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})=\{f(\mathbf{z};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta\},\ \mathbf{z}\in\mathbb{R}^n_Z,\ \Theta\subset\mathbb{R}^m,\ m<n. \tag{1}$$

Unlike Pearson (1920), who would proceed from data $\mathbf{z}_0$:=$(z_1,...,z_n)$ in search of a frequency curve $f(z;\boldsymbol{\vartheta})$ to describe its histogram, Fisher proposed to begin with a prespecified $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ (a 'hypothetical infinite population'), and view $\mathbf{z}_0$ as a realization thereof. He envisaged the specification of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ as a response to the question: "Of what population is this a random sample?" (ibid., p. 313), underscoring that: "the adequacy of our choice may be tested a posteriori." (p. 314) He identified the 'problems of statistics' to be: (1) specification, (2) estimation and (3) distribution and emphasized that addressing (2)-(3) depended crucially on dealing with (1) successfully first. A misspecified $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ would vitiate any procedure relying on $f(\mathbf{z};\boldsymbol{\theta}),\ \mathbf{z}\in\mathbb{R}^n_Z$, (or the likelihood function), and render all inductive inferences unreliable by giving rise to *actual error probabilities* that are invariably different from the *nominal* ones.

Despite its manifest importance, specification received only scant attention in the statistics literature since the 1920s (Spanos, 2006a). Although the formal apparatus of frequentist statistical inference was largely in place by the late 1930s, the nature of the underlying *inductive reasoning* was clouded by disagreements. Fisher argued for 'inductive inference' spearheaded by his significance testing (Fisher, 1955), and Neyman argued for 'inductive behaviour' based on N–P testing (Neyman, 1956). However, neither account gave a satisfactory answer to the canonical question:

when do data $\mathbf{z}_0$ provide evidence for (or against) a substantive claim H?

Their testing procedures were plagued by *the fallacy of acceptance* [*no* evidence against the null is misinterpreted as evidence *for* it], and *the fallacy of rejection* [evidence *against* the null is misinterpreted as evidence *for* the alternative]. Indeed, several crucial foundational problems reverberate largely unresolved to this day:

(I) how could a (possibly) infinite set $\mathcal{P}(\mathbf{z})$, of all possible models that could have given rise to data $\mathbf{z}_0$, be narrowed down to a single statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$?

(II) how could the adequacy of a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ be tested *a posteriori*?

(III) what is the role of *substantive* information in statistical modeling?

(IV) what is the role of pre-data vs. post-data error probabilities? (Hacking, 1965),

(V) how could the fallacies of acceptance and rejection be addressed in practice?

These issues created endless confusions in the minds of practitioners concerning the appropriate use and interpretation of the frequentist approach to inference.

Over the last three decades, Fisher's specification problem has been recast in the form of *model selection* where questions (IV)-(V) are ignored, but (I)-(III) are dealt with in specific ways. Question (I) is handled by separating the problem into two stages where, a broader family of models $\{\mathcal{M}_{\boldsymbol{\varphi}_i}(\mathbf{z}),\ i=1,...m\}$ is selected first, and then a best model $\mathcal{M}_{\boldsymbol{\varphi}_k}(\mathbf{z})$ within this family is chosen. The problem raised in (II) is treated by trading goodness-of-fit against parsimony, and the issue in (III) is often handled by using substantive information (including mathematical approximation

theory) in selecting the family of models. The quintessential example of such a procedure is based on the *Akaike Information Criterion* (AIC) and variations/extensions thereof, such as the Bayesian (BIC), the Schwarz (SIC), the Hannan-Qinn (HQIC) and the Minimum Description Length (MDL), as well as Cross-Validation criteria; Rao and Wu (2001), Burnham and Anderson (2002), Konishi and Kitagawa (2008).

These Akaike-type procedures are widely used in econometrics, and other applied disciplines, as offering objective methods for selecting parsimonious models (see Greene, 2003). In philosophy of science they are viewed as providing a pertinent way to address the curve fitting problem; see Forster and Sober (1994), Kieseppa (1997).

The primary objective of this paper is to make a case that the Akaike-type model selection procedures invariably give rise to unreliable inferences because:

(a) they ignore the preliminary step of validating the prespecified family of models,

(b) their selection amounts to testing comparisons among the models within the prespecified family but without 'controlling' the relevant error probabilities.

To deal with problems (a)-(b) and address questions (I)-(V) the paper argues for a return to the original statistical model specification problem, as envisaged by Fisher (1922), but view it in the context of a modification/extension of the original frequentist framework, known as *error statistics* (see Mayo, 1996), which emphasizes the probing of the different ways an inference might be in error. A key difference is that error statistics pays due attention to potential errors at the two points of nexus between statistical inference and the real-world phenomenon of interest:

[A] from the phenomenon of interest to an adequate statistical model,

[B] from inference results to evidence for or against substantive claims.

In the error statistical framework, Fisher's specification is carried out by distinguishing, *ab initio*, between *substantive* and *statistical* information and devising a purely probabilistic construal of a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ by viewing it as a parameterization of the stochastic process $\{\mathbf{Z}_k, k\in\mathbb{N}:=(1,...,n,...)\}$ whose probabilistic structure is chosen so as to render data $\mathbf{z}_0$ a *truly typical realization* thereof; see Spanos (1986). The specification of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ in $\mathcal{P}(\mathbf{z})$ is guided solely by *statistical adequacy*: the probabilistic assumptions making up the model are valid for data $\mathbf{z}_0$. Securing the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ enables one to deal with problems (a)-(b), and address questions (I)-(V) by employing ascertainable error probabilities (pre-data and post-data) to evaluate the reliability and pertinence of inductive inferences, including the inferential appraisal of substantive claims; see Mayo and Spanos (2006).

Section 2 discusses briefly the curve-fitting problem with a view to bring out the main features of the mathematical approximation perspective that motivates and underlays the Akaike-type selection procedures and most nonparametric methods in econometrics. The error statistical perspective is used to argue that undue reliance on mathematical approximation often undermines the reliability of inference because it relies on non-testable premises that often ignore the regularities in the data. In section 3, the Akaike-type selection procedures are discussed and their unreliability illustrated using empirical examples. Section 4 discusses the general question of rec-

onciling statistical and substantive information by embedding a structural model into a validated statistical model. Section 5 contrasts the Akaike-type selection procedures with securing statistical adequacy using thorough Mis-Specification (M-S) testing and respecification, and section 5 discusses certain charges leveled against M-S testing, like double-use of data, infinite regress, circularity and pre-test bias.

2 Mathematical approximation and curve fitting

The curve-fitting problem is often viewed in both statistics and philosophy of science as an exemplar that encapsulates the multitude of dimensions and problems associated with *inductive inference* (learning from data), including *underdetermination* – more than one curve can 'account for the regularities in the data' equally well – and the *reliability* of inference, which is often assumed to depend on a priori stipulations, such as the uniformity of nature; see Skyrms (2000).

2.1 The curve-fitting problem: a brief summary

In its simplest form *the curve-fitting problem* has three basic components:

(i) The existence of a 'true' but *unknown* function**:** $y=h(x),\ (x,y)\in\mathbb{R}_X\times\mathbb{R}_Y$.

(ii) A *data* set in the form of n observations: $\mathbf{z}_0:=\{(x_k,y_k),\ k=0,1,\ldots,n\}$.

(iii) Seeking an approximating function $g_m(x)$ of $h(x)$ over $\mathbf{z}_0$ that is 'best' in the sense that it 'accounts for the regularities in $\mathbf{z}_0$'; see Skyrms (2000).

Example 1. In an attempt to motivate some of mathematical apparatus needed to address this problem, consider a familiar example where $g_m(x)$ is chosen to be:

$$g_m(x;\boldsymbol{\alpha})=\alpha_0+\textstyle\sum_{k=1}^{m}\alpha_i x^k, \tag{2}$$

i.e. an ordinary polynomial of degree m, and 'best' is selected via the minimization:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^m}:\ell(\boldsymbol{\alpha})=\textstyle\sum_{k=1}^{n}\left(y_k-\alpha_0-\sum_{k=1}^{m}\alpha_i x^k\right)^2, \tag{3}$$

giving rise to the Least-Squares estimator $\widehat{\boldsymbol{\alpha}}_{LS}=(\widehat{\alpha}_0,..,\widehat{\alpha}_m)$ of $\boldsymbol{\alpha}=(\alpha_0,..,\alpha_m)$.

In this example a number of decisions (choices) were made that need to be formally (mathematically) justified in light of the information in (i)-(iii) above:

[a] What guarantees the existence and uniqueness of the 'best' fitted curve $g_m(x;\widehat{\boldsymbol{\alpha}}_{LS})$?

[b] Why choose an ordinary polynomial for $g_m(x;\boldsymbol{\alpha})$, and not some other function(s)?

[c] Why minimize the sum of squares in (3), and not some other objective function?

2.2 The mathematical approximation perspective

Mathematical approximation theory, which began with a profound theorem by Weierstrass in 1885 (see Powell, 1981), is concerned with how *unknown* functions like $h(x)$ can best be approximated with simpler *known* functions $\{\phi_i(x),\ i=1,2,...\}$, and with *quantitatively* characterizing the errors introduced thereby. This theory proposes specific answers to questions [a]-[c] stemming from three interrelated premises:

(a) the structure and 'smoothness' of $y=h(x),\ (x,y)\in\mathbb{R}_X\times\mathbb{R}_Y$ (continuity, etc.),

(b) a family $\mathcal{G}$ of approximating functions $\{\phi_i(x),i=1,...\}$ used as building blocks,

(c) a concept of 'distance' determining the notion of a 'best' approximation.

Mathematically the most effective way to render the curve-fitting tractable is to view the problem in the context of a *normed linear space* (see Powell, 1981), such as the set of all continuous functions over the interval $[a,b] \subset \mathbb{R}$, denoted by $(C[a,b], \| \, . \, \|_p)$, where the p-norm is defined by:

$$\| \, f \, \|_p = (\int_a^b |f(x)|^p \, dx)^{\frac{1}{p}}, \; p \geq 1, \text{ for all } f \in C[a,b]. \tag{4}$$

Interesting special cases are $p=1,2$ and $p=\infty$: $\| \, f \, \|_\infty = \max_{a \leq f \leq b} | \, f \, |$, $f \in C[a,b]$.

It can be argued that the traditional view of curve-fitting is largely the result of imposing a *mathematical approximation* perspective on the problem. Hence, it's no accident that the approximating function often takes the linearized form:

$$g_m(x;\boldsymbol{\alpha}) = \alpha_0 + \sum_{i=1}^m \alpha_i \phi_i(x), \tag{5}$$

where $\{\phi_i(x), \; i=1,...m\}$ is a *base set* of generalized polynomials defined on $[a,b]$; see Appendix for a summary of the basic results on mathematical approximation.

It is important to note that when the approximation is over a *net* of points $\mathbb{G}_n(\mathbf{x}) := \{x_k, \; k=1,\ldots,n\}$, the results in the appendix need to be modified, because the relevant metric is associated with the *discrete* p-norm:

$$\| \, h(x_k) - g_m(x_k;\boldsymbol{\alpha}) \, \|_p = (\sum_{k=1}^n w(x_k) \, |h(x_k) - g_m(x_k;\boldsymbol{\alpha})|^p)^{\frac{1}{p}}, \quad p \geq 1, \tag{6}$$

where $w(x_k) > 0$, $\sum_{k=0}^n w(x_k) = 1$, denotes the weight function. This is commonly justified on the basis that the net $\mathbb{G}_n(\mathbf{x})$ is *dense* in the interval $[a,b]$, in the sense that as $n \to \infty$ the discrete metric converges to its continuous analogue (see Rivlin, 1981):

$$\lim_{n \to \infty} (\sum_{k=1}^n w(x_k) \, |h(x_k) - g_m(x_k;\boldsymbol{\alpha})|^p)^{\frac{1}{p}} = (\int_a^b w(x) \, |h(x) - g_m(x;\boldsymbol{\alpha})|^p \, dx)^{\frac{1}{p}}, \; p \geq 1.$$

Example 2. Let $y=h(x)$, $x \in \mathbb{R}$, $y \in \mathbb{R}$, be a square integrable function defined on $\mathbb{R} := (-\infty, \infty)$, i.e. $\int_{-\infty}^{\infty} |h(x)|^2 \, dx < \infty$, and denoted by $h(x) \in L_2(-\infty, \infty)$. Mathematical approximation theory asserts that the *Hermite orthogonal polynomials:*

$$\left\{ h_0(x)=1, \; h_1(x)=2x, \; h_2(x)=4x^2-2, \; h_3(x)=x^3-3x, ..., h_m(x)=(-1)^m e^{x^2} \frac{d^m(e^{-x^2})}{dx^m} \right\},$$

provide a *complete* base set for the normed linear (Hilbert) space $(L_2(-\infty,\infty), \| \, . \, \|_2)$, and thus, for the approximating function $g_m(x_k;\boldsymbol{\alpha}) = \sum_{k=0}^m \alpha_i h_i(x_k)$, the 2-norm ensures both uniqueness and convergence in mean to $h(x)$; see Luenberger (1969). Hence, one can estimate $\boldsymbol{\alpha}$ using Weighted Least Squares (Hildebrand, 1982):

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^m} : \ell(\boldsymbol{\alpha}) = \sum_{k=1}^n e^{-x_k^2} \left(y_k - g_m(x_k;\boldsymbol{\alpha})\right)^2, \tag{7}$$

giving rise to the *residuals*: $\widehat{\varepsilon}(x_k,m) = y_k - \sum_{k=0}^m \widehat{\alpha}_i h_i(x_k)$, $k=1,2,...,n$, where:

$$\widehat{\boldsymbol{\alpha}}_{WLS} = (\widehat{\alpha}_0, \widehat{\alpha}_1, ...\widehat{\alpha}_m), \; \widehat{\alpha}_\ell = (\tfrac{1}{2^\ell \ell! \sqrt{\pi}}) \sum_{k=1}^n [e^{-x_k^2}] h_\ell(x_k) y_k, \; , \; \ell = 0,1,...,m. \tag{8}$$

2.3 Mathematical approximation vs. inductive inference

2.3.1 Goodness-of-fit vs. parsimony: is that the issue?

The curve-fitting problem discussions are dominated by the mathematical approximation perspective to such an extent that the primary problem in selecting (5) is

thought to be *overfitting*, stemming from the fact that one can make the error in (3) as small as desired by increasing m. Indeed, the argument goes, one can make the approximation error $\varepsilon(x_k;m)=y_k-\sum_{i=0}^{m}\alpha_i\phi_i(x_k)$ equal to *zero* by choosing $m=n-1$; see Skyrms (2000). This reasoning suggests that goodness-of-fit cannot be the sole criterion for 'best'. To avoid *overfitting* one needs to supplement goodness-of-fit with pragmatic criteria such as *simplicity,* which can be justified on prediction grounds, since simpler curves enjoy better predictive accuracy; see Forster and Sober (1994).

A closer look at the above argument reveals that 'trading goodness-of-fit against parsimony' ignores two potential *unreliability* of inference problems when the fitted curve $g_m(x_k;\widehat{\boldsymbol{\alpha}})=\sum_{i=0}^{m}\widehat{\alpha}_i\phi_i(x_k)$ is used as a basis for inductive inference.

