Received: 17 December 2020 | Revised: 26 September 2021 | Accepted: 28 October 2021

DOI: 10.1111/joes.12483

JOURNAL OF ECONOMIC SURVEYS WILEY

ARTICLE

# Statistical modeling and inference in the era of Data Science and Graphical Causal modeling

**Aris Spanos**

Department of Economics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, USA

**Correspondence**
Aris Spanos, Department of Economics, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA.
Email: aris@vt.edu

**Abstract**

The paper discusses four paradigm shifts in statistics since the 1920s with a view to compare their similarities and differences, and evaluate their effectiveness in giving rise to 'learning from data' about phenomena of interest. The first is Fisher's 1922 recasting of Karl Pearson's descriptive statistics into a model-based $[\mathcal{M}_\theta(\mathbf{x})]$ statistical induction that dominates current statistics (frequentist and Bayesian). A crucial departure was Fisher's replacing the curve-fitting perspective guided by goodness-of-fit measures with a model-based perspective guided by the *statistical adequacy*: the validity of the probabilistic assumptions comprising $\mathcal{M}_\theta(\mathbf{x})$. Statistical adequacy is pivotal in securing trustworthy evidence since it underwrites the reliability of inference. The second is the nonparametric turn in the 1970s aiming to broaden $\mathcal{M}_\theta(\mathbf{x})$ by replacing its distribution assumption with weaker mathematical conditions relating to the unknown density function underlying $\mathcal{M}_\theta(\mathbf{x})$. The third is a two-pronged development initiated in Artificial Intelligence (AI) in the 1990s that gave rise to Data Science (DS) and Graphical Causal (GC) modeling. The primary objective of the paper is to compare and evaluate the other competing approaches with a refined/enhanced version of Fisher's model-based approach in terms of their effectiveness in giving rise to genuine "learning from data;" excellent

  

goodness-of-fit/prediction is neither necessary nor sufficient for statistical adequacy, or so it is argued.

**KEYWORDS**
Big Data, data patterns, Data Science, functional analysis, Graphical Causal modeling, Machine Learning, mathematical approximation, model-based inference, probably approximately correct (PAC), statistical adequacy, Statistical Learning Theory, trustworthy evidence

## 1 | INTRODUCTION: PARADIGM SHIFTS IN STATISTICS

Statistical modeling and inference have gone through several major changes during the 20th century, including radical paradigm shifts, that have often been overlooked by practitioners. The initial paradigm shift was Fisher (1922) recasting of Karl Pearson's descriptive statistics into a model-based statistical induction, which was subsequently extended by Neyman and Pearson (1933) and Neyman (1937). The next major shift was the *nonparametric* turn in the late 1970s aiming to broaden the scope of the Fisher–Neyman–Pearson (F–N–P) model-based approach. The last two paradigm shifts came in the form of Data Science (DS) and Graphical Causal (GC) modeling. This neglect resulted in an overlap of different, and often incompatible perspectives on statistics, creating endless confusions about the nature and proper interpretation of inferential claims associated with different perspectives, and how to secure the reliability of inference and the trustworthiness of ensuing empirical evidence.

The discussion that follows sketches four paradigm shifts in statistics since the 1920s and focuses, not on the usefulness of their methods in general, but their effectiveness (optimality and reliability) in accomplishing the primary objective of "*learning from data about phenomena of interest*." The notion of a paradigm is used broadly to denote a conceptual framework that includes theories, beliefs, values, research methods, and standards for what constitutes genuine contributions to a field as established by its professional and educational structure; see Kuhn (1970).

The primary objective of the discussion that follows is threefold. First, in an attempt to address the data-driven versus theory-driven quandary and the trustworthiness of evidence problem, a refined/enhanced version of the F–N–P frequentist model-based approach, as enunciated in Spanos (2006), is articulated to provide the broad framework for evaluating the other competing perspectives. The refinement of the F–N–P approach comes in the form of separating the *modeling* from the *inference facet*, and the enhancement in the form of distinguishing, ab initio, between the *statistical* (data-based) and the *substantive* (theory-based) information/model.

Second, the proposed refined/enhanced F–N–P approach is used to evaluate the other competing perspectives to statistics in terms of their effectiveness (reliability and precision) in giving rise to genuine "learning from data" by securing the trustworthiness of their ensuing evidence about phenomena of interest; reliability of inference refers to the actual error probabilities approximating the nominal (assumed) ones closely, and precision refers to the optimality of such procedures with a sufficiently large sample size. A key argument is that "learning from data" stems from trustworthy evidence stemming from statistically valid inductive premises.

Third, to argue that DS and GC modeling can achieve their potential constructive roles in "learning from data" more effectively when integrated within the proposed refined/enhanced

F–N-P model-based statistical framework where the modeling and inference are guided by statistical adequacy that secures the reliability of inference and the trustworthiness of the ensuing evidence. In contrast, excellent goodness-of-fit/prediction is neither necessary nor sufficient for statistical adequacy; see Spanos (2007a).