- The *first* relates to an inherent tension between the sample size n (statistical information) and the degree of approximation m of $g_m(x_k;\boldsymbol{\alpha})$ (section 2.3.2).
- The *second* has to do with the notion of assessing when 'a fitted curve accounts for the regularities in data $\mathbf{z}_0$' (section 2.3.3).

2.3.2 The inherent tension between $m\to\infty$ and $n\to\infty$

The tension between n and m can be brought out by scrutinizing the soundness of the claim that the choice $m=n-1$ renders the approximation error zero. Although this claim is mathematically sound, it is inferentially fallacious because the estimated coefficients $\widehat{\boldsymbol{\alpha}}:=(\widehat{\alpha}_0,...,\widehat{\alpha}_m)$ will be *inconsistent* estimators of $\boldsymbol{\alpha}:=(\alpha_0,...,\alpha_m)$, and thus, any inference based on $g_m(x_k;\widehat{\boldsymbol{\alpha}})$ will be totally unreliable.

Indeed, this highlights an inherent tension between *mathematical convergence* results that concern $m\to\infty$, and *probabilistic convergence* results concerned with $n\to\infty$. The former enhances fit by making the approximation error smaller, but consistency (convergence in probability), $\widehat{\boldsymbol{\alpha}}\xrightarrow{\mathbb{P}}\boldsymbol{\alpha}$ as $n\to\infty$, is necessary (but not sufficient) for the reliability of any form of inductive inference; see Spanos (2007a).

2.3.3 The qualitative characterization of the approximation error

When does a fitted curve $g_m(x_k;\widehat{\boldsymbol{\alpha}})$ account for the regularities in data $\mathbf{z}_0$? Intuition suggests that $g_m(x_k;\widehat{\boldsymbol{\alpha}})$ has captured the systematic information in $\mathbf{z}_0$ when what is left, the residuals $\widehat{\varepsilon}(x_k;m)$, $k=1,2,...,n$, constitute *white-noise*. More formally, when the approximation error $\varepsilon(x_k;m)$ is *non-systematic* in a probabilistic sense:

$$\text{[i]}\ \varepsilon_k(x_k,m)\backsim \mathsf{IID}(0,\sigma^2),\quad \text{[ii]}\ E\left[\varepsilon_k(x_k,m)\cdot g_m(x_k;\boldsymbol{\alpha})\right]=0,\quad \forall\,(x_k,k)\in\mathbb{R}_{\mathbf{x}}\times\mathbb{N},\qquad (9)$$

where 'IID' stands for 'Independent and Identically Distributed'. A closer look at the qualitative characterization of $\varepsilon(x_k;m)$ reveals that being 'small' in a mathematical sense (see (56) in the Appendix), does not entail (9), and neither does trading goodness-of-fit against overfitting. Indeed, reflecting on the results of mathematical approximation theory, suggests that conditions [i]-[ii] will often be invalid since the qualitative characterization of $\varepsilon(x_k;m)$ renders it a function of k, x and m. This can potentially devastate the reliability of any inference based on $g_m(x_k;\widehat{\boldsymbol{\alpha}})$.

To be more specific, the potential violation of conditions [i]-[ii] stems from the very mathematical results that specify *necessary* and *sufficient conditions* (iff) for

the existence and uniqueness of the best approximating function $g_m(x;\widehat{\boldsymbol{\alpha}})$, known as *oscillation-type theorems*; see Powell (1981). According to theorem 3 in the Appendix, the *iff* conditions require the *residuals* $\widehat{\varepsilon}(x_k;m)=h(x_k)-g_m(x_k;\widehat{\boldsymbol{\alpha}}),\ k=1,...,n$ to *alternate in sign* at least $m+2$ times over $\mathbb{G}_n(\mathbf{x})$. Analogous oscillation theorems are needed, not only for approximations based on different norms, but also for local (piecewise) polynomial approximation using splines (see Watson, 1980).

This potential unreliability problem can be brought out using a simple runs test for *randomness* (IID), which states that under IID the expected number of runs R is approximately $E(R)=(2n-1)/3$; a *run* is a sub-sequence of one type (+ or −) immediately preceded and succeeded by an element of the other type. For $n=100,\ E(R)\simeq 66$, the expected number of runs in the residuals should be considerably higher than $(m+2)$, otherwise they are likely to exhibit Markov dependence; see Spanos (1999).

This problem is illustrated in Spanos (2007a) by comparing the estimated Kepler's first law of planetary motion (statistically adequate) with Ptolemy's model (statistically inadequate), shown to represent a quintessential form of curve fitting whose residuals exhibit strong temporal dependence/heterogeneity.

2.3.4 The missing inductive understructure

Example 3. Consider fitting (by least squares) the line $y=\alpha_0+\alpha_1 x$, suggested by a certain theory (see Mayo and Spanos, 2004), through the scatter-plot of data $\mathbf{z}_0$:

$$\widehat{y}_t \ = 167.115+1.907x_t, \quad s=1.77,\ n=35. \tag{10}$$

The end result is that, as it stands, (10) provides *no basis* for inductive inference (learning from data). At best, mathematical approximation can provide very crude *Jackson-type* upper bounds (see Appendix) in terms of m (the degree of $g_m(x_k;\boldsymbol{\alpha})$). The residuals $\widehat{\varepsilon}_k=y_k-\widehat{\alpha}_0-\widehat{\alpha}_1x_k$ can be used to construct goodness-of-fit measures like:

$$R^2=1-\left[\textstyle\sum_{k=1}^n \widehat{\varepsilon}_k^2/\sum_{k=1}^n(y_k-\overline{y})^2\right]. \tag{11}$$

However, this framework does not provide the necessary understructure for inductive inference, i.e. it does not delineate the *probabilistic premises* stating the conditions under which the statistics $(\widehat{\alpha}_0,\widehat{\alpha}_1,s^2,R^2)$ are inferentially reliable, as opposed to mathematically justifiable or theoretically meaningful.

The above discussion brings out a *crucial weaknesses* of the mathematical approximation perspective: its answers to questions [a]-[c] (section 2.1) rely on premises (a)-(c) (section 2.2) that are largely *non-testable* and often ignore the regularities in data $\mathbf{z}_0$. This can potentially undermine the reliability of all inductive inference methods that rely on this perspective. Indeed, the non-testability of the invoked mathematical premises can severely undermine the reliability of many *nonparametric/semiparametric* methods that are misleadingly appealed to as a way to circumvent the statistical misspecification problem; see Li and Racine (2006).

The question that naturally arises is whether there is way to relate the mathematical premises (a)-(c) (section 2.2) to the regularities in data $\mathbf{z}_0$ that would ensure the validity of assumptions [i]-[ii]. Such a connection is provided by embedding the

mathematical approximation problem into a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ whose probabilistic assumptions are chosen with a view to account for the regularities in data $\mathbf{z}_0$. This embedding offers a way to transform the non-testable mathematical premises (a)-(c) into the *testable* [vis-a-vis $\mathbf{z}_0$] inductive premises of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$; see section 4.

3 Model selection in statistics

3.1 Gauss's pioneering contribution

Historically, Gauss (1809) should be credited with the first attempt to *embed* the mathematical approximation formulation into a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$, by making explicit *probabilistic assumptions* pertaining to a generic error term ε_k, specifying the **Gauss Linear model** (see Spanos, 1986, ch. 18):

$$y_k = \beta_0 + \textstyle\sum_{i=1}^{m} \beta_i x_{ik} + \varepsilon_k, \quad \varepsilon_k \backsim \mathsf{NIID}(0, \sigma^2), \ k=1,2,\ldots,n,... \tag{12}$$

What makes his contribution all-important is that (12) provides the inductive premises for assessing the reliability of inference based on the fitted model $\widehat{y}_k=\widehat{\beta}_0+\sum_{i=1}^{m}\widehat{\beta}_i x_{ik}$.

Example 3 (continued). Gauss's error assumptions in (12) provide specific premises for inductive inference, in the sense that the estimated model:

$$y_k \underset{(.610)}{= 167.115} + \underset{(.024)}{1.907} x_k + \underset{(1.771)}{\widehat{u}_k}, \ R^2=.995, \ s=1.77, \ n=35, \tag{13}$$

can be used to generate inferential statistics, such as the standard errors (SE) given in parentheses, to draw inferences, including testing the significance of the coefficients (β_0, β_1). For instance, if one were to take these SE at face value, the t-ratios $\tau(\widehat{\beta}_0)$=273.96[.000], $\tau(\widehat{\beta}_1)$=79.458[.000] (p-values in square brackets), would suggest that both coefficients are significantly different from zero. As argued below, however, such inferences turn out to be untrustworthy when the presumed probabilistic structure in (12) is invalid for data $\mathbf{z}_0=(y_1,...,y_n;x_1,...,x_n)$.

In general, the appended probabilistic structure raises two crucial questions:

(A) How does the approximation error $\varepsilon(x_k;m)$ *relate* to the statistical error ε_k?

(B) How does one *validate* the probabilistic assumptions of the error ε_k?

These questions are crucial when reconciling substantive and statistical information with a view to ensure the reliability of inductive inference; see section 4.

To illustrate the inbuilt connection between the mathematical (a)-(c) (section 2.2) and inductive premises, consider the choice of a norm $\|\ .\ \|$ and the distributional assumption underlying a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$.

3.2 The choice of a norm and distributional assumptions

The majority of statistical models motivated by the mathematical approximation perspective revolve around a generalized form of (12):

$$y_k = \alpha_0 + \textstyle\sum_{k=1}^{m} \alpha_i \phi_i(x_k) + \varepsilon_k, \quad \varepsilon_k \backsim \mathsf{NIID}(0, \sigma^2), \ k=1,2,\ldots,n,..., \tag{14}$$

where $g_m(x_k;\boldsymbol{\alpha})=\alpha_0+\sum_{k=1}^{m}\alpha_i\phi_i(x_k)$ is chosen using substantive information relating to $y=h(x)$, $(x,y)\in[\mathbb{R}_X\times\mathbb{R}_Y]$. In this case, the assumed distribution is Normal and the norm implicitly used is $\|\ .\ \|_2$, inducing the *inner product*:

$$\langle h(x_k)-g_m(x_k;\boldsymbol{\alpha})\rangle_2 = \textstyle\sum_{k=1}^{n}\left(y_k-\alpha_0-\sum_{k=1}^{m}\alpha_i\phi_i(x_k)\right)^2, \tag{15}$$

and *convergence in mean*, giving rise the Least-Squares estimator $\widehat{\boldsymbol{\alpha}}_{LS}$ of $\boldsymbol{\alpha}$. Indeed, it is no accident that in example 2 a natural base set for $L_2(-\infty,\infty)$ is provided by the Hermite polynomials whose weight function for orthogonality $w(x)=e^{-x^2}$ is in essence [by rescaling] the standard Normal density $\phi(x)$, i.e. the Rodrigues formula for $h_m(x)$ can be written equivalently as $h_m^*(x)=\frac{(-1)^m}{\phi(x)}\frac{d^m\phi(x)}{dx^m}$, $m=0,1,....$

What is often not made explicit in the statistics literature is the connection between other p-norms and the implicit distributional assumption. The 1-norm $\parallel h(x_k)-g_m(x_k;\boldsymbol{\alpha})\parallel_1$, inducing the *inner product* (Powell, 1981):

$$\langle h(x_k)-g_m(x_k;\boldsymbol{\alpha})\rangle_1=\textstyle\sum_{k=1}^n|y_k-\alpha_0-\sum_{k=1}^m\alpha_i\phi_i(x_k)|\,, \tag{16}$$

and *pointwise convergence*, giving rise to the *Least Absolute Deviation* estimator of $\boldsymbol{\alpha}$ (see Shao, 2003), is related to the *Laplace* distribution, $f(\varepsilon_k)=(\frac{1}{2\sigma})e^{(-|\varepsilon_k|/\sigma)}$, $\varepsilon_k\in\mathbb{R}$.

The ∞-norm $\parallel h(x_k)-g_m(x_k;\boldsymbol{\alpha})\parallel_\infty$, inducing the *inner product*:

$$\langle h(x_k)-g_m(x_k;\boldsymbol{\alpha})\rangle_\infty=\sup_{\varepsilon_k\in[-\sigma,\sigma]}|y_k-\alpha_0-\textstyle\sum_{k=1}^m\alpha_i\phi_i(x_k)|\,, \tag{17}$$

and *uniform convergence*, giving rise to the *Minimax* estimator of $\boldsymbol{\alpha}$ (see Shao, 2003), is related to the *Uniform* distribution, $f(\varepsilon_k)=\frac{1}{2\sigma}$, $\varepsilon_k\in[-\sigma,\sigma]$. Note that the natural base set for $C[-1,1]$ is provided by the Legendre polynomials whose weight function $w(x)$ for orthogonality relates to the uniform density; see Hildebrand (1982).

The importance of this connection stems from the fact that when the choice of a norm is based on the appropriateness of the corresponding distributional assumption vis-a-vis data $\mathbf{z}_0$, the resulting estimator $\widehat{\boldsymbol{\alpha}}_{ML}$ is the *Maximum Likelihood Estimator* of $\boldsymbol{\alpha}$, which enjoys certain desirable statistical properties, including parameterization invariance, which is particularly crucial in econometrics since the structural parameters are often reparametrizations/restrictions of the statistical parameters; see section 4.4. This suggests that in practice the choice of a norm should not be a matter of convenience, mathematical expediency or even *robustness* (see Shao, 2003, p. 346), but needs to be justified on statistical adequacy grounds.

3.3 Akaike-type model selection procedures

Akaike-type model selection procedures assume a prespecified family of models $\{\mathcal{M}_{\boldsymbol{\varphi}_i}(\mathbf{z}),$ $i=1,2,...m\}$ based on some variation/extension of (14). The stated objective of these procedures is motivated by the curve-fitting perspective and the selection is guided by trading goodness-of-fit against *overfitting*.

The initial attempt to curb overfitting was made by introducing the *adjusted* R^2:

$$\overline{R}^2_m=1-\left(\tfrac{n-1}{n-m-1}\right)(1-R^2), \tag{18}$$

that 'penalizes' the R^2 in (11) for increasing as m increases. The statistical literature continued with Akaike's (1970) Final Prediction Error (FPE), Mallows's (1973) C_p and Allen's (1971) Cross-Validation criteria:

$$\mathsf{FTE}_K=\tfrac{RSS_m}{n}\left(1+\tfrac{2K}{n-K}\right),\quad C_p=(\tfrac{nRSS_p}{RSS_m}-n+2p),\quad CV(1)\simeq RSS_m(\tfrac{n}{(n-K)^2}), \tag{19}$$

where $K=(m+1)$, $RSS_p=\sum_{k=1}^n(y_k-\widehat{\beta}_0-\sum_{i=1}^p\widehat{\beta}_i x_{ik})^2$ for $p\leq m$.

The model selection literature came of age with Akaike's Information Criterion (AIC) and related procedures such as the BIC, SIC, HQIC in the 1970s and 1980s; see Rao and Wu (2001). The AIC is based on curbing overfitting by penalizing the log-likelihood function ($\ln L(\boldsymbol{\theta})$) – measuring goodness-of- fit – using the number of unknown parameters (K) in $\boldsymbol{\theta}$: $\mathsf{AIC} = -2\ln L(\boldsymbol{\theta}) + 2K$.

The formal justification for this criterion is that AIC provides a pertinent estimator of the risk of the *Kullback-Leibler* (K-L) loss function:

$$\Delta_{K-L}(M_0, \widehat{M}_1) = E_0\left(\ln\left(\frac{f_0(\mathbf{x};\boldsymbol{\theta}_0)}{f_1(\mathbf{x};\widehat{\boldsymbol{\theta}}_1)}\right)\right) = \int_{\mathbf{z}\in\mathbb{R}^n_Z}(\ln[f(\mathbf{z};\boldsymbol{\theta}_0)/f_1(\mathbf{z};\widehat{\boldsymbol{\theta}}_1)])f(\mathbf{z};\boldsymbol{\theta}_0)\mathbf{dz}, \quad (20)$$

viewed as the expected discrepancy between the true model M_0 and an estimated model M_1. In particular, Akaike (1973) showed that when one compares different models within a prespecified family $\{\mathcal{M}_{\boldsymbol{\theta}_i}(\mathbf{z}),\ i{=}1, 2, ...m\}$:

$$\mathsf{AIC}(i) = -2\ln f_i(\mathbf{z};\widehat{\boldsymbol{\theta}}_i) + 2K_i,\ \ i{=}1, 2, ..., m, \quad (21)$$

provides an *asymptotically unbiased* estimator of $\Delta_{K-L}(M_0, \widehat{M}_i)$, K_i being the number of parameters in $\boldsymbol{\theta}_i$. For example, in the case of (12):

$$\mathsf{AIC} = n\ln(\widehat{\sigma}^2) + 2K, \quad \text{or} \quad \mathsf{AIC}_n = \ln(\widehat{\sigma}^2) + \tfrac{2K}{n}, \quad (22)$$

where $\widehat{\sigma}^2{=}\frac{1}{n}\sum_{k=1}^n(y_k{-}\widehat{\beta}_0{-}\sum_{i=1}^m\widehat{\beta}_i x_{ik})^2$, and $K{=}m{+}2$. The AIC procedure selects $\mathcal{M}_{\boldsymbol{\theta}_k}(\mathbf{z})$ when $\mathsf{AIC}(k)$ yields the minimum value. Early studies on the effectiveness of the AIC raised two crucial weaknesses (see McQuarrie and Tsai, 1998):

(i) The value $\widehat{K}$ chosen by AIC is an *inconsistent* estimator of the 'true' K^*,

(ii) In small samples AIC often leads to overfitting, i.e. ($\widehat{K}{>}\ K^*$).

Attempts to deal with the *inconsistency problem* gave rise to several modifications of the AIC_n, with the most widely used being the BIC (see Schwarz, 1978), Hannan and Quinn (1979) and Rissanen's (1978) Minimum Description Length criteria:

$$\mathsf{BIC}_n = \ln(\widehat{\sigma}^2) + \tfrac{K\ln(n)}{n}, \quad \mathsf{HQIC}_n = \ln(\widehat{\sigma}^2) + \tfrac{2K\ln(\ln(n))}{n}, \quad \mathsf{MDL}_n{=}(\mathsf{BIC}_K/2). \quad (23)$$

These are special cases of the generalized criterion: $\mathsf{GIC}_n = -2\ln(f(\mathbf{z};\widehat{\boldsymbol{\theta}})) + a(n)\cdot K$, where $a(n)$ is a smooth function of the sample size, which also includes Tacheuchi's (TIC) criterion; see Konishi and Kitaqawa (2008).

Attempts to deal with the *small sample overfitting* problem led to a several modifications of AIC, such as (see McQuarrie and Tsai, 1998):

$$\mathsf{M\text{-}AIC}_n = \ln(\widehat{\sigma}^2) + ([n{+}K]/[n{\text{-}}K{\text{-}}2])\,. \quad (24)$$

In what follows it is argued that (i)-(ii) above are the least of AIC's difficulties. The unreliability of inference problems (a)-(b) (section 1) are considerably more serious, and undermine equally all the modifications/extensions of the AIC, including cross-validation criteria and nonparametric methods which rely solely on mathematical approximation theory (see Konishi and Kitagawa, 2008). In all these areas, as well as curve-fitting (see Forster and Sober, 1994), the assumptions defining the likelihood function are often treated as an afterthought, and not as crucial stipulations whose inappropriateness vis-a-vis data $\mathbf{z}_0$ will undermind the reliability of inference.

3.4 Potential errors in model selection

Viewing the Akaike-type model selection procedures in the context of error statistics (Mayo and Spanos, 2009), calls for probing all the different ways the final inference:

$$\boxed{\mathcal{M}_{\boldsymbol{\varphi}_k}(\mathbf{z}) \text{ is the 'best' model within the prespecified family } \{\mathcal{M}_{\boldsymbol{\varphi}_i}(\mathbf{z}),\ i=1,2,...m\}}$$

might be in error, as well as delineating the notion of a 'best' model as it relates to the various objectives associated with using $\mathcal{M}_{\boldsymbol{\varphi}_k}(\mathbf{z})$.

It is argued that the primary objective of a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is to provide a sound basis for both, inductive inference (estimation, testing, prediction and simulation), as well as risk evaluation associated with different decision rules. In light of that, a minimum requirement for a 'best' model $\mathcal{M}_{\boldsymbol{\varphi}_k}(\mathbf{z})$ is *statistical adequacy*, because without it all inductive inferences, as well as risk evaluations, are likely to be misleadingly *unreliable*, voiding any practical usage of $\mathcal{M}_{\boldsymbol{\varphi}_k}(\mathbf{z})$.

This stems from the fact that a *misspecified* $f(\mathbf{z};\boldsymbol{\theta})$ will vitiate the sampling distribution $F_n(t)$ of any statistic $T_n{=}g(Z_1,...,Z_n)$, since by definition:

$$F_n(t)=\mathbb{P}(T_n\leq t)=\int\cdots\int_{\{\mathbf{z}:\ g(\mathbf{z})\leq t\}} f(\mathbf{z};\boldsymbol{\theta})d\mathbf{z},\ \ t{\in}\mathbb{R}. \tag{25}$$

In turn, $F_n(t)$ will undermine the reliability of any inference based on it by giving rise to actual error probabilities that are *different* from the nominal ones.

Example 4. Let $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ be a simple Normal model, i.e. $Z_k \backsim \mathsf{NIID}(\mu,\sigma^2)$, $k{=}1,...,n$, and consider testing: H_0: $\mu{=}\mu_0$ vs. H_1: $\mu{>}\mu_0$, using the t-test:

$$\tau(\mathbf{Z}){=}\frac{\sqrt{n}(\overline{Z}_n-\mu_0)}{s},\ C_1{:=}\{\mathbf{z}\text{: }\tau(\mathbf{z}){>}c_\alpha\},\ \overline{Z}_n{=}\tfrac{1}{n}\textstyle\sum_{k=1}^n Z_k,\ s^2{=}(\frac{1}{n-1})\sum_{k=1}^n(Z_k{-}\overline{Z}_n)^2.$$

What would the effect on the relevant error probabilities be if *Independence* is false? Instead, assume that $Corr(Z_i,Z_j){=}\rho$, $0{<}\rho{<}1$, for $i{\neq}j,\ i,j{=}1,...n$. For $n{=}100$, and a nominal $\alpha{=}.05$ ($c_\alpha{=}2.01$), even a small value $\rho{=}.1$ will yield an *actual* type I error of $\alpha^*{=}.317$; a sixfold increase! Similarly, the distortions in power, ranging from positive for tiny discrepancies $\gamma{=}\mu-\mu_0$, $\pi^*(\gamma{=}.01){-}\pi(\gamma{=}.01){=}.266$ [where more power is *not* needed], to negative for larger discrepancies, $\pi^*(\gamma{=}.3){-}\pi(\gamma{=}.3){=}-.291$ [where more power *is* needed], will render the test completely unreliable. Moreover, the distortions in both, the type I error and the power, get much larger as $\rho{\longrightarrow}1$; see Spanos (2009).

Similarly, the evaluation of the risk function for a decision rule $\widehat{\boldsymbol{\theta}}(\mathbf{Z})$ using a loss function $L(.)$, is likely to be highly misleading because the averaging is with respect to a *misspecified* $f(\mathbf{z};\boldsymbol{\theta})$ vitiating the result:

$$\mathcal{R}[\widehat{\boldsymbol{\theta}}(\mathbf{Z}),\boldsymbol{\theta}]=\int\cdots\int_{\mathbf{z}\in\mathbb{R}^n_Z} L(\widehat{\boldsymbol{\theta}}(\mathbf{z}),\boldsymbol{\theta})\cdot f(\mathbf{z};\boldsymbol{\theta})d\mathbf{z},\ \ \boldsymbol{\theta}\in\boldsymbol{\Theta}. \tag{26}$$

In the Akaike-type model selection literature the notion of 'best' is sometimes identified with the 'true' model, whatever that might mean (see Burnham and Anderson, 2002). In the error statistical framework that would require one to secure both statistical and *substantive adequacy* –$\mathcal{M}_{\boldsymbol{\varphi}_k}(\mathbf{z})$ provides a veritable explanation for the phenomenon of interest – which calls for additional probing of (potential) errors in bridging the gap between the two. In the present context, the notion of a 'best' model is assumed to include statistical adequacy [accounting for the *chance* regularities in data $\mathbf{z}_0$], since it is a necessary for assessing substantive adequacy.

The *first* potential error arises when the prespecified family $\{\mathcal{M}_{\boldsymbol{\varphi}_i}(\mathbf{z}),\ i{=}1,...m\}$ does *not* include an adequate model $\mathcal{M}_0(\mathbf{z})$. This is a variant of the statistical adequacy problem that arises when Akaike-type model selection procedures take the likelihood function at face value. This will invariably lead astray infering a best model because any use of error probabilities (or risks) will be misguiding; see section 3.5.

A *second* error arises when the family $\{\mathcal{M}_{\boldsymbol{\varphi}_i}(\mathbf{z}),\ i{=}1,...m\}$ *does* include a statistically adequate model, say $\mathcal{M}_{\boldsymbol{\varphi}_j}(\mathbf{z})$, but is different from the selected model $\mathcal{M}_{\boldsymbol{\varphi}_k}(\mathbf{z})$, $j{\neq}k$, which is inadequate. Model selection procedures ignore this potential error because, despite edicts to the contrary (see Burnham and Anderson, 2002), their norm-based minimization is tantamount to testing comparisons among the models within the prespecified family using Neyman-Pearson (N-P) testing, but without 'controlling' the relevant error (type I and II) probabilities; see section 3.6.

3.5 Selection within a misspecified family of models

Example 3 (continued). Consider using the Normal/Linear Regression model (table 2) to estimate the relationship between y_t-the population of the USA and x_t-a special predictor, using annual data for 1955-1989 (see Mayo and Spanos, 2004):

$$M(1):\quad y_t = \underset{(.610)}{167.115}+\underset{(.024)}{1.907}x_t+\widehat{u}_t,\ R^2{=}.995,\ s{=}1.77,\ n{=}35; \tag{27}$$

Despite the high R^2, consider exploring the possibility of using Akaike-type procedures, to choose $m{=}K$-1 within the broader family of models:

$$M(m):\quad y_k=\alpha_0+\textstyle\sum_{i=1}^{m}\alpha_i\psi_i(x_k)+\varepsilon_k, \tag{28}$$

where $\{\psi_i(x_k),\ i{=}1,...,m\}$ are *orthogonal Chebyshev* polynomials; see Powell (1981).

The results in table 1 indicate that all three criteria AIC, BIC and HQIC, as well as the modified AIC in (24), select the same model: $M(4)$

Applying certain simple Mis-Specification (M-S) tests (see Spanos and McGuirk, 2001), it can be shown (table 3) that (27) is statistically misspecified (assumptions [4]-[5] are invalid); p-values are reported in square brackets. Because of the statistical inadequacy of (28), the Akaike-type model selection procedures will always make fallacious selections; this issue was first highlighted by Lehmann (1990), p. 162. Note that this criticism extends to all the above selection procedures (see Rao and Wu, 2001). Statistical misspecification will give rise to unreliable inferences concerning the value of m with probability one.

Table 1 - AIC, BIC and HQIC based on (28)

Model	$\mathsf{AIC}_n=\ln(\widehat{\sigma}^2)+[2K/n]$,	rank	BIC_n	rank	HQIC_n	rank
$M(1)$	$\ln(2.9586)+[2(3)/35]{=}1.256$	5	1.389	4	1.302	4
$M(2)$	$\ln(2.5862)+[2(4)/35]{=}1.179$	3	1.357	3	1.240	3
$M(3)$	$\ln(2.5862)+[2(5)/35]{=}1.236$	4	1.458	5	1.313	5
$M(4)$	$\ln(1.8658)+[2(6)/35]{=}0.967$	1	1.233	1	1.059	1
$M(5)$	$\ln(1.8018)+[2(7)/35]{=}0.989$	2	1.300	2	1.096	2

Table 2 - Normal/Linear Regression Model

Statistical GM:		$y_t = \beta_0 + \boldsymbol{\beta}_1^\top \mathbf{x}_t + u_t, \quad t \in \mathbb{N}.$	
[1]	Normality:	$(y_t \mid \mathbf{X}_t = \mathbf{x}) \backsim \mathsf{N}(.,.),$	$\left.\begin{array}{l}\\\\\\\\\end{array}\right\}$ $t \in \mathbb{N}.$
[2]	Linearity:	$E(y_t \mid \mathbf{X}_t = \mathbf{x}_t) = \beta_0 + \boldsymbol{\beta}_1^\top \mathbf{x}_t,$	
[3]	Homoskedasticity:	$Var(y_t \mid \mathbf{X}_t = \mathbf{x}) = \sigma^2,$	
[4]	Independence:	$\{(y_t \mid \mathbf{X}_t = \mathbf{x}),\ t \in \mathbb{N}\}$ indep. process,	
[5]	t-invariance:	$(\beta_0, \beta_1, \sigma^2)$ are *not* changing with t,	

Table 3 - Mis-Specification (M-S) testing results	
Normality:	$D'AP(s) = -.482[.314],$
Linearity:	$F(1,29) = 4.608[.049]$
Heteroskedasticity:	$F(2,30) = .802[.458],$
Independence:	$F(1,31) = 6.608[.015]^*,$
t-invariance:	$F(1,29) = 156.273[.000]^*$

3.6 Selection within a well-specified family of models

Although the scenario that the prespecified family of models is well-specified is highly unlikely in practice, it is interesting to demonstrate that even in this best case scenario model selection procedures are likely give rise to misleading inferences.

Example 3 (continued). Using *statistical adequacy* as the criterion for 'best', one is led to the family of statistical models (see Mayo and Spanos, 2004):

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

which is the Dynamic Linear Regression (DLR) model with trends; see Spanos (1986). The question is whether Akaike-type criteria will choose the correct model, knowing that an adequate model [it accounts for the chance regularities in the data] exists.