All approaches to statistics begin with two basic elements: (i) "question(s) of interest" to be posed to (ii) "a data set." The questions are often motivated by some substantive (theory-based) information, however primitive (simple conjectures) or highly sophisticated (a system of stochastic difference equations). The data are chosen to provide trustworthy answers to these questions, that is, genuine *learning from data*. Alternative approaches to statistics differ primarily in terms of the following features:

[a] **Underlying framework**: the underlying probabilistic and broader mathematical framework.
[b] **Inductive premises**: the framing of the inductive premises (probabilistic assumptions imposed on the data), and the interpretation of the assumed model.
[c] **Model choice**: the selection of the "best" model—inductive premises— for the particular data.
[d] **Quantification**: the estimation of the inductive premises using the data.
[e] **Inductive inference**: the underlying inductive reasoning that includes the invoked interpretation of probability and the nature and interpretation of the ensuing inferences.
[f] **Substantive versus statistical**: how the approach conciliates the substantive (theory-based) and statistical (data-based) information.

The first paradigm shift occurred with Fisher (1922) recasting of Karl Pearson's *descriptive statistics* for data $\mathbf{x}_0 := (x_1, \dots, x_n)$ (Yule, 1916) into a *model-based inferential statistics* framed around the concept of a prespecified (parametric) statistical model, generically defined by:

$$\mathcal{M}_\theta(\mathbf{x}) = \{f(\mathbf{x};\theta),\ \theta \in \Theta \subset \mathbb{R}^m\}, \quad \mathbf{x} \in \mathbb{R}^n_X, \quad n > m; \tag{1}$$

$f(\mathbf{x};\theta),\ \mathbf{x} \in \mathbb{R}^n_X$ denotes the joint distribution of the sample $\mathbf{X} := (X_1, \dots, X_n)$, $\mathbb{R}^n_X$ is the sample space, $\Theta$ the parameter space, and $\theta$ denotes the unknown statistical parameters. It is important to emphasize that $f(\mathbf{x};\theta)$ encapsulates the probabilistic assumptions of $\mathcal{M}_\theta(\mathbf{x})$ from three broad categories (Spanos, 1986), [D] Distribution, [M] Dependence and [H] Heterogeneity, imposed on the stochastic process $\{X_t,\ t \in \mathbb{N} := (1, 2 \dots, n, \dots)\}$ underlying data $\mathbf{x}_0$.

**Example 1.** The *simple Normal model* is specified by:

$$\mathcal{M}_\theta(\mathbf{x}) : X_t \backsim \mathsf{NIID}(\mu, \sigma^2), \quad \theta := (\mu, \sigma^2) \in \Theta := (\mathbb{R} \times \mathbb{R}_+), \quad x_t \in \mathbb{R}, \quad t \in \mathbb{N}\}, \tag{2}$$

where "NIID" stands for "Normal [D], Independent [M], and Identically Distributed [H]." Equation (2) and its several variations/extensions provided the relevant statistical models for most of Fisher's theoretical and empirical work; see Fisher (1925, 1935).

The revolutionary nature of Fisher's recasting stems from specifying explicitly the probabilistic assumptions imposed on data $\mathbf{x}_0$, and viewing $\mathcal{M}_\theta(\mathbf{x})$ as a *stochastic mechanism* that could have given rise to data $\mathbf{x}_0$, as opposed to describing $\mathbf{x}_0$ using summary statistics. Fisher replaced Pearson's curve-fitting perspective, guided by goodness-of-fit measures, with a model-based

perspective guided by the *statistical adequacy*: the validity of the probabilistic assumptions comprising $\mathcal{M}_\theta(\mathbf{x})$, for example, NIID in the case of (2). Statistical adequacy is pivotal in securing trustworthy evidence since it underwrites the reliability of inference by ensuring the optimality and reliability of inference procedures, including assuring that the actual error probabilities approximate closely the nominal (assumed) ones. Applying a .05 significance level test when the actual type I error (due to a misspecified $\mathcal{M}_\theta(\mathbf{z})$) is closer to .9, is likely to yield untrustworthy evidence; see Spanos and McGuirk (2001).

The second paradigm shift originated in the *nonparametric* turn in the late 1970s whose primary aim was to broaden the scope of $\mathcal{M}_\theta(\mathbf{x})$ by replacing the direct distribution assumption (Normal, Poisson, etc.) with a family of densities $\mathcal{F}$ specified in terms of *indirect* distribution assumptions, such as: (i) the existence of moments up to order $p \geq 1$, (ii) smoothness restrictions on the *unknown* density function $f(x), x \in \mathbb{R}_X$; see Thompson and Tapia (1990). As stated by Wasserman (2006), p. 1: "The basic idea of nonparametric inference is to use data to infer an unknown quantity while making as few assumptions as possible. Usually, this means using statistical models that are infinite-dimensional." The allure of "making as few assumptions as possible" is highly misleading because nonparametric models retain strong Dependence and Heterogeneity assumptions, often IID. That is, the weakening concerns only the Distribution [D] assumption to render the inferences less vulnerable to such departures. As an aside, it is important to note that the original nonparametric turn of the late 1940s, based on order and rank statistics (Lehmann, 1975), can be viewed as part of the *robustness* literature relating to Fisher's model-based statistics, where IID is retained but Normality might be inappropriate for the particular data $\mathbf{x}_0$; see Spanos (2001).