Table 4 - AIC, BIC and HQIC based on (29)

Model	$\mathsf{AIC}_n = \ln(\widehat{\sigma}^2) + (2K/n),$	rank	BIC_n	rank	HQIC_n	rank
$M(1,1)$	ln(.057555)+[2(6)/35]=-2.512	9	−2.246	9	−2.420	9
$M(1,2)$	ln(.034617)+[2(8)/35]=-2.906	3	−2.551	1	−2.784	2
$M(1,3)$	ln(.033294)+[2(10)/35]=-2.831	5	−2.387	6	−2.678	6
$M(2,1)$	ln(.040383)+[2(7)/35]=-2.809	6	−2.498	3	−2.702	5
$M(2,2)$	ln(.033366)+[2(9)/35]=-2.886	4	−2.486	4	−2.748	4
$M(2,3)$	ln(.032607)+[2(11)/35]=-2.795	7	−2.306	8	−2.626	7
$M(3,1)$	ln(.042497)+[2(8)/35]=-2.701	8	−2.346	7	−2.578	8
$M(3,2)$	ln(.029651)+[2(10)/35]=-2.947	1	−2.502	2	−2.793	1
$M(3,3)$	ln(.026709)+[2(12)/35]=-2.937	2	−2.404	5	−2.753	3

The results in table 4 indicate that the model selected by the AIC and HQIC criteria is $M(3,2)$, and $M(1,2)$ is selected by the BIC. It turns out that the model selected on statistical adequacy grounds is $M(1,2)$:

$$y_t = \underset{(5.122)}{17.687} + \underset{(.080)}{.193}t - \underset{(.036)}{.000}x_t + \underset{(.147)}{1.496}y_{t-1} + \underset{(.037)}{.013}x_{t-1} - \underset{(.134)}{.596}y_{t-2} + \\ + \underset{(.035)}{.014}x_{t-2} + \widehat{\varepsilon}_t,\ R^2 = .9999,\ s = .154,\ n = 35. \tag{30}$$

The statistical adequacy of (30) is established via thorough Mis-Specification (M-S) testing where all the assumptions underlying the DLR model are validated vis-a-vis the data in question; see Mayo and Spanos (2004).

Parenthetically, the N-P test for the joint significance of the coefficients of (x_t, x_{t-1}, x_{t-2}) in (30) yields $F(3, 26)$=.142[.934], which indicates that the special predictor is totally *unrelated* to the US population; negating the earlier (unreliable) inference that x_t is an excellent predictor, based on the misspecified model (27).

The questions that naturally arise are 'what led to the different choices?', and 'why did the BIC select the correct model?' The answer is that Akaike-type procedures are often unreliable because their minimization of a normed-based function is tantamount to comparisons among the models within the prespecified family $\{\mathcal{M}_{\boldsymbol{\varphi}_i}(\mathbf{z})$, i=1, 2, ...$m\}$, based on Neyman-Pearson (N-P) hypothesis testing with *unknown error probabilities*. To illustrate that consider the question of choosing between:

$$\begin{array}{ll} M_2: & y_t = \beta_0 + \beta_1 x_t + \beta_2 x_t^2 + \beta_3 x_t^3 + u_t, \\ M_1: & y_t = \alpha_0 + \alpha_1 x_t + u_t, \end{array} \tag{31}$$

and assume that on the basis of the AIC procedure model M_2 was selected, i.e.

$$\left[n \ln(\widehat{\sigma}_1^2) + 2K_1\right] > \left[n \ln(\widehat{\sigma}_2^2) + 2K_2\right]. \tag{32}$$

This selection implies that $(\widehat{\sigma}_1^2/\widehat{\sigma}_2^2) > \exp([2(K_2\text{-}K_1)]/n)$, and one can relate the AIC decision in favor of M_2 to the rejection of the null:

$$H_0: \beta_2 = \beta_3 = 0, \quad \text{vs.} \quad H_1: \beta_2 \neq 0, \text{ or } \beta_3 \neq 0,$$

by the N-P test based on the F statistic (see Spanos, 1986, p. 426):

$$F(\mathbf{z}) = ([\widehat{\sigma}_1^2\text{-}\widehat{\sigma}_2^2]/\widehat{\sigma}_2^2)(\tfrac{n\text{-}K_2}{K_2\text{-}K_1}), \qquad C_1 := \{\mathbf{z}: F(\mathbf{z}) > c_\alpha\}, \tag{33}$$

where c_α denotes the critical value for significance level α; e.g. for α=.05 $\Rightarrow$ c_α=3.32. This suggests that the AIC procedure amounts to rejecting H_0 when:

$$F(\mathbf{z}) > k_{AIC}, \text{ where } k_{AIC} = (\tfrac{n\text{-}K_2}{K_2\text{-}K_1}) \left[\exp\left(\tfrac{2(K_2\text{-}K_1)}{n}\right) \text{-}1\right].$$

For n=35, k_{AIC}=1.816 implies that the implicit type I error is .180. Note that this coincides with the probability that the AIC will overfit by 2 parameters. The AIC procedure is inconsistent because, asymptotically the probability of selecting M_1 when true is less than one: $\lim_{n\to\infty} \mathbb{P}(F(\mathbf{Z}) \leq k_{AIC}; H_0) < 1$; see McQuarrie and Tsai (1998).

The same argument as in (32) yields the implicit critical values for BIC and HQIC:

$$k_{BIC} = (\tfrac{n-K_2}{K_2\text{-}K_1}) \left[\exp\left(\tfrac{(K_2\text{-}K_1)\ln(n)}{n}\right) \text{-}1\right], \quad k_{HQIC} = (\tfrac{n-K_2}{K_2\text{-}K_1}) \left[\exp\left(\tfrac{2(K_2\text{-}K_1)\ln(\ln(n))}{n}\right) \text{-}1\right],$$

where for n=35, k_{BIC}=3.379, k_{HQIC}=2.340. Hence, the implicit type I error probabilities for BIC and HQIC are .047 and .114, respectively. The BIC error probability of .047 is close to the traditional significance levels used in establishing the statistical adequacy of (30), and might explain why the BIC selected the correct model in *this*

particular example. However, this result is coincidental, and should not be used as an argument in favor of the BIC because these implicit error probabilities are generally unknown, and they depend crucially on both n and K. As shown in section 5.2, the BIC leads the inference astray when K changes.

The main conclusion one can draw from the above empirical examples is that the Akaike-type procedures will only *circumstantially* select a statistically adequate model and always unbeknown to the modeler. The next section summarizes the error statistical modeling framework where statistical adequacy can be *deliberately* secured with a view to ensure the reliable appraisal of substantive information.

4 Bridging over statistical and structural models

In the error statistical framework the reconciliation between the statistical and substantive information is viewed in the broader context of bridging the gap between theory and data $\mathbf{z}_0$ using a sequence of interconnecting models: theory, structural (estimable) and statistical models; see Spanos (1986).

4.1 Statistical vs. substantive information

It is generally acknowledged that both substantive and statistical information play important roles in learning from data, and in practice empirical models constitute an amalgam of both sources of information, but the role of each type of information has not been delineated in modern statistics; Lehmann (1990), Spanos (2006a).

In the curve-fitting problem, the family of models is determined mainly by approximation theory (substantive) information based on the 'smoothness' of $y{=}h(x)$, $(x,y){\in}\mathbb{R}_X{\times}\mathbb{R}_Y$. For instance, in example 2 this theory asserts that when $h(x){\in}L_2(\text{-}\infty,\infty)$, the *Hermite orthogonal polynomials* provide a complete base set, and the 2-norm ensures the existence and uniqueness of $g_m(x;\boldsymbol{\alpha}){=}\sum_{i=0}^{m}\alpha_i h_i(x)$. None of these choices, however, has anything to do with the chance regularities in data $\mathbf{z}_0$.

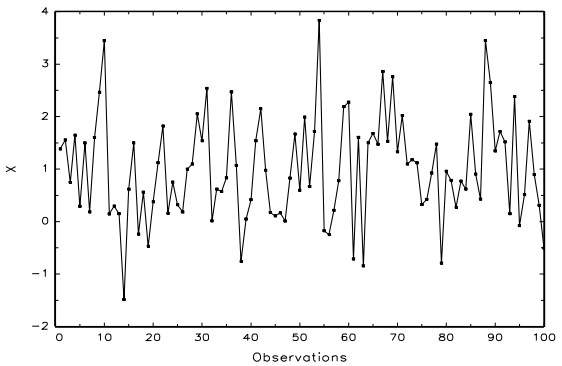

Fig. 1: t-plot of x_t

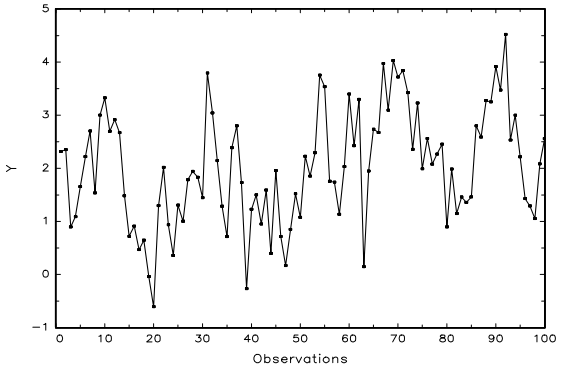

Fig. 2: t-plot of y_t

Broadly speaking, statistical information refers to the *chance regularity patterns* exhibited by the data when viewed as realizations of *generic stochastic processes,* without any information pertaining to what they represent (measure) substantively. For instance, in figure 1 one can see a typical realization of a process $\{X_k,\ k{\in}\mathbb{N}\}$

assumed to be NIID, but figure 2 depicts a typical realization of a process $\{Y_k,\ k\in\mathbb{N}\}$ assumed to be Normal, Markov and Stationary. This can be discerned from these plots using analogical reasoning; see Spanos (1999), ch. 5.

4.2 From a theory to a structural model

The term theory is used generically as any claim conjectured to elucidate a phenomenon of interest. When one proposes a *theory* 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 $\mathbf{x}_k$, aware that there might be numerous other potentially relevant factors $\boldsymbol{\xi}_k$ (observable and unobservable) influencing the behavior of y_k.

A *theory model* corresponds to an idealized mathematical representation of a phenomenon of interest that facilitates 'learning' whose generic form is:

$$y_k = h^*(\mathbf{x}_k, \boldsymbol{\xi}_k), \quad k\in\mathbb{N}. \tag{34}$$

The guiding principle in selecting the variables in $\mathbf{x}_k$ is to ensure that they account for the *systematic* behavior of y_k, and the omitted factors $\boldsymbol{\xi}_k$ represent non-essential disturbing influences which, collectively, have only a non-systematic effect on y_k. The potential presence of a large number of contributing factors $(\mathbf{x}_k, \boldsymbol{\xi}_k)$ explains the conjuring of *ceteris paribus* clauses. This line of reasoning transforms the theory model (34) into a *structural (estimable) model* of the form:

$$y_k = h(\mathbf{x}_k;\boldsymbol{\varphi}) + \epsilon(\mathbf{x}_k,\boldsymbol{\xi}_k), \quad k\in\mathbb{N}, \tag{35}$$

where $h(.)$ denotes the postulated functional form, $\boldsymbol{\varphi}$ stands for the structural parameters of interest. The *substantive error term*, defined to represent all unmodeled influences, is often a function of both $\mathbf{x}_k$ and $\boldsymbol{\xi}_k$:

$$\{\epsilon(\mathbf{x}_k,\boldsymbol{\xi}_k) = y_k - h(\mathbf{x}_k;\boldsymbol{\varphi}),\ k\in\mathbb{N}\}. \tag{36}$$

How does this relate to the approximation error in (9)? As argued above, the choice of $g_m(x_k;\boldsymbol{\alpha})=\sum_{i=0}^{m}\alpha_i\phi_i(x_k)$ relied on substantive (mathematical) information and thus $\{\varepsilon_k(x_k,m)=y_k-g_m(x_k;\boldsymbol{\alpha}),\ k\in\mathbb{N}\}$, constitutes a special case of (36), where the degree m of $g_m(x_k;\boldsymbol{\alpha})$ plays the same role as $\boldsymbol{\xi}_k$ in (35), in the sense that m stands for the omitted higher order terms $\phi_i(x)$, $i=m+1, m+2, ...$

For (35) to provide a meaningful model for y_k, $\epsilon(\mathbf{x}_k,\boldsymbol{\xi}_k)$ needs to be non-systematic:

$$\text{[i]}\quad \epsilon(\mathbf{x}_k,\boldsymbol{\xi}_k) \backsim \mathsf{IID}(0,\sigma^2),\ \forall(\mathbf{x}_k,\boldsymbol{\xi}_k)\in\mathbb{R}_{\mathbf{x}}\times\mathbb{R}_{\boldsymbol{\xi}}. \tag{37}$$

In addition, one needs to ensure that the GM (35) is 'nearly isolated' (Spanos, 1995):

$$\text{[ii]}\quad E(\epsilon(\mathbf{x}_k,\boldsymbol{\xi}_k)\cdot h(\mathbf{x}_k;\boldsymbol{\phi}))=0,\ \forall(\mathbf{x}_k,\boldsymbol{\xi}_k)\in\mathbb{R}_{\mathbf{x}}\times\mathbb{R}_{\boldsymbol{\xi}}. \tag{38}$$

The assumptions [i]-[ii] are clearly non-testable vis-a-vis data $\mathbf{z}_0:=(\mathbf{z}_1,\mathbf{z}_2,...,\mathbf{z}_n)$ because their confirmation would involve *all possible values* of both $\mathbf{x}_k$ and $\boldsymbol{\xi}_k$. To render them testable vis-a-vis data $\mathbf{z}_0$ one needs to embed the structural model (35) into a statistical model built on chance regularities in $\mathbf{z}_0$ reflecting the probabilistic structure of $\{\mathbf{Z}_k:=(y_k,\mathbf{X}_k),\ k\in\mathbb{N}\}$; a crucial modeling move that also addresses questions (A)-(B) of section 3.1. The embedding depends crucially on the nature of the available data $\mathbf{z}_0$ and their relation to the theory in question; sometimes the gap between them might be unbridgeable (see Spanos, 1995).

4.3 Embedding a structural into a statistical model

The nature of the embedding itself depends on whether the data $\mathbf{z}_0$ are the result of an experiment or they are observational in nature, but the aim in both cases is to find a way to transform the substantive error $\epsilon(\mathbf{x}_k,\boldsymbol{\xi}_k)$, for all $(\mathbf{x}_k,\boldsymbol{\xi}_k)\in\mathbb{R}_{\mathbf{x}}\times\mathbb{R}_{\boldsymbol{\xi}}$ into a *generic IID process* without the quantifier; see Spanos (2006a).

Experimental data. In the case where one can perform experiments, controls and 'experimental design' techniques such as *replication*, *randomization* and *blocking,* can often be 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 objective is to transform ($\rightsquigarrow$) $\epsilon(\mathbf{x}_k,\boldsymbol{\xi}_k)$ into a generic IID error:

$$(\epsilon(\mathbf{x}_k,\boldsymbol{\xi}_k)\,\|\text{experimental controls \& designs})\rightsquigarrow\varepsilon_k\backsim\mathsf{IID}(0,\sigma^2),\quad k=1,...,n. \tag{39}$$

This, in effect, embeds the structural model (35) into a *statistical model* $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$:

$$y_k=h(\mathbf{x}_k;\boldsymbol{\theta})+\varepsilon_k,\quad \varepsilon_k\backsim\mathsf{IID}(0,\sigma^2),\ k=1,2,...,n, \tag{40}$$

where the statistical error term ε_k in (40) is qualitatively very different from the substantive error term $\epsilon(\mathbf{x}_k,\boldsymbol{\xi}_k)$ in (35), because ε_k is no longer a function of $(\mathbf{x}_k,\boldsymbol{\xi}_k)$, and its assumptions are rendered testable vis-a-vis data $\mathbf{z}_0$; see Spanos (2006a). A widely used special case of (40) is the *Gauss Linear model* (12).

Observational data. In the case of observational (non-experimental) data $\mathbf{z}_0$ the embedding takes a different form in the sense that the experimental control and intervention are replaced by judicious *conditioning* on an appropriate information set $\mathfrak{D}_k$; often generated by an observable process, say $\sigma(\mathbf{X}_k)$. The generating mechanism of the embedding statistical model takes the general form:

$$y_k=E(y_k\mid\mathfrak{D}_k)+u_k,\quad k\in\mathbb{N}, \tag{41}$$

where $\mu_k=E(y_k\mid\mathfrak{D}_k)$, denotes the *systematic component* and $\mathfrak{D}_k$ is a proper subset of the σ-field $\mathfrak{F}$ in the probability space $(S,\mathfrak{F},\mathbb{P}(.))$, on which the process $\{\mathbf{Z}_k,\ k\in\mathbb{N}\}$ is defined. $\mathfrak{D}_k$ is chosen in such as way so as to render the error process $\{u_k,\ \mathfrak{D}_k,\ k\in\mathbb{N}\}$:

$$u_k=y_k-E(y_k|\mathfrak{D}_k), \tag{42}$$

non-systematic in the sense of being a *Martingale-difference (M-d)* process relative to $\mathfrak{D}_k$; see Doob (1953). The statistical error term is not treated as as an autonomous but as a derived process whose probabilistic structure is determined by that of the observable stochastic process $\{\mathbf{Z}_k,\ k\in\mathbb{N}\}$. For instance, in the case of the process $\{\mathbf{Z}_k:=(y_k,\mathbf{X}_k),k\in\mathbb{N}\}$, where $E(\mathbf{Z}_k)<\infty$, the conditioning information set $\mathfrak{D}_k=\sigma(y_{k-1},y_{k-2},...,y_1,\mathbf{X}_k,...,\mathbf{X}_1)$, defines a M-d process in the sense that $E(u_k|\mathfrak{D}_k)=0$, and $E(u_k\mu_k|\mathfrak{D}_k)=0$, $k\in\mathbb{N}$, follows from (42).

In specifying a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ one has a *twofold objective* in mind:
[I] to account for the chance regularities in data $\mathbf{z}_0$ by choosing a probabilistic structure for the stochastic process $\{\mathbf{Z}_k,\ k\in\mathbb{N}\}$ so as to render the data $\mathbf{z}_0$ a *truly typical realizations thereof*, and
[II] to *parameterize* the probabilistic structure of $\{\mathbf{Z}_k,\ k\in\mathbb{N}\}$ in an attempt to specify an adequate statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ that would embed (nest) the structural model of interest $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ in its context.

It is important to emphasize that the functional form of the systematic component $E(y_k|\mathfrak{D}_k)$ will be determined exclusively by the statistical information described by the joint $D(\mathbf{Z}_1,\mathbf{Z}_2,...,\mathbf{Z}_n;\boldsymbol{\phi})$. For instance, when $\{\mathbf{Z}_k,\ k\in\mathbb{N}\}$ is IID and $\mathfrak{D}_k=\sigma(\mathbf{X}_k)$, the functional form of the regression and skedastic functions, $E(y_k|X_k)=h(X_k)$, $Var(y_k|X_k)=g(X_k)$, are determined by $D(y_k,X_k;\boldsymbol{\psi})$ since:

$$D(y_k|X_k;\boldsymbol{\theta})=[D(y_k,X_k;\boldsymbol{\psi})/\textstyle\int_{y_k\in\mathbb{R}_Y} D(y_k,X_k;\boldsymbol{\psi})dy_k],\ \forall(y_k,x_k)\in\mathbb{R}_y\times\mathbb{R}_X.$$

When $D(y_k,X_k;\boldsymbol{\psi})$ is assumed to be Normal, $E(y_k|X_k)=\beta_0+\beta_1X_k$ and $Var(y_k|X_k)=\sigma^2$, giving rise to the Normal/Linear Regression (N/LR) model (table 2). Indeed, the N/LR model can be formally viewed as a tripartite *reduction* of the form:

$$D(\mathbf{Z}_1,\mathbf{Z}_2,...,\mathbf{Z}_n;\boldsymbol{\phi})\overset{\mathsf{NIID}}{\rightsquigarrow}\textstyle\prod_{k=1}^n D(y_k|\ \mathbf{X}_k;\boldsymbol{\theta}).$$

This gives rise to the N/LR model, specified in terms of assumptions [1]-[5] (table 2). The relationship between the observable process $\{(y_k|\ \mathbf{X}_k=\mathbf{x}_k),\ k\in\mathbb{N}\}$ and the statistical error process, as defined in (42), is:

$$(y_k|\ \mathbf{X}_k=\mathbf{x}_k)\backsim\mathsf{NI}(\beta_0+\boldsymbol{\beta}_1^\top\mathbf{x}_t,\sigma^2)\Rightarrow(u_k|\ \mathbf{X}_k=\mathbf{x}_k)\backsim\mathsf{NIID}(0,\sigma^2).$$

In view of the relationship between the probabilistic structure of the process $\{\mathbf{Z}_k,\ k\in\mathbb{N}\}$ and the model assumptions, one can use the chance regularities exhibited by data $\mathbf{z}_0$ to guide the selection of an appropriate statistical model. For example, if the scatter plot of $\mathbf{z}_0$ looks like figure 3, the N/LR model would be an appropriate choice, but if it looks like figure 4, N/LR model will be inappropriate, irrespective of the structural model. This is because fig. 4 exhibits a typical realization of an Exponential IID process whose regression and skedastic functions are (Spanos, 1999):

$$E(y_k|X_k)=\tfrac{(1+\theta+\theta X_k)}{(1+\theta X_k)^2},\quad Var(y_k|X_k)=\tfrac{[(1+\theta+\theta X_k)^2-2\theta^2]}{[1+\theta X_k]^4},\quad x_k\in\mathbb{R}_+,\theta>0.$$

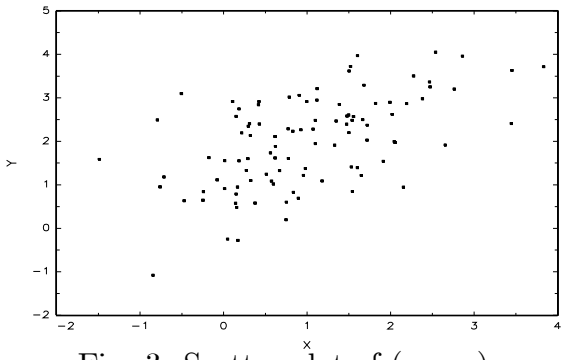

Fig. 3: Scatter-plot of (x_k,y_k)

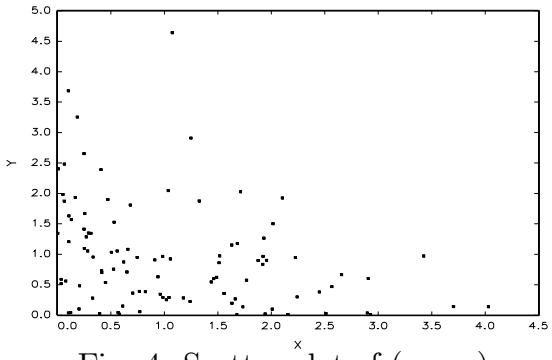

Fig. 4: Scatter-plot of (x_k,y_k)

In this sense the statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is built primarily on statistical information, and has 'a life of its own' in the sense that it constitutes a parameterization of a stochastic process $\{\mathbf{Z}_k,\ k\in\mathbb{N}\}$, underlying data $\mathbf{z}_0$, chosen to account for the chance regularities in data $\mathbf{z}_0$, which can be gleaned in figures 1-4. In this sense, a statistically adequate model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ provides a form of *statistical knowledge*, against which the substantive information can be appraised. Hence, substantive information enhances learning from data when it does not contravene statistical knowledge.

4.4 Reconciling substantive and statistical information

Substantive subject matter information is crucially important in learning from data about phenomena of interest, but no systematic learning can take place in the context of a statistically misspecified model. For reliable assessment of substantive questions of interest it is imperative to have a statistically adequate model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ which is built on separate information, and therefore, can be used to provide the broader inductive premises for evaluating *substantive adequacy*. The latter requires the structural model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ to provide a veritable explanation for the phenomenon of interest, and necessitates probing (potential) errors in bridging the gap between the two; this includes external validity, confounding effects and other concerns (see Spanos, 2006b-c).

The first step in assessing substantive information is to embed the structural $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ into a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ via reparametrization/restriction, in the form of the implicit function $\mathbf{G}(\boldsymbol{\varphi},\boldsymbol{\theta})=\mathbf{0}$ where $\boldsymbol{\varphi}$ and $\boldsymbol{\theta}$ denote the structural and statistical parameters, respectively. This provides a link between $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ and the phenomenon of interest that takes the form of **identification**:

Does the implicit function $\mathbf{G}(\boldsymbol{\varphi},\boldsymbol{\theta})=\mathbf{0}$ define $\boldsymbol{\varphi}$ *uniquely* in terms of $\boldsymbol{\theta}$?

Often, there are more statistical than structural parameters, and that enables one to test the additional substantive information using the *overidentifying restrictions*:

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

This error statistical view of identification differs from the traditional textbook notion (see Greene, 2003) in so far as it requires that the underlying statistical model (the reduced form) be validated vis-a-vis data $\mathbf{z}_0$ for the link between structural parameters and the phenomenon of interest to be rendered trustworthy; Spanos (1990).

5 Model specification vs. model selection

5.1 Statistical model specification

Statistical information refers the chance regularities (recurring patterns) exhibited by data $\mathbf{y}_0$ when viewed as a realization of a *generic* stochastic process $\{Y_k,\ k\in\mathbb{N}\}$, irrespective of what the data quantify (substantively). This enables one to provide a purely probabilistic construal of a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{y})$, by viewing it as a parameterization of the probabilistic structure of the process $\{Y_k,\ k\in\mathbb{N}\}$, as summarized by $f(\mathbf{y};\boldsymbol{\theta})$. $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{y})$ is viewed as an *idealized probabilistic description* of the stochastic mechanism that gave rise to data $\mathbf{y}_0$. In this sense, $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{y})$ is built exclusively on statistical systematic information in data $\mathbf{y}_0$, and its parameterization chosen so as to embed the structural model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{y})$ in its context; see section 4.

Example 5. Specification begins with a data set $\mathbf{y}_0{:=}(y_{-p},y_{-p+1},...,y_1,...,y_n)$, say figure 2, and poses the question 'what kind of probabilistic structure for $\{Y_t,\ t\in\mathbb{N}\}$ would render that data a typical realization thereof?' Using analogical reasoning, one can conjecture that $\{Y_t,\ t\in\mathbb{N}\}$ being Normal (N), Markov(p) (M) and Stationary (S), would do just that. Imposing these assumptions yields the following reduction:

$$f(y_1,y_2,...,y_n;\boldsymbol{\phi})\ \overset{\text{M\&S}}{=}\ f_p(y_1,y_2,...,y_p;\boldsymbol{\theta}_p)\textstyle\prod_{t=p+1}^{n} f(y_t|y_{t-1},...,y_{t-p};\boldsymbol{\theta}),$$

where the Normality of $\{Y_t,\ t\in\mathbb{N}\}$ implies that for $\mathbf{Y}_{t-1}^{p}:=(Y_{t-1},...,Y_{t-p})$:

$$(Y_t \mid \mathbf{Y}_{t-1}^{p}) \ \backsim \mathsf{N}\left(\alpha_0+\textstyle\sum_{t=1}^{p}\alpha_t Y_{t-t},\ \sigma_0^2\right),\ \ t\in\mathbb{N}.$$

This reduction gives rise to the AR(p) model in terms of complete and internally consistent set of testable [vis-à-vis data $\mathbf{y}_0$] probabilistic assumptions [1]-[5] (table 5); this being necessary for statistical adequacy purposes.

This form of specification often demands (i) recasting assumptions about *unobservable errors* into equivalent assumptions in terms of the observable process $\{Y_k,\ k\in\mathbb{N}\}$, as well as (ii) unveiling *hidden assumptions*; see Spanos (1986). To illustrate this issue note that assumptions [1]-[5] (table 5) imply that the error term $\{(u_t|\mathbf{Y}_{t-1}^{m}),\ t\in\mathbb{N}\}$ defines a [i] Normal (N), [ii] Martingale-difference (M-d) process:

$$(u_t \mid \mathbf{Y}_{t-1}^{p}) \ \backsim \mathsf{NM\text{-}d}\left(0,\ \sigma_0^2\right),\ \ t\in\mathbb{N}, \tag{43}$$

but the converse is not true. The error assumptions in (43) do not provide a complete set of assumptions for the AR(p) model, because [i[-[ii] allow for t-varying coefficient parameters, in the sense that they hold even if $u_t=Y_t-\alpha_0(t)-\sum_{k=1}^{p}\alpha_k(t)Y_{t-k}$. Hence, in terms of (43), the crucial assumption [5] is veiled, and thus rarely tested in practice.

Table 5 - Normal/AutoRegressive Model

Statistical GM:		$Y_t=\alpha_0+\sum_{k=1}^{p}\alpha_k Y_{t-k}+u_t,\ t\in\mathbb{N}.$	
[1]	Normality:	$\left(Y_t \mid \mathbf{Y}_{t-1}^{p}\right)\backsim\mathsf{N}(.,.),\ y_t\in\mathbb{R},$	$\Big\}t\in\mathbb{N}.$
[2]	Linearity:	$E\left(Y_t \mid \mathbf{Y}_{t-1}^{p}\right)=\alpha_0+\sum_{k=1}^{p}\alpha_k Y_{t-k},$	
[3]	Homoskedasticity:	$Var\left(Y_t \mid \mathbf{Y}_{t-1}^{p}\right)=\sigma_0^2,$	
[4]	Markov dependence:	$\{Y_t,\ t\in\mathbb{N}\}$ is a Markov(p) process,	
[5]	t-invariance:	$(\alpha_0,\alpha_1,...,\alpha_p,\sigma_0^2)$ are *not* changing with t,	

5.2 Statistical adequacy and model selection

In securing statistical adequacy the optimal value of p is decided as part of the validation process for all [1]-[5] assumptions (table 5), with particular emphasis placed on [4]; see Andreou and Spanos (2003). It can be shown that thorough M-S testing can guard against both underfitting and overfitting since both induce detectable systematic information in the residuals; see Spanos (1986).