The third and fourth paradigm shifts in the 1980s and 1990s were initiated in artificial intelligence (AI). The first was *Big Data and* Data Science (DS), which includes Machine Learning (ML), Statistical Learning Theory (SLT), pattern recognition, data mining, and the second was Graphical Causal (*GC) modeling*. These developments created two separate paradigms with their own terminology, procedures, scope, and primary aims, adding further to the overall confusion in statistical modeling and inference.

As perceived by Vapnik (2000), a pioneer in SLT:

"Between 1960 and 1980 a revolution in statistics occurred: Fisher's paradigm, introduced in the 1920s and 1930s, was replaced by a new one. This paradigm reflects a new answer to the fundamental question: what must one know a priori about an unknown functional dependency in order to estimate it on the basis of observations? In Fisher's paradigm, the answer was very restrictive—one must know almost everything. Namely, one must know the desired dependency up to the values of a finite number of parameters. ... In the new paradigm ... it is sufficient to know some general properties of the set of functions to which the unknown dependency belongs." (ix).

The appeal of DS stems primarily from its claim that potential "learning from data" can be impeded, not by the desideratum of reliable substantive (theory-based) subject matter information, but by the lack of ingenuity and coding skills of practitioners in mining large data sets effectively; see Hastie et al. (2017).

In contrast, the literature on GC modeling argues that without substantive causal information, empirical modeling degenerates into fiddling with correlations that yield no real learning from data in the sense of explaining (or understanding) phenomena of interest; see Pearl (2009), Spirtes et al. (2000), Koller and Friedman (2009). Pearl and Mackenzie (2018) call it a "causal revolution" and criticize the DS's data-driven modeling:

"We live in an era that presumes Big Data to be a solution to all our problems. Courses in 'DS' are proliferating in our universities, and jobs for 'data scientists' are lucrative in the companies

that participate in that 'data economy.' But I hope, in this book, to convince you that data are profoundly dumb." (p. 6). "In certain circles there is an almost religious faith that we can find the answers to these questions [Is there a gene that causes lung cancer? What kinds of solar systems are likely to harbor Earth-like planets, etc.] in the data itself, if only we are sufficiently clever at data mining. However, readers of this book will know that this hype is likely to be misguided. The questions that I have just asked are all causal, and causal questions can never be answered by data alone. They require us to formulate a model of the process that generates the data, or at least some aspects of that process." (p. 351).

All three recent paradigm shifts view their proposed perspectives as major improvements over the F–N–P model-based approach that has dominated traditional statistics (frequentist and Bayesian) since the 1930s, but "do they?" is a question to be deliberated in the sequel.

Although DS and GC modeling place themselves at opposite ends of the data-driven versus theory-driven spectrum, both perspectives share with the Pearson and the nonparametric approaches a common lineament in so far as they all view empirical modeling as:

(i) curve-fitting based on a prespecified family of mathematical functions $\mathcal{F}$,
(ii) subject to stochastic approximation errors (often white-noise), that
(iii) impart the implicit probabilistic assumptions imposed on data $\mathbf{x}_0$, and
(iv) the "best" curves in $\mathcal{F}$ are selected on goodness-of-fit/prediction grounds.

In particular, DS and nonparametric modeling replaces the Pearson family of distributions $\mathcal{F}_P$ with a set of mathematically smooth functions $g(\mathbf{x};\boldsymbol{\psi}_n)$ in $\mathcal{G}$ that live in "a normed linear space " bestowed with sufficient mathematical structure to ensure the existence and uniqueness of approximating functions $g_m(\mathbf{x};\widehat{\boldsymbol{\psi}}_n)$; see Aggarwal (2020), Iske (2018), and Murphy (2012). On the other hand, GC modeling relies on structural (functional) causal models, say $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$, where $\boldsymbol{\varphi}$ denotes the structural (causal) parameters, specified in terms of mathematical equations among the relevant variables in $\mathbf{Z}_t := (Y_t, \mathbf{X}_t)$, with white-noise approximation error terms attached; see Pearl (2000).

## 2 | PEARSON'S CURVE-FITTING DESCRIPTIVE STATISTICS

In Karl Pearson's approach to descriptive statistics (Yule, 1916), one begins with the raw data $\mathbf{x}_0$, whose initial rough summary takes the form of a histogram with $m \geq 10$ bins; see Spanos (2019), p. 445. To provide a more compact descriptive summary of the histogram Pearson would use the first four raw moments of $\mathbf{x}_0$ to select a frequency curve within a particular family known today as *the Pearson family* ($\mathcal{F}_P$). Members of this family are generated by the differential equation

$$\mathcal{F}_P: \frac{d\ln f(x;\boldsymbol{\psi})}{dx} = \left[(x-\psi_1)/\left(\psi_2+\psi_3 x+\psi_4 x^2\right)\right], \quad \boldsymbol{\psi}\in\Psi\subset\mathbb{R}^4, \quad x\in\mathbb{R} := (-\infty,\infty), \quad (3)$$

that includes well-known distributions, such as the Normal, the Student's $t$, the Beta, the Gamma, and so forth. Pearson also proposed an analogous difference equation for discrete random variables that includes distributions such as the Binomial, Poisson, and so forth. $\mathcal{F}_P$ is characterized by the four unknown parameters $\boldsymbol{\psi} := (\psi_1,\psi_2,\psi_3,\psi_4)$ that are estimated using $\widehat{\mu}_k(\mathbf{x}_0) = \frac{1}{n}\sum_{t=1}^{n} x_t^k$, with $\widehat{\mu}_{-1} = 0$, $\widehat{\mu}_0 = 1$, by solving four equations stemming from