This perspective questions the role of Akaike-type model selection procedures in determining p, which are widely used in econometrics (see Greene, 2003, Lutkepohl, 2005). As argued above, one of the reasons why the capacity of these procedures to ensure the reliability of inference is severely impaired is because statistical adequacy is ignored. What if one were to secure the *statistical adequacy* of the prespecified family of models $\{\mathcal{M}_{\boldsymbol{\varphi}_i}(\mathbf{z}),\ i=1,2,...m\}$ first, and then apply these model selection procedures? The fact of the matter is that establishing statistical adequacy renders these model selection procedures redundant. But even if we ignore that, model selection is likely to yield unreliable inference because, as argued above, their selection is tantamount to applying N-P testing with unknown error probabilities. Indeed, the fact that BIC selected the correct model in table 4 was largely coincidental.

Example 3 (continued). To illustrate the unreliability of Akaike-type model selection procedures further, let y_t - the US population annual data (1955-1989), and consider the selection of a model within the $AR(k,p)$ family:

$$AR(k,p): \quad Y_t = \delta_0 + \sum_{i=1}^{k} \delta_i t + \sum_{i=1}^{p} a_i Y_{t-i} + \varepsilon_t. \tag{44}$$

In total accord with (30), $AR(1,2)$ is chosen on statistical adequacy grounds:

$$Y_t = \underset{(3.507)}{12.279} + \underset{(.047)}{.153}t + \underset{(.123)}{1.560}Y_{t-1} - \underset{(.113)}{.628}Y_{t-2} + \widehat{\varepsilon}_t, \ R^2{=}.999, \ s{=}.1477. \tag{45}$$

In contrast, all three procedures, AIC, BIC and HQIC (see table 6) selected the $AR(3,2)$ model. By comparing the results of tables 4 and 6, it becomes clear that dropping the insignificant terms (x_t, x_{t-1}, x_{t-2}) from (30), to render the model more parsimonious, resulted in the BIC switching its selection to the wrong model!

Table 6 - AIC, BIC and HQIC based on (44)

Model	$\mathsf{AIC}_n = \ln(\widehat{\sigma}^2) + [2K/n]$,	rank	BIC_n	rank	HQIC_n	rank
$AR(1,1)$	ln(.038545)+[2(4)/35]=-3.027	8	−2.850	7	−2.966	7
$AR(1,2)$	ln(.019320)+[2(5)/35]=-3.661	3	−3.439	2	−3.584	3
$AR(1,3)$	ln(.018818)+[2(6)/35]=-3.630	4	−3.363	3	−3.538	4
$AR(2,1)$	ln(.036340)+[2(5)/35]=-3.029	7	−2.807	8	−2.952	8
$AR(2,2)$	ln(.019150)+[2(6)/35]=-3.613	5	−3.346	5	−3.521	5
$AR(2,3)$	ln(.018795)+[2(7)/35]=-3.574	6	−3.263	6	−3.467	6
$AR(3,1)$	ln(.035610)+[2(6)/35]=-2.992	9	−2.726	9	−2.900	9
$AR(3,2)$	ln(.015525)+[2(7)/35]=-3.765	1	−3.454	1	−3.658	1
$AR(3,3)$	ln(.015525)+[2(8)/35]=-3.708	2	−3.353	4	−3.585	2

5.3 Mis-Specification (M-S) testing and Respecification

The question that one might naturally pose at this stage is that, despite the apparent differences sketched above, both model selection and the model specification procedures come down to comparing one statistical model to another to find out which one is more appropriate. Such a view represents a misleading oversimplification.

A closer look at the above specification argument for AR(p), reveals that one is *not* choosing a statistical model as such, but a probabilistic structure for the stochastic process $\{Y_k, \ k{\in}\mathbb{N}\}$ that would render data $\mathbf{y}_0$, a typical realization thereof; $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{y})$ is a particular parameterization of this structure. This standpoint sheds very different light on the problem of *underdetermination* in this context. There can be two statistically adequate models only when they represent two alternative parametrizations of the same probabilistic structure; see Spanos (2007a). The choice between them is made using other criteria, including the substantive questions of interest.

The selected model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{y})$ is viewed as an element of the set $\mathcal{P}(\mathbf{y})$ which includes all possible statistical models that could have given rise to data $\mathbf{y}_0$. But how does one narrow down a possibly infinite set $\mathcal{P}(\mathbf{y})$ to one model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{y})$? The narrowing down is attained by partitioning $\mathcal{P}(\mathbf{y})$ using probabilistic assumptions from three broad

categories: Distribution (D), Dependence (M) and Heterogeneity (H); see Spanos (1995).

Example 6. The partitioning by reduction is illustrated in figure 5 in the case the simple Normal model which is based on the reduction assumptions that $\{Y_k,\ k\in\mathbb{N}\}$ is (D) Normal, (M) Independent and (H) Identically Distributed, denoted by $Y_k \backsim \mathsf{NIID}(\mu, \sigma^2),\ k\in\mathbb{N}$; a model that seems appropriate for the data in figure 1.

The tripartite partitioning also plays a crucial role in **M-S testing** based on:

$$H{:}f^*(\mathbf{y})\in\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{y}) \quad \text{vs.} \quad \overline{H}{:}f^*(\mathbf{y})\in [\mathcal{P}(\mathbf{y})-\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{y})]\,, \tag{46}$$

where $f^*(\mathbf{y})$ denotes the 'true' distribution of the sample. The probing *beyond* the boundaries of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{y})$ raises several conceptual and technical issues concerning its effectiveness and reliability. The partitioning of $\mathcal{P}(\mathbf{y})$ creates a framework wherein one can formally assess the model assumptions, relating to $\{Y_k,\ k\in\mathbb{N}\}$, using informed M-S testing, because it provides an *exhaustively complete* probing strategy. Changing the original reduction assumptions in deliberative ways, in light of the information one can glean from exploratory data analysis, gives rise to effective M-S tests which can eliminate an *infinite* number of alternative models at a time; see Spanos (1999). The most *inefficient* way to do this is to attempt to probe $[\mathcal{P}(\mathbf{y})-\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{y})]$ one model at a time $\mathcal{M}_{\boldsymbol{\varphi}_i}(\mathbf{y})$, $i=1,2,..$ since there is an infinity of models to search through.

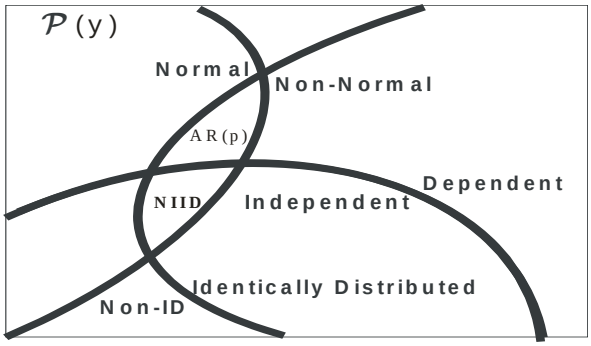

Fig. 5 - Specification by partitioning

Respecification amounts to returning to $\mathcal{P}(\mathbf{y})$ and recasting the original reduction assumptions in an attempt to account for statistical systematic information unaccounted for by the original model. For instance, the Normal, AR(p) model in table 5 can be viewed as a respecification of the simple Normal model where the NIID assumptions have been replaced by the assumptions that $\{Y_k,\ k\in\mathbb{N}\}$ is (D) Normal, (M) Markov and (H) Stationary; see figure 5.

This error statistical strategy of M-S testing and respecification by re-partitioning is in complete contrast to the traditional textbook approach based on ad hoc diagnostics and 'furbishing up' the original model using 'error-fixing' techniques. It can be shown that ad hoc and partial M-S testing can easily give rise to unreliable diagnoses, and the traditional error-fixing strategies, such as error-autocorrelation and heteroskedasticity corrections, as well as the use of heteroskedasticity consistent standard errors (see Greene, 2003), do *not* address the unreliability of inference problem. If anything, they often make matters worse; see Spanos and McGuirk (2001).

5.4 The error statistical approach: taking stock

Returning to the methodological questions (I)-(V) raised in the introduction, one can summarize the answers proposed by the error statistical approach as follows.

(I)* The set of all possible models $\mathcal{P}(\mathbf{z})$ can be narrowed down to a single model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ using a three-way partitioning based on probabilistic assumptions pertaining to the process $\{\mathbf{Z}_k,\ k\in\mathbb{N}\}$ underlying data $\mathbf{z}_0$. By definition $\mathcal{P}(\mathbf{z})$ includes a true probabilistic structure for $\{\mathbf{Z}_k,\ k\in\mathbb{N}\}$ – one that would render data $\mathbf{z}_0$ a truly typical realization thereof – but there is no guarantee that one will always be able to parameterize this structure to get an operational model, whatever the chance regularities in data $\mathbf{z}_0$. That will depend on whether this true probabilistic structure has sufficient *invariant* features to be adequately parameterized by constant parameters $\boldsymbol{\theta}$.

(II)* The adequacy of such a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is assessed *a posteriori* by probing its probabilistic assumptions (e.g. [1]-[5] in tables 2 and 5) vis-a-vis data $\mathbf{z}_0$ using thorough M-S testing to secure their validity. Statistical adequacy answers the question 'when $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ accounts for the chance regularities in data $\mathbf{z}_0$.'

(III)* Foisting the substantive information on the data by estimating the structural model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ directly, is invariably a rash strategy because statistical specification errors are likely to undermine the prospect of reliably evaluating the relevant errors for primary inferences. When modeling with observational data, the estimated $\mathcal{M}_{\widehat{\boldsymbol{\varphi}}}(\mathbf{z})$ is often both statistically and substantively inadequate, and one has no way to delineate the two; is the theory wrong or are the (implicit) inductive premises invalid for data $\mathbf{z}_0$? To avert this impenetrable dilemma, error statistics proposes to distinguish, *ab initio*, between statistical and substantive information and then bridge the gap between them by a sequence of interconnecting models which enable one to delineate and probe for the potential errors at different stages of modeling. From the theory side, the substantive information is initially encapsulated by a theory model and then modified into a structural one $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ to render it estimable with data $\mathbf{z}_0$. From the data side, the statistical information is distilled by a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ whose parameterization is chosen with a view to render $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ a reparametrization/restriction thereof. The statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is secured first in order to ensure the reliability of the procedures for appraising *substantive* claims.

(IV)* Error statistics proposes a blending in of the Fisherian and Neyman-Pearson (N-P) perspectives that weaves a coherent frequentist inductive reasoning anchored firmly on *error probabilities*. The key is provided by realizing that the p-value is a *post-data* and the type I and II are *pre-data* error probabilities, and that they fulfill crucial complementary roles. Pre-data error probabilities are used to appraise the generic capacity of inference procedures, and post-data error probabilities are used to bridge the gap between the coarse 'accept/reject' and evidence provided by data $\mathbf{z}_0$ for or against substantive claims; see Cox and Mayo (2009).

(V)* Post-data error probabilities can be used to address both *the fallacy of acceptance* and *the fallacy of rejection*, using a *post-data* evaluation of inference based on severe testing reasoning. This amounts to establishing the smallest (largest) discrepancy $\gamma \geq 0$ from H_0 warranted by data $\mathbf{z}_0$, associated with the N-P decision to accept (reject) H_0; see Mayo and Spanos (2006), Mayo and Cox (2006).

6 Methodological issues raised by M-S testing

As argued above, statistical adequacy renders the relevant error probabilities *ascertainable* by ensuring that the *nominal* error probabilities are approximately equal to the *actual* ones. Spanos and McGuirk (2001) demonstrated that even seemingly minor departures from the assumptions of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ can have devastating effects on the reliability of inference; see also Spanos (2009). In light of these, why is there such unwillingness to secure statistical adequacy using M-S testing in applied econometrics?

One possible explanation is that M-S testing is invariably viewed as undefendable against several methodological charges including double-use of data, infinite regress, circularity and pre-test bias; see Kennedy (2008). Let us revisit these issues.

6.1 M-S testing and double-use of data

In the context of the error statistical approach it is certainly true that the same data $\mathbf{z}_0$ are being used for two different purposes: (a) to test primary hypotheses in terms of the unknown parameter(s) $\boldsymbol{\theta}$, and (b) to assess the validity of the prespecified model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$, but 'does that constitute an illegitimate double-use of data?' The short answer is *no*, because, *first*, (a) and (b) pose very different questions to data $\mathbf{z}_0$, and *second*, the probing takes place within vs. outside $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$, respectively.

Neyman-Pearson testing assumes that $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is adequate, and poses questions within its boundaries. In contrast, the question posed by M-S testing is whether or not the particular data $\mathbf{z}_0$ constitute a '*truly typical realization*' of the stochastic mechanism described by $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$, and the probing takes place outside its boundaries, i.e. in $[\mathcal{P}(\mathbf{z})-\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})]$; see Spanos (2000). Indeed, one can go as far as to argue that the answers to the questions posed in (a) and (b) rely on distinct information in $\mathbf{z}_0$.

Spanos (2007b) showed that, for many statistical models, including the simple Normal and the Normal/Linear Regression (table 2) models, M-S testing can be based solely on a *maximal ancillary* statistic $\mathbf{R}(\mathbf{Z}):=(R_1,...,R_{n-m})$, which is independent of a *complete sufficient* statistic $\mathbf{S}(\mathbf{Z}):=(S_1,...,S_m)$ used solely for *primary inferences*. This is the case when the distribution of the sample $f(\mathbf{z};\boldsymbol{\theta})$ simplifies as follows:

$$f(\mathbf{z};\boldsymbol{\theta})=|J|\cdot f(\mathbf{s},\mathbf{r};\boldsymbol{\theta})=|J|\cdot f(\mathbf{s};\boldsymbol{\theta})\cdot f(\mathbf{r}),\ \forall\,(\mathbf{s},\mathbf{r})\in\mathbb{R}_s^m\times\mathbb{R}_r^{n-m}, \tag{47}$$

where $|J|$ denotes the Jacobian of the transformation $\mathbf{Z}\rightarrow(\mathbf{S}(\mathbf{Z}),\ \mathbf{R}(\mathbf{Z}))$. This means that all primary inferences can be based exclusively on $f(\mathbf{s};\boldsymbol{\theta})$, and $f(\mathbf{r})$ (free of $\boldsymbol{\theta}$) can be used to appraise the validity of the statistical model in question.

Example 4 (continued). For $Z_k\backsim\mathsf{NIID}(\mu,\sigma^2),\ k=1,...,n$, the minimal sufficient statistic is $\mathbf{S}:=(\overline{Z}_n,s^2)$ and the maximal ancillary statistics is $\mathbf{R}(\mathbf{Z})=(\widehat{v}_3,..,\widehat{v}_n)$, where $\widehat{v}_k=(\sqrt{n}(Z_k-\overline{Z}_n)/s)$, $k=1,2,..,n$, are known as the *studentized* residuals.

This view calls into question the argument by Claeskens & Hjort (2008), p. xi:

"Uncertainties involved in the first step [specification] must be taken into account when assessing distributions, confidence intervals, etc. That such themes have been largely underplayed in theoretical and practical statistics has been called 'the quiet scandal of statistics.' ... Model averaging can help to develop methods for better assessment and better construction of confidence intervals, p-values, etc."

As argued above, there is nothing scandalous about separating the two steps, it's a matter of judicious modeling. Allowing specification errors to vitiate inferential error probabilities will derail any learning from data about the underlying mechanism, and no amount of averaging over misspecified models can redeem the reliability of inference. Even in the best case scenario where $\mathcal{M}_{\boldsymbol{\phi}}(\mathbf{z})=\lambda\mathcal{M}_{\boldsymbol{\theta}_0}(\mathbf{z})+(1\text{-}\lambda)\mathcal{M}_{\boldsymbol{\theta}_1}(\mathbf{z})$, $0<\lambda<1$, $\mathcal{M}_{\boldsymbol{\theta}_0}(\mathbf{z})$ is statistically adequate but $\mathcal{M}_{\boldsymbol{\theta}_1}(\mathbf{z})$ is misspecified, the result of averaging is to ruin the reliability of inference based on $\mathcal{M}_{\boldsymbol{\theta}_0}(\mathbf{z})$; $\mathcal{M}_{\boldsymbol{\phi}}(\mathbf{z})$ is a misspecified model. Learning from data can only occur when the inferential error probabilities relate directly to an adequate description of the underlying mechanism, and not when the assumed model $\mathcal{M}_{\boldsymbol{\phi}}$ includes $\mathcal{M}_{\boldsymbol{\theta}_1}(\mathbf{z})$ which could not have contributed in generating data $\mathbf{z}_0$. This is not to deny that model averaging can play a role in the context of substantive adequacy, when dealing with statistically adequate models based on different data$\{\mathcal{M}_{\boldsymbol{\theta}_i}(\mathbf{z}_i), i=1,...,m\}$, e.g. regressions with different regressors.

6.2 M-S testing and infinite regress/circularity charges

The *infinite regress* charge is often articulated by claiming that each M-S test relies on a set of assumptions, and thus it assesses the assumptions of the model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ by invoking the validity of its own assumptions, trading one set of assumptions with another *ad infinitum*. Indeed, this reasoning is often *circular* because some M-S tests inadvertently assume the validity of the very assumption being tested!

A closer look at the reasoning underlying M-S testing reveals that both charges are misplaced. *First*, the scenario used in evaluating the type I error invokes no assumptions beyond those of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$, since every M-S test is evaluated under:

$$H_0\text{: } \textit{all} \text{ the probabilistic assumptions of } \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z}) \text{ are valid.}$$

Example 7. The *runs test,* using the residuals from an AR(p) model $\{\widehat{\varepsilon}_t,\ t=1,2,...,n\}$, is an example of an omnibus M-S test for assumptions [4]-[5] (table 5) based a test statistic: $Z_R(\mathbf{Y})=[R-E(R)]/\sqrt{Var(R)}$; see Spanos (1999). For $n\geq 40$, the type I error probability evaluation is based on:

$$Z_R(\mathbf{Y})=\frac{R-([2n-1]/3)}{\sqrt{[16n-29]/90}}\overset{[1]\text{-}[5]}{\backsim}\mathsf{N}(0,1).$$

Second, the type II error (and power), for any M-S test, is determined by evaluating the test statistic under certain forms of departures from the assumptions being appraised [no circularity], but retaining the rest of the model assumptions, or choose M-S tests which are insensitive to departures from the retained assumptions.

For the runs test, the evaluation under the alternative takes the form:

$$Z_R(\mathbf{Y})\overset{[1]-[3]\&\overline{[4]}\text{-}\overline{[5]}}{\backsim}\mathsf{N}(\delta,\tau^2),\ \delta\neq 0,\ \tau^2>0,$$

where $\overline{[4]}$ and $\overline{[5]}$ denote specific departures from these assumptions considered by the test in question; note that the runs test is *insensitive* to departures from Normality. The type of departures implicitly or explicitly considered by the M-S test in question will affect the power of the test in a variety of ways, and one needs to apply a battery of different M-S tests to ensure broad probing capacity and self-correcting in the sense that the effect of any departures from the maintained assumptions is also detected.

In practice, potential problems such as circular reasoning, inadequate probing and erroneous diagnoses can be circumvented by employing:

(a) Judicious combinations of parametric, non-parametric, omnibus and simulation-based tests, probing as broadly as possible and invoking dissimilar assumptions.

(b) Astute *ordering* of M-S tests so as to exploit the interrelationship among the model assumptions with a view to 'correct' each other's diagnosis.

(c) *Joint M-S tests* (testing several assumptions simultaneously) designed to avoid 'erroneous' diagnoses as well as minimize the maintained assumptions.

These strategies enable one to argue with severity that when no departures from the model assumptions are detected, the validated model provides a reliable basis for appraising substantive claims; see Spanos (2000), Mayo and Spanos (2004).

6.3 M-S testing/respecification and pre-test bias

The question sometimes raised is whether the above error statistical strategies of M-S testing and respecification are vulnerable to the charge of *pre-test bias*. To discuss the merits of this charge, consider the Durbin-Watson test for assessing the assumption of no autocorrelation for the linear regression errors, based on (see Greene, 2003):

$$H_0 : \rho = 0, \text{ vs. } H_1 : \rho \neq 0.$$

Step 1. The pre-test bias perspective interprets this M-S test as equivalent to choosing between the following two models:

$$\begin{array}{ll} \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}): & y_t = \beta_0 + \beta_1 x_t + u_t, \\ \mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z}): & y_t = \beta_0 + \beta_1 x_t + u_t,\ u_t{=}\rho u_{t-1} + \varepsilon_t. \end{array} \tag{48}$$

Step 2. This is then formalized into a choice between two estimators of β_1 in decision-theoretic terms using the *pre-test estimator*:

$$\ddot{\beta}_1{=}\lambda\widehat{\beta}_1 + (1{-}\lambda)\widetilde{\beta}_1, \text{ where } \lambda{=}\begin{cases} 1, & \text{if } H_0 \text{ is accepted} \\ 0, & \text{if } H_0 \text{ is rejected;} \end{cases} \tag{49}$$

$\widehat{\beta}_1$ is the OLS estimator under H_0, and $\widetilde{\beta}_1$ is the GLS estimator under H_1.

Step 3. This perspective claims that the relevant error probabilities revolve around the Mean Square Error (MSE) of $\ddot{\beta}_1$, whose sampling distribution is usually non-Normal, biased and has a highly complicated variance; see Leeb and Pötscher (2005).

When viewed in the context of the error-statistical approach, the pre-test bias argument, based on (49), seems highly questionable on several grounds.

First, it misconstrues M-S testing by recasting it as a decision-theoretic estimation problem based on a loss function. As argued discerningly by Hacking (1965), pp. 31:

"Deciding that something *is* the case differs from deciding to *do* something."

M-S testing poses the canonical question whether $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ *is* statistically adequate, i.e. it accounts for the chance regularities in data $\mathbf{z}_0$ or not; it is not concerned with selecting one of two models come what may. Having said that, one can potentially construct a M-S test with a view to assess a subset of the model assumptions by viewing an alternative model $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z})$ as a result of narrowing $[\mathcal{P}(\mathbf{z}){-}\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})]$ (see (46)) down to a single alternative model which (parametrically) encompasses $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$; see Spanos (1999). As argued in section 3.4, however, when the ultimate inference is

concerned with whether $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is statistically adequate, the relevant errors are:

(i) the selected model is inadequate but the other model is adequate, or

(ii) both models are inadequate.

In contrast, $E(\ddot{\beta}_1-\beta_1)^2$ evaluates the expected loss resulting from the modeler's supposedly tacit intention to use $\ddot{\beta}_1$ as an estimator of β_1. Is there a connection between $E(\ddot{\beta}_1-\beta_1)^2$, for all $\beta_1\in\mathbb{R}$, and the errors (i)-(ii)? The short answer is none. The former evaluates the expected loss stemming from one's (misguided) *intentions*, but the latter pertain to the relevant error probabilities (type I & II) associated with the inference that one of the two models is statistically adequate; Spanos (2009).

Second, the case where a M-S test supposedly selects the alternative ($\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z})$), the implicit inference is that $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z})$ is statistically adequate; the raison d'être of model validation. This constitutes a classic example of *the fallacy of rejection* [evidence *against* H_0 is misinterpreted as evidence *for* H_1]. The validity of $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z})$ needs to be established separately by testing its own assumptions. Hence, in a M-S test one should *never* accept the alternative without further testing; see Spanos (2000).

Third, the case where a M-S test supposedly selects the null ($\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$), the implicit inference is that $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is statistically adequate. This inference is problematic for two reasons. First, given the multitude of assumptions constituting a model, there is no single M-S test based on a parametrically encompassing model $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z})$, that could, by itself, establish the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$. Second, the inference is vulnerable to *the fallacy of acceptance* [*no* evidence against H_0 is misinterpreted as evidence *for* it]. It is possible that the particular M-S test did not reject $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ because it had very low power to detect an existing departure. In practice this can be remedied using additional M-S tests with higher power to cross-check the results, or/and use a post-data evaluation of inference to establish the warranted discrepancies from H_0.

In summary, instead of devising ways to circumvent the fallacies of rejection and acceptance to avoid erroneous inferences, the pre-test bias argument embraces them by recasting the original problem (in step 1), formalizes them (in step 2), and evaluates risks (in step 3) that have no bearing on erroneously inferring that the selected model is statistically adequate. The pre-test bias charge is ill-conceived because it misrepresents model validation as a choice between two models come what may.

7 Summary and conclusions

Akaike-type model selection procedures often give rise to unreliable inferences primarily because they: (a) assume away the problem of statistical model specification, and (b) ignore the relevant error probabilities for the inferences reached.

Both problems can be addressed in the context of the error statistical framework, which emphasizes the probing of the different ways an inference might be in error. Using statistical adequacy as *the* sole criterion for assessing when a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ 'accounts for the chance regularities in data $\mathbf{z}_0$', renders the relevant error probabilities ascertainable and can obviate statistical misspecification, thus securing the reliability of inference. The key is provided by viewing $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ in terms of the

probabilistic structure of the process $\{\mathbf{Z}_k, k\in\mathbb{N}\}$ underlying data $\mathbf{z}_0$, rendering its assumptions testable and nesting parametrically the structural model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$.

Trenchant M-S testing and informed respecification can secure the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$, which can be subsequently used as a basis for reliable inductive inference in probing substantive questions of interest; the disentanglement of the two facets being particularly important to prevent specification errors from vitiating the inferential error probabilities and thus forestall any learning from data.

References

[1] Akaike, H. (1970), "Statistical Predictor for Identification," *Annals of the Institute of Statistical Mathematics*, **22**: 203-217.

[2] 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.

[3] Allen, D. M. (1971), "Mean square error of prediction as a criterion for selecting variables," *Technometrics*, **13**: 469-475.

[4] Andreou, E. and A. Spanos (2003), "Statistical adequacy and the testing of trend versus difference stationarity", *Econometric Reviews*, **22**: 217-252.

[5] Burnham, K. P., and D. R. Anderson (2002), *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed., Springer, NY.

[6] Cheney, E. W. (1982), *Introduction to Approximation Theory*, AMS Chelsea Publishing, RI.

[7] Claeskens, G. and N. L. Hjort (2008), *Model Selection and Model Averaging,* Cambridge University Press, Cambridge.

[8] Cox, D. R. and D. G. Mayo (2009), "Objectivity and Conditionality in Frequentist Inference," in Mayo, D. G and A. Spanos (eds), *Error and Inference*, forthcoming Cambridge University Press, Cambridge.

[9] Doob, J. L. (1953), *Stochastic Processes*, Wiley, New York.

[10] Fisher, R. A. (1922), "On the mathematical foundations of theoretical statistics", *Philosophical Transactions of the Royal Society A*, **222**: 309-368.

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

[12] Fisher, R. A. (1955), "Statistical Methods and Scientific Induction," *Journal of The Royal Statistical Society,* B, **17**: 69-78.

[13] 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.

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

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

[16] Hacking, I. (1965), *Logic of Statistical Inference*, Cambridge University Press, Cambridge.

[17] 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.

[18] Hildebrand, F. B. (1982), *Introduction to Numerical Analysis*, Dover, NY.

[19] Kennedy, P. (2008), *A Guide to Econometrics*, 6th ed., MIT Press, Cambridge, MA.

[20] 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.

[21] Leeb, H. and B. M. Pötscher (2005), "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, **21**: 21-59.

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

[23] Li, Q. and J. S. Racine (2006), *Nonparametric Econometrics: Theory and Practice*, Princeton University Press, Princeton, NJ.

[24] Luenberger, D. G. (1969), *Optimization by Vector Space Methods*, Wiley, NY.

[25] Lutkepohl, H. (2005), *New Introduction to Multiple Time Series Analysis*, Springer, NY.

[26] Mallows, C. L. (1973), "Some Comments on C_p," *Technometrics*, 15, 661-675.

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

[28] Mayo, D. G. and D. R. Cox (2006), "Frequentist statistics as a theory of inductive inference," pp. 77-97 in *Optimality: The Second Erich L. Lehmann Symposium*, edited by J. Rojo, Lecture Notes-Monograph Series, vol. 49, Institute of Mathematical Statistics.

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

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

[31] Mayo, D. G. and Spanos, A. (2009), "Error Statistics," forthcoming in *Philosophy of Statistics, Handbook of Philosophy of Science*, Elsevier, (eds.) D. Gabbay, P. Thagard, and J. Woods.

[32] McQuarrie, A. D. R. and C-L. Tsai (1998), *Regression and Time Series Model Selection*, World Scientific, NJ.

[33] Neyman, J. (1956), "Note on an Article by Sir Ronald Fisher," *Journal of the Royal Statistical Society*, B, **18**: 288-294.

[34] Pearson, K. (1920), "The Fundamental Problem of Practical Statistics," *Biometrika*, **XIII**: 1-16.

[35] Powell, M. J. D. (1981), *Approximation Theory and Methods*, Cambridge University Press, Cambridge.

[36] Rao, C. R. and Y. Wu (2001), "On Model Selection," pp. 1-64 in *Model Selection*, ed. by P. Lahiri, Institute of Mathematical Statistics, Lecture Notes-Monograph series, vol. 38, Beachwoodm OH.

[37] Rissanen, J. (1978), "Modeling by the shortest data description," *Automatica*, **14**: 465-471.

[38] Rivlin, T. J. (1981), *An Introduction to the Approximation of Functions*, Dover, NY.

[39] Skyrms, B. (2000), *Choice and Chance: an introduction to inductive logic*, 4th ed., Wadsworth, US.

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

[41] Shao, J. (2003), *Mathematical Statistics*, 2nd ed., Springer, NY.

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

[43] Spanos, A. (1990), "The Simultaneous Equations Model revisited: statistical adequacy and identification", *Journal of Econometrics*, **44**: 87-108.

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

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

[46] Spanos, A. (2000), "Revisiting Data Mining: 'hunting' with or without a license," *The Journal of Economic Methodology*, **7**: 231-264.

[47] Spanos, A. (2006a), "Where Do Statistical Models Come From? Revisiting the Problem of Specification," pp. 98-119 in *Optimality: The Second Erich L. Lehmann Symposium*, edited by J. Rojo, Lecture Notes-Monograph Series, vol. 49, Institute of Mathematical Statistics.

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

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

[50] Spanos, A. (2007a), "Curve-Fitting, the Reliability of Inductive Inference and the Error-Statistical Approach," *Philosophy of Science*, **74**: 1046–1066.

[51] Spanos, A. (2007b), "Sufficiency and Ancillarity Revisited: Testing the Validity of a Statistical Model" Working Paper, Virginia Tech.