Equation (3) to yield $\widehat{\boldsymbol{\psi}}(\mathbf{x}_0) := (\widehat{\psi}_1, \widehat{\psi}_2, \widehat{\psi}_3, \widehat{\psi}_4)$; see Ord (1972). One would then use the estimates $\widehat{\boldsymbol{\psi}}(\mathbf{x}_0)$ to select $f(x; \widehat{\boldsymbol{\psi}}) \in \mathcal{F}_P$ that "best" fits over the histogram. Its "appropriateness" is evaluated using Pearson (1900) goodness-of-fit test:

$$\eta(\mathbf{X}) = \sum_{i=1}^{m} \left[ (\widehat{f}_i - f_i)^2 / f_i \right] \underset{n\to\infty}{\backsim} \chi^2(m), \tag{4}$$

where $(\widehat{f}_i, i = 1, 2, \ldots, m)$ and $(f_i, i = 1, 2, \ldots, m)$ denote the empirical and assumed [as specified by $f_0(x)$] frequencies.

Similarly, Pearson's approach to correlation and regression amounts to curve-fitting guided by goodness-of-fit aiming to describe succinctly the association between data series, say $\mathbf{Z}_0 := \{(x_t, y_t), t = 1, 2, \ldots, n\}$. How the Pearson-type statistics was understood in the 1920s is summarized by Mills (1924), who distinguishes between "statistical description" versus "statistical induction." In statistical description measures such as the "sample" mean variance, and correlation coefficient:

$$\overline{x}_n = \frac{1}{n}\sum_{t=1}^{n} x_t,\ \overline{y}_n = \frac{1}{n}\sum_{t=1}^{n} y_t,\ \widehat{\sigma}_x^2 = \frac{1}{n}\sum_{t=1}^{n}(x_t - \overline{x}_n)^2,\ \widehat{\sigma}_y^2 = \frac{1}{n}\sum_{t=1}^{n}(y_t - \overline{y}_n)^2,$$

$$\widehat{\rho}_{xy} = \left[ \left( \sum_{t=1}^{n}(x_t - \overline{x}_n)(y_t - \overline{y}_n) \right) \Big/ \sqrt{\left[ \sum_{t=1}^{n}(x_t - \overline{x}_n)^2 \right] \left[ \sum_{t=1}^{n}(y_t - \overline{y}_n)^2 \right]} \right], \tag{5}$$

"provide just a summary for the data in hand" and "may be used to perfect confidence, as accurate descriptions of the given characteristics" (p. 549). However, when the results are to be extended *beyond* the data in hand—statistical induction—their validity depends on certain underlying a priory stipulations such as: (a) the "uniformity" for the *population* and (b) the "representativeness" of the *sample* (pp. 550–552). That is, the statistical description does not invoke the validity of any assumptions, but if the same data are used to go beyond the data in hand (inductive inference), one needs to invoke (a) and (b).

What Pearson and Mills did not appreciate sufficiently in the 1920s is that, irrespective of whether one is summarizing the data for descriptive or inferential purposes, one implicitly imposes probabilistic assumptions on the data. For instance, the move from the raw data $\mathbf{x}_0$ to a histogram invokes a "random" (IID) sample $\mathbf{X} := (X_1, \ldots, X_n)$ underlying data $\mathbf{x}_0$. This provided a probabilistic framing of assumptions (a) and (b) above. When the IID assumptions are invalid for $\mathbf{x}_0$, not only the data moments in Equation (5) but also the frequency curve chosen $f(x; \widehat{\boldsymbol{\psi}})$ will be highly misleading; see Spanos (2019). Similarly, Pearson's approach to correlation and regression is viewed as curve-fitting that (implicitly) assumes that the data $\mathbf{Z}_0 := \{(x_t, y_t),\ t = 1, 2, \ldots, n\}$ are IID over the ordering $t$.

In terms of the [a]–[f] criteria in the introduction, the Pearson approach can be summarized as follows:

[a] **Underlying framework**: the probabilistic framework is narrowed by the (implicit) IID assumptions and the choice of a distribution from the Pearson family $\mathcal{F}_P(x; \boldsymbol{\psi})$ in Equation (3).
[b] **Inductive premises**: the IID assumptions are taken at face value, and $f(x; \boldsymbol{\psi}) \in \mathcal{F}_P(x; \boldsymbol{\psi})$. $f(x; \boldsymbol{\psi})$ is viewed as a model that summarizes the information in data $\mathbf{x}_0$.

[c] **Model choice**: the "best" fitted model $f(x;\widehat{\boldsymbol{\psi}}),\ x\in\mathbb{R}$, upon which inductive inference is based, is evaluated on goodness-of-fit grounds tested using Pearson's chi-square test in Equation (4).
[d] **Quantification**: data-driven curve-fitting based on Pearson's method of moments using $\widehat{\mu}_k,\ k=1,2,3,4$, to estimate $f(x;\widehat{\boldsymbol{\psi}})$.
[e] **Inductive reasoning**: inferences based on descriptive statistics, $\widehat{\mu}_k,\ k=1,2,3,4$, sample correlation $(r_{xy})$ and estimated linear regression (LR) coefficients $(y_t=\widehat{\beta}_0+\widehat{\beta}_1x_t+\widehat{u}_t)$, are based on "large $n$ approximations" (as $n\to\infty$ ). The approach was a hybrid of large sample $(n)$ frequentist descriptive statistics with a tinge of Bayesian flavor based on a "uniform" *prior* distribution, with the emphasis placed on "probable errors," a 0.5 Bayesian credible interval, for example, $\widehat{\mu}_1\pm.6745\sqrt{Var(\widehat{\mu}_1)}$; see Yule (1916). Note that that "uniform" prior was considered widely acceptable (Pearson, 1920), and the conventional wisdom was that as "$n\to\infty$" the prior 'washes out'; see Edgeworth (1884).
[f] **Substantive versus statistical**: the questions of interest need to be framed in terms of the relevant descriptive statistics.

## 3 | FISHER'S RECASTING OF STATISTICS

Regrettably, the statistics literature until the 1920s conflated the sample $\mathbf{X}$ with the sample realization (the observed data) $\mathbf{x}_0$, and the estimator $\widehat{\theta}(\mathbf{X})$, the estimate $\widehat{\theta}(\mathbf{x}_0)$, and the unknown parameter $\theta$. Fisher (1922), p. 311, was first to delineate these different concepts and explain the importance of these distinctions.

### 3.1 | Model-based frequentist approach

Inspired by Gossett (1908) derivation of the Student's t distribution based on unveiling the implicit probabilistic assumptions invoked for that derivation, Fisher (1922) recast Pearson's curve-fitting into modern model-based induction by viewing the data $\mathbf{x}_0$ as a "typical realization" of stochastic Generating Mechanism (GM) in the form of a prespecified *parametric statistical model* in Equation (1).

**Example 2.** The simple Bernoulli model:

$$X_k\backsim \mathsf{BerIID}(\theta,\theta(1-\theta)),\quad x_k=0,1,\quad E(X_k)=\theta\in[0,1],\ Var(X_k)=\theta(1-\theta),\quad k\in\mathbb{N}. \tag{6}$$

Fisher proposed a complete reformulation of statistical induction from generalizing the observed "relative frequencies" in $\mathbf{x}_0$ to unobserved probabilities in $f(x;\widehat{\boldsymbol{\psi}})$, to modeling the statistical underlying GM $[\mathcal{M}_\theta(\mathbf{x})]$ specified in terms of the observable stochastic process $\{X_t,\ t\in\mathbb{N}\}$ underlying data $\mathbf{x}_0$.. According to Fisher (1922), $\mathcal{M}_\theta(\mathbf{x})$ is chosen by responding to the question: "Of what population is this a random sample?" (p. 313), and adding that "and the adequacy of our choice may be tested posteriori." (314), emphasizing the importance of model validation:

"For empirical as the specification of the hypothetical population $[\mathcal{M}_\theta(\mathbf{x})]$ may be, this empiricism is cleared of its dangers if we can apply a rigorous and objective test of the adequacy with which the proposed population $[\mathcal{M}_\theta(\mathbf{x})]$ represents the whole of the available facts." (p. 314).

The idea is that the probabilistic assumptions comprising $\mathcal{M}_\theta(\mathbf{x})$ account for the chance regularity patterns exhibited by $\mathbf{x}_0$, rendering $\mathbf{x}_0$ a typical realization of $\mathcal{M}_\theta(\mathbf{x})$ by pairing chance regularities with appropriate probabilistic assumptions from the three broad categories: Distribution, Dependence, and Heterogeneity. The "typicality" can be validated using Mis-Specification (M-S) testing; see Spanos (2018).

Phenomena of interest are amenable to model-based statistical inference when the data they give rise to exhibit inherent chance regularity patterns. As argued by Neyman (1952): "The application of the theory involves the following steps: (i) if we wish to treat certain phenomena by means of the theory of probability we must find some element of these phenomena that could be considered as random, following the law of large numbers (LLN). This involves the construction of a mathematical model of the phenomena involving one or more probability sets. (ii) The mathematical model is found satisfactory, or not. This must be checked by observation. (iii) If the mathematical model is found satisfactory, then it may be used for deductions concerning phenomena to be observed in the future." (p. 27). In (i) and (ii), Neyman raises the *frequentist interpretation* of probability being grounded on the strong law of large numbers (SLLN), and its validity is ensured when the statistical adequacy of $\mathcal{M}_\theta(\mathbf{x})$ is established (Spanos, 2013). In (iii), he indicates that the validity of the probabilistic assumptions comprising $\mathcal{M}_\theta(\mathbf{x})$ is evaluated before any inferences are drawn.

The *primary objective* of frequentist inference is to use the sample information, as summarized by $f(\mathbf{x};\theta)$, $\mathbf{x}\in\mathbb{R}^n_X$, in conjunction with data $\mathbf{x}_0$, to *narrow down* $\Theta$ as much as possible, ideally, to a single point $\theta^*$— the "true" value of $\theta$ in $\Theta$, which is shorthand for saying that $\mathcal{M}^*(\mathbf{x})=\{f(\mathbf{x};\theta^*)\}$, $\mathbf{x}\in\mathbb{R}^n_X$, could have generated data $\mathbf{x}_0$. In practice, this ideal situation is unlikely to be reached, except by happenstance, but that does not prevent learning from $\mathbf{x}_0$. To ensure the uniqueness of $\theta^*$, it is assumed that $f(\mathbf{x};\theta)=f(\mathbf{x};\boldsymbol{\psi})$ if and only if $\theta=\boldsymbol{\psi}\in\Theta$, referred to as *statistical identifiability*.

The optimality and reliability (effectiveness) in learning from data about $\theta^*$ are evaluated using *error probabilities* associated with different procedures stemming from their sampling distributions. The sampling distribution, $f(y_n;\theta)$, $\forall y_n\in\mathbb{R}$, of a statistic (estimator, test, predictor) $Y_n=g(X_1,X_2,\dots,X_n)$ is derived using prespecified values of $\theta$ based on two different forms of reasoning, (i) factual (estimation and prediction): presuming that $\theta=\theta^*$, whatever that value happens to be in $\boldsymbol{\Theta}$, and (ii) hypothetical (hypothesis testing): various hypothetical scenarios based on $\theta$ taking different prespecified values under $H_0$: $\theta\in\boldsymbol{\Theta}_0$ (presuming that $\theta\in\boldsymbol{\Theta}_0$) versus $H_1$: $\theta\in\boldsymbol{\Theta}_1$ (presuming that $\theta\in\boldsymbol{\Theta}_1$), where $\boldsymbol{\Theta}_0\cup\boldsymbol{\Theta}_1=\boldsymbol{\Theta}$, $\boldsymbol{\Theta}_0\cap\boldsymbol{\Theta}_1=\varnothing$; see Spanos (2019), p. 576. The derivation of $f(y_n;\theta)$, $\forall y_n\in\mathbb{R}$ takes the form of:

$$F_n(y)=\mathbb{P}(Y_n\leq y)=\underbrace{\int\int\cdots\int}_{\{\mathbf{x}:\ g(\mathbf{x})\leq y\}} f(\mathbf{x};\theta)d\mathbf{x},\quad \forall y\in\mathbb{R}. \tag{7}$$

It is important to emphasize that neither form of reasoning involves conditioning on $\theta$ since the latter makes no mathematical or logical sense; $\theta$ is an *unknown constant*. Also, a misspecified $\mathcal{M}_\theta(\mathbf{x})$ is based on an erroneous $f(\mathbf{x};\theta)$ and thus the derivations in Equation (7) are likely to distort $f(y_n;\theta)$, $\forall y_n\in\mathbb{R}$, inducing inconsistency in estimators and sizeable discrepancies between the actual error probabilities and the nominal (assumed) ones in confidence intervals (CIs) and testing procedures.

In summary, the main features of the Fisher model-based approach are:

[a] **Underlying framework**: the underlying probabilistic framework is grounded in a stochastic process $\{X_t,\ t \in \mathbb{N}\}$ defined on a probability space $(S, \mathfrak{F}, \mathbb{P}(.))$ that satisfies the well-known three axioms; see Kolmogorov (1933) and Doob (1953). The modeling and inference revolve around the concept of a prespecified parametric statistical model $\mathcal{M}_\theta(\mathbf{x})$ viewed as a particular parametrization of the stochastic process $\{X_t,\ t \in \mathbb{N}\}$ underlying data $\mathbf{x}_0$.
[b] **Inductive premises**: the probabilistic assumptions comprising $\mathcal{M}_\theta(\mathbf{x})$ constitute a complete, internally consistent, and testable set of probabilistic assumptions from three broad categories, Distribution, Dependence, and Heterogeneity. $\mathcal{M}_\theta(\mathbf{x})$ is viewed as a statistical model that could have given rise to data $\mathbf{x}_0$, aiming to account for all the chance regularity patterns exhibited by the data $\mathbf{x}_0$.
[c] **Model choice**: the appropriate $\mathcal{M}_\theta(\mathbf{x})$ is chosen on *statistical adequacy* grounds (the validity of its premises for data $\mathbf{x}_0$) established by using M-S testing.
[d] **Quantification**: the estimation of $\theta \in \Theta$ $[\mathcal{M}_\theta(\mathbf{x})]$ usually relies on the maximum likelihood (ML) method because of the optimal properties of its estimators that include reparametrization invariance.
[e] **Inductive inference**: the interpretation of probability is frequentist and the underlying inductive reasoning comes in two forms: factual (estimation, prediction) and hypothetical (testing), giving rise to two different forms of inference both of which relate to learning from data about $\theta^*$, the true value of $\theta$ in $\Theta$. The effectiveness (optimality) of inference procedures is calibrated using error probabilities based on the sampling distributions of statistics (estimators, tests, predictors).

## 4 | FOUNDATIONAL ISSUES IN FREQUENTIST MODELING

### 4.1 | Fisher's reduction of data: a model-based perspective

Fisher (1925), p. 1: "Statistics may be regarded as (i) the study of **populations**, (ii) as the study of **variation**, (iii) as the study of methods of the **reduction of data**."

He then proceeds to elaborate on the reduction of data

""The problems which arise in reduction of data may be conveniently divided into three types: (i) problems of **Specification**. These arise in the choice of the mathematical form of the population. This is not arbitrary, but requires an understanding of the way in which the data are supposed to, or did in fact, originate. Its further discussion depends on such fields as the theory of Sample Survey, or that of Experimental Design. (ii) When the specification has been obtained, problems of **Estimation** arise. . . . (iii) Problems of **Distribution** include the mathematical deduction of the exact nature of the distributions in random samples of our estimates of the parameters, and of other statistics designed to test the validity of the specification (tests of **Goodness of Fit**). (p. 8).

The last sentence, however, gave rise to confusion because Fisher did not distinguish between significance testing *within* the statistical model $\mathcal{M}_\theta(\mathbf{x})$, say testing $H_0 : \mu = \mu_0$, assuming that Equation (2) is valid, and *M-S testing* that probes *outside* $\mathcal{M}_\theta(\mathbf{x})$, that is, tests the validity the NIID assumptions. M-S testing is concerned with whether the $\mathbf{x}_0$ constitutes a "typical realization" of $\mathcal{M}_\theta(\mathbf{x})$, by assessing how well $\mathcal{M}_\theta(\mathbf{x})$ accounts for all the systematic statistical information (reflected in the chance regularity patterns exhibited by data $\mathbf{x}_0$); see Spanos (2018).

Due to his focus on data from experimental designs and sample surveys, Fisher's writings do not elaborate on how the substantive (theory-based) information relates to the statistical information

(data-based). In an attempt to shed some light on that issue, Neyman (1939/1952) distinguished between "interpolatory" [statistical] and "explanatory" [substantive] models, where

"The latter try to provide an explanation of the mechanism underlying the observed phenomena; Mendelian inheritance was Neyman's favorite example. ... to develop a "genuine explanatory theory" requires substantial knowledge of the scientific background of the problem." (Lehmann, 1990, p. 956).

Unfortunately, Neyman's distinction between "explanatory" and "interpolatory" models does not explain the connection, if any, between the two, and their respective link to the data; see Spanos (2006).

The F–N–P model-based frequentist approach has been plagued by several foundational problems that have bedeviled its proper implementation and interpretation of its results since the 1930s, including the following two.

**Foundational issue 1**. How to secure statistical adequacy: the validity of the probabilistic assumptions comprising the invoked $\mathcal{M}_\theta(\mathbf{x})$ vis-a-vis data $\mathbf{x}_0$.

**Foundational issue 2**. How to relate the substantive (subject matter) and statistical information in the context of model-based statistics aiming to give rise to reliable learning from data about observable phenomena of interest.

Neyman (1939/1952), p. 42, underscored the crucial gap between a substantive model and the phenomenon of interest and proposed to bridge it using the data:

"...it is my strong opinion that no mathematical theory refers exactly to happenings in the outside world and that any application requires a solid bridge over an abyss. The construction of such a bridge consists first, in explaining in what sense the mathematical model provided by the theory is expected to 'correspond' to certain actual happenings and second, in checking empirically whether or not the correspondence is satisfactory."

It turns out that the two foundational problems are inextricably bound up because the current perspective on empirical modeling in most disciplines, including economics, is dominated by hybird-models (a mixure of substantive and statistical asumptions) whose curve-fitting is guided by ad hoc error term assumptions, which are evaluated on goodness-of-fit/prediction grounds; see Spanos (2010c). As a result, the overwhelming majority of "fitted curves" are both statistically and substantively misspecified and there is no way to separate the two types of misspecification (statistical vs. substantive) so that one would know how to address them. This is a variant of *Duhem's conundrum*: no hypothesis can be tested separately from the set of auxiliary hypotheses needed for such empirical tests; see Mayo (1996). This renders the task of establishing either form of adequacy impossible, unless the two perspectives, statistical and substantive, are initially untangled by defining the statistical model to comprise only the probabilistic assumptions imposed (directly or indirectly) on the data. This will enable one to secure the statistical adequacy of $\mathcal{M}_\theta(\mathbf{x})$—ensuring the statistical reliability of inference—before any questions of interest are posed to $\mathbf{x}_0$ or the cogency of the substantive information is evaluated.

## 4.2 | Establishing the statistical adequacy of $\mathcal{M}_\theta(\mathbf{z})$

The first step in establishing the statistical adequacy of $\mathcal{M}_\theta(\mathbf{z})$ is to refine Fisher's reduction of data into "specification," "estimation," and "distribution," by distinguishing between the "modeling" and "inference" facets, where the former includes the cycle "specification,""estimation," "M-S testing," and "respecification" until a statistically adequate $\mathcal{M}_\theta(\mathbf{z})$ is reached, and the latter includes all inferences grounded on $\mathcal{M}_\theta(\mathbf{z})$; see Spanos (2006). The primary aim of this

**TABLE 1** The LR model: Traditional specification

| |
|---|
| $\boldsymbol{Y_t = \beta_0 + \beta_1^\top x_t + \varepsilon_t, t \in \mathbb{N}}$, |
| $\{1\}\,(\varepsilon_t \mid \mathbf{X}_t = \mathbf{x}_t) \backsim \mathsf{N}(.,.)$, $\{2\}\,E(\varepsilon_t \mid \mathbf{X}_t = \mathbf{x}_t) = 0, \mathbf{X}_t : (m \times 1)$, |
| $\{3\}\,E(\varepsilon_t^2 \mid \mathbf{X}_t = \mathbf{x}_t) = \sigma_\varepsilon^2$, $\{4\}\,E(\varepsilon_t \cdot \varepsilon_s \mid \mathbf{X}_t = \mathbf{x}_t) = 0, t \neq s, t, s \in \mathbb{N}$. |

**TABLE 2** Normal, linear regression (LR) model

| **Statistical GM:** $\boldsymbol{Y_t = \beta_0 + \beta_1^\top x_t + u_t,\ t \in \mathbb{N}}$, | |
|---|---|
| [1] Normality : $(Y_t \mid \mathbf{X}_t = \mathbf{x}_t) \backsim \mathsf{N}(.,.)$, | $t \in \mathbb{N}$. |
| [2] Linearity : $E(Y_t \mid \mathbf{X}_t = \mathbf{x}_t) = \beta_0 + \boldsymbol{\beta}_1^\top \mathbf{x}_t$, | |
| [3] Homosk/city : $Var(Y_t \mid \mathbf{X}_t = \mathbf{x}_t) = \sigma^2$, | |
| [4] Independence : $\{(Y_t \mid \mathbf{X}_t = \mathbf{x}_t), t \in \mathbb{N}\}$ independent process, | |
| [5] $t$-invariance : $\theta := (\beta_0, \boldsymbol{\beta}_1, \sigma^2)$ are *not* changing with $t$, | |
| $\boldsymbol{\beta} := (\beta_0, \boldsymbol{\beta}_1)^\top$, $\beta_0 = E(Y_t) - \boldsymbol{\beta}_1^\top E(\mathbf{X}_t)$, $\boldsymbol{\beta}_1 = [Cov(\mathbf{X}_t)]^{-1} Cov(\mathbf{X}_t, Y_t)$, | |
| $\sigma^2 = Var(Y_t) - Cov(\mathbf{X}_t, Y_t)^\top [Cov(\mathbf{X}_t)]^{-1} Cov(\mathbf{X}_t, Y_t)$ | |

*Notes*: GM: generating mechanism.

distinction is to provide a framework where the statistical adequacy of $\mathcal{M}_\theta(\mathbf{z})$ is established before the inference facet to ensure its reliability.

The second step needed is to ensure that $\mathcal{M}_\theta(\mathbf{z})$ is specified in terms of a complete, internally consistent, and testable set of probabilistic assumptions; internal consistency implies that $\mathcal{M}_\theta(\mathbf{z})$ is not well-defined. As argued in Spanos (1986), a way to ensure all these preconditions is to view a statistical model $\mathcal{M}_\theta(\mathbf{z})$ from a probabilistic reduction (PR) perspective as a particular parametrization of the observable (vector) process $\{\mathbf{Z}_t,\ t \in \mathbb{N}\}$ underlying data $\mathbf{Z}_0$. This places the problem of specifying $\mathcal{M}_\theta(\mathbf{z})$ squarely in the context of stochastic processes where Kolmogorov (1933) extension (existence) theorem guarantees that that joint distribution $f(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n; \boldsymbol{\phi})$ provides a complete description of any stochastic process $\{X_t,\ t \in \mathbb{N}\}$ under some mild regularity conditions; see Billingsley (1995).

**Example 3.** The traditional specification of the LR model is given in Table 1, where {2}–{4} are the well-known Gauss–Markov assumptions, with {1} Normality being viewed as optional. It can be shown that {1}–{4} do not constitute a complete set of testable probabilistic assumptions for the LR model by comparing it to the specification in Table 2, stemming from the PR perspective viewing $\mathcal{M}_\theta(\mathbf{z})$ as a particular parametrization of the observable process $\{\mathbf{Z}_t = (Y_t, \mathbf{X}_t),\ t \in \mathbb{N}\}$.

That is, one can derive the LR model by imposing the reduction assumptions of NIID to give rise to the Probabilistic Reduction (PR):

$$f(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n; \boldsymbol{\phi}) \overset{\text{IID}}{=} \prod_{t=1}^{n} f(y_t \mid \mathbf{x}_t; \theta) \cdot f(\mathbf{x}_t; \boldsymbol{\phi}_2), \quad \mathbf{z}_t \in \mathbb{R}^{m+1}, \tag{8}$$

with the LR model framed in terms of $f(y_t \mid \mathbf{x}_t; \theta