[52] Spanos, A. (2009), "Philosophy of Econometrics," forthcoming in the *Handbook of the Philosophy of Science*, edited by D. Gabbay, P. Thagard, and J. Woods, Elsevier, North Holland.

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

[54] Watson, G. A. (1980), *Approximation Theory and Numerical Methods*, Wiley, NY.

8 Appendix - Mathematical approximation theory: a brief summary of relevant results

Mathematically the most effective way to render the curve-fitting problem tractable is to view it in the context of a *normed linear space*, say $\mathcal{L}$, where the 'true' function $h \in \mathcal{L}$, and the approximating function $g_m(.)$ belongs to a subset $\mathcal{G} \subset \mathcal{L}$. The notion of 'best' is defined in terms of a **norm** $\| \, . \, \| : \mathcal{L} \to [0, \infty)$, and satisfies certain properties generalizing the notion of length; see Luenberger (1969).

Example 8. Let $C[a,b]$ consist of all real-valued continuous functions on the real interval $[a,b] \subset \mathbb{R}$, together with the ∞-norm:

$$\| \, f \, \|_\infty = \max_{a \leq f \leq b} | \, f \, | \, . \tag{50}$$

It can be shown that $(C[a,b], \| \, . \, \|_\infty)$ defines a normed linear space. Other norms in the context of $C[a,b]$ are special cases of:

$$\| \, f \, \|_p = \left(\int_a^b |f(x)|^p \, dx \right)^{\frac{1}{p}} , \; p \geq 1. \tag{51}$$

The norm defines a *distance* between any two elements (f,g) in $\mathcal{L}$ via the *metric*:

$$d(f,g) = \| \, f - g \, \| \text{ for any } (f,g) \in \mathcal{L},$$

where $d(.,.) : (\mathcal{L} \times \mathcal{L}) \to [0, \infty)$, and satisfies certain properties. The pair $(\mathcal{L}, \; d(.,.))$ defines a *metric space* induced by $(\mathcal{L}, \| \, . \, \|)$. Similarly, one can define the notion of an *inner (scalar) product* via:

$$\sqrt{\langle f, f \rangle} = \| \, f \, \|, \text{ for any } f \in \mathcal{L},$$

inducing an *inner product space* $(\mathcal{L}, \; \langle ., . \rangle)$, where $d(f,g) = \langle f - g, f - g \rangle^{\frac{1}{2}}$; see Powell (1981). The additional mathematical structure induced by an inner product (to define angles, and thus orthogonality via $\langle f, g \rangle = 0$) is needed when the problem requires one to go beyond existence and uniqueness results to construct the approximating function $g_m(.)$ explicitly.

Example 9. For the normed linear space $(C[a,b], \| \, . \, \|_\infty)$, the induced metric space is $(C[a,b], d_\infty(.,.))$, where $d_\infty(.,.) = \max_{a \leq x \leq b} |f(x) - g(x)|$.

In a normed linear space one can pose three interrelated questions of interest:

(i) Does there *exist* in $\mathcal{G}$ a best approximation $g^*(x)$ of $f(x)$?

(ii) When it exists, is $g^*(x)$ *unique*?

(iii) How can one *construct* $g^*(x)$?

(iv) How *adequate* is the approximation rendered by the constructed $g^*(x)$?

Let us consider each of these issues briefly.

The *existence result* is easy to ensure when $\mathcal{G}$ is chosen to be a *compact* subset of a normed linear space $\mathcal{L}$. The cornerstone of such existence results is a famous theorem whose general form is.

Theorem 1. Let $f(x)$ be an upper semicontinuous functional on $\mathcal{G}$, a compact subset of a normed linear space $(\mathcal{L}, \| \, . \, \|)$, then $f(x)$ achieves a maximum on $\mathcal{G}$; see Luenberger (1969). A special case of this theorem is known as Weierstrass' theorem which ensures that when $h(x)$ is in $(C[a,b], \| \, . \, \|_\infty)$, for any $\epsilon > 0$ there exists an integer $N(\epsilon)$ such that for $m > N(\epsilon)$:

$$|h(x) - g_m(x; \boldsymbol{\alpha}_m^*)| < \epsilon, \text{ for all } x \in [a,b], \tag{52}$$

where $g_m(x; \boldsymbol{\alpha}_m) = \sum_{i=0}^m \alpha_{im} x^i$, $m \geq 1$. The norm underlying (52) is the $\infty-$norm in (50), giving rise to the metric:

$$d_\infty(h(x), g_m(x; \boldsymbol{\alpha}_m)) = \max_{a \leq x \leq b} |h(x) - g_m(x; \boldsymbol{\alpha}_m)|$$

and the mode of convergence is known as *uniform*. Note that the coefficients $\boldsymbol{\alpha}_m := (\alpha_{0m}, \alpha_{1m}, ..., \alpha_{mm})$ depend crucially on m, and as the degree of the polynomial increases these coefficients change with it.

The *uniqueness result* depends crucially on the convexity of both $\mathcal{G}$ and $\| \, . \, \|$.

Theorem 2. Let $\mathcal{G}$ be a convex set in a normed linear space $(\mathcal{L}, \| \, . \, \|)$, whose norm is strictly convex. Then for all $f \in \mathcal{L}$, there is a unique best approximation in $\mathcal{G}$.

Example 8 (continued). In the case of $C[a,b]$, when $\mathcal{G}$ is a finite dimensional linear subspace (convex set), the 2-norm:

$$\| \, f \, \|_2 = \left(\int_a^b |f(x)|^2 \, dx \right)^{\frac{1}{2}},$$

ensures uniqueness because $\| \, f \, \|_2$is strictly convex. However, the $\infty-$norm in (50) or the 1-norm:

$$\| \, f \, \|_1 = \left(\int_a^b |f(x)| \, dx \right),$$

are *not* strictly convex. Note that in general (Powell, 1981):

$$\| \, f \, \|_1 \leq (b-a)^{\frac{1}{2}} \, \| \, f \, \|_2 \leq (b-a) \, \| \, f \, \|_\infty \text{ for all } f \in C[a,b].$$

This means that in the case of the normed linear spaces $(C[a,b], \| \, . \, \|_1)$ and $(C[a,b], \| \, . \, \|_\infty)$ one needs to impose further restrictions on $\mathcal{G}$ or $h(x)$ to ensure uniqueness. One such restriction on a linear subspace $\mathcal{G}$ of $C[a,b]$, $\dim(\mathcal{G}) = m+1$, is the *Haar condition.*

Haar condition. For any $\phi(x) \in \mathcal{G}$, that is not the zero element, the number of roots of the equation $\{\phi(x) = 0, \; x \in [a,b]\}$ is at most m. In the case where $\mathcal{G}$ is a set of polynomials of the form $g_m(x; \boldsymbol{\alpha}) = \sum_{i=0}^m \alpha_i x^i$, this condition is satisfied because $\dim(\mathcal{G}) = m+1$ and $g_m(x; \boldsymbol{\alpha})$ can have at most m distinct zeros.

Chebyshev set. A closely related condition is to exchange the base functions $\{1, x, x^2, ..., x^m\}$ with a *Chebyshev set* of generalized polynomials: $\{\phi_i(x), \; i = 0, 1, ..., m\}$, defined on $[a,b]$, such that every non-trivial linear combination $g_m(x; \boldsymbol{\alpha}) = \sum_{i=0}^m \alpha_i \phi_i(x)$ has at most m distinct zeros on $[a,b]$; $\alpha_0 = \cdots = \alpha_m = 0$ is the trivial case; Powell (1981).

Necessary and sufficient conditions. In the context of the normed linear space $(C[a,b], \| \, . \, \|_\infty)$, where a Chebyshev set $\{\phi_i(x), \; i = 0, 1, ..., m\}$ spans $\mathcal{G}_m$, the

Haar condition ensures that the best approximating polynomial $g_m(x;\boldsymbol{\alpha}^*)$ is unique if and only if the error function: $\varepsilon(x;m) = h(x) - g_m(x;\boldsymbol{\alpha}^*),\ x{\in}[a,b]$, changes sign more than $m+1$ times as x varies over $[a,b]$; see Cheney (1982). It is important to emphasize that this result does not say that points $\{x_{[k]},\ k{=}0,1,...,m\}$, where the successive change of sign in the residuals $\{\varepsilon(x_{[k]};m),\ k=0,1,...,m\}$ occurs, is unique, or that it does not happen more than $m{+}1$ times.

Theorem 3 - Oscillation. Let $\mathcal{G}$ be a finite dimensional linear space in a normed linear space $(\mathcal{L}, \| \ . \ \|)$ that satisfies the Haar condition. For the best approximation $g_m(x;\boldsymbol{\alpha}^*)$ on $[a,b]$ to $h(x)$ in $\mathcal{G}$, there exist $m{+}2$ points $\{x_{[k]},\ k=0,1,...,m+1\}$: $a \le x_{[0]} \le x_{[1]} \cdots \le x_{[m+1]} \le b$, such that the error function $\varepsilon(x_{[k]};m) = h(x_{[k]}) - g_m(x_{[k]};\boldsymbol{\alpha}^*)$, satisfies the condition (see Powell, 1981):

(O) $\varepsilon(x_{[k]};m)$, $k=0,1,...,m$ alternate in sign at least $m+2$ times.

Piecewise approximation. Of particular interest in mathematical approximation theory are the results pertaining to local (piecewise) approximation using splines, where the interval $[a,b]$ is partitioned at $N{+}2$ knots $a{=}x_0 < x_1 < x_2 < \cdots < x_N < x_{N+1}{=}b$, and a polynomial of degree m, say $g_m(x)$ is fitted over each interval $[x_k, x_{k+1}]$, $k=0,1,...,N$. The most widely used piecewise polynomials are *splines* of degree m (often $m=2$), which have $(m-1)$ continuous derivatives at the knots. Analogous oscillation theorems are also applied to the case of local (piecewise) approximation using splines; see Watson (1980), pp. 157-171.

It turns out that in cases where the approximating function $g^*(x)$ is unique, it can be represented by a mapping from $\mathcal{L}$ to $\mathcal{G}$, say $g^*(x) = \mathsf{P}_{\mathcal{G}}(h(x))$, and the structure of $\mathsf{P}_{\mathcal{G}}(.)$ is of value in considering the question of constructing such best approximations. This mapping is often a *linear projection operator* which is characterized by the property that:

$$\mathsf{P}_{\mathcal{G}}[\mathsf{P}_{\mathcal{G}}(f(x))] = \mathsf{P}_{\mathcal{G}}(f(x)),\ \text{for all } f \in \mathcal{L}. \tag{53}$$

A sufficient condition for $\mathsf{P}(.)$ to be a projection is to satisfy the condition:

$$\mathsf{P}_{\mathcal{G}}[g(x)] = g(x),\ \text{for all } g \in \mathcal{G}. \tag{54}$$

Theorem 4. Let $\mathcal{G}$ be a finite dimensional linear space in a normed linear space $(\mathcal{L}, \| \ . \ \|)$, such that for every $f \in \mathcal{L}$, there is a unique best approximation in $\mathcal{G}$, say $\mathsf{P}_{\mathcal{G}}(f)$, then the operator $\mathsf{P}_{\mathcal{G}}$ is continuous. Moreover, when $\mathsf{P}(f)$ satisfies condition (54), then for $d^* = \min_{f \in \mathcal{G}} \| \ f - g \ \|$, the error of the approximation $\mathsf{P}_{\mathcal{G}}(f)$ satisfies the bound (Powell, 1981): $\| \ f - \mathsf{P}_{\mathcal{G}}(f) \ \| \ \ \le [1 + \| \ \mathsf{P}_{\mathcal{G}} \ \|] d^*$.

This result is of interest in constructing $g^*(x) = \mathsf{P}_{\mathcal{G}}(f)$ because rounding errors can have substantial effects on the constructed approximations if $\mathsf{P}_{\mathcal{G}}$ is discontinuous. Viewing the approximating function as a projection also sheds additional light on both questions (iii)-(iv).

(iii) *Constructing* an approximating function often involves (a) the choice of the appropriate family of building block functions spanning $\mathcal{G}$ which, for reasons of existence and uniqueness, often take the form:

$$g_m(x;\boldsymbol{\alpha}) = \textstyle\sum_{i=0}^{m} \alpha_i \phi_i(x), \tag{55}$$

where $\boldsymbol{\alpha} := (\alpha_0, \alpha_1, ..., \alpha_m)$, and $\{\phi_i(x),\ i{=}0, 1, ...m\}$ is a *base (Chebyshev) set* of generalized polynomial functions, and (b) an algorithm that chooses the best approximating function within this family; see Powell (1981).

(iv) Assessing the *adequacy* of the best approximating function involves constructing upper bounds for the approximating error or even deriving explicit forms of the error function under certain circumstances where additional smoothness conditions are imposed on $h(x)$. Typical results in this context are the following.

Theorem 5. Consider the problem of approximating $h(x)$, an element of $(C[a,b], \| \ . \ \|_\infty)$, using $g_m(x; \boldsymbol{\alpha})$ in (55), and define the *modulus of continuity* of $h(x)$:

$$\omega(\delta) = \sup_{|x_1 - x_2| \leq \delta} |h(x_1) - h(x_2)|, \text{ for } (x_1, x_2) \in [a,b],\ \delta > 0.$$

The approximation error $\varepsilon(x; m)$ satisfies the Jackson-type upper bound:

$$\varepsilon(x; m) \ \leq \ 6\omega(\delta) \left(\tfrac{b-a}{2m}\right) \tag{56}$$

For more accurate bounds one needs to impose additional smoothness conditions on $h(x)$, such as the existence of derivatives up to order k. Such bounds are useful in appraising the behavior of $\varepsilon(x; m)$ as $m \to \infty$, and hence the appropriateness of the functions spanning $\mathcal{G}$; see Rivlin (1981), Cheney (1982).

In the case of piecewise (local) approximation, such as splines, these Jackson-type upper bounds also depend on the length of the intervals $[x_k, x_{k+1}]$, say $h_k = [x_{k+1} - x_k]$, $k = 0, 1, ..., N$. When one defines a uniform partition,i.e. $x_{k+1} = hx_k$, the length h is known as the smoothing parameter which plays an important role in convergence results; see Watson (1980).

Theorem 6. Let $h(x) \in C[a,b]$. 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:

$$g_m(x; \boldsymbol{\alpha}) = \textstyle\sum_{i=0}^m y_i \prod_{j=0, j\neq i}^m \left(\frac{x - x_j^*}{x_i^* - x_j^*}\right), \quad x \in [a,b], \tag{57}$$

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:

$$\varepsilon(x, m) = \tfrac{d^{m+1} h(\xi)}{d^{m+1} x} \tfrac{1}{(m+1)!} \textstyle\prod_{j=0}^m \left(x - x_j^*\right), \ \xi \in [a,b]. \tag{58}$$

(58) suggests that the error curve $\varepsilon(x, m)$ behaves like a polynomial in x, with $m+1$ roots $(x_0^*, x_1^*, \cdots, x_m^*)$: $\varepsilon(x, m) = ax^{m+1} + b_m x^m + \cdots + b_1 x + b_0$.
Such an *oscillating curve* is also typical for error term arising from the least squares approximation; see Watson (1980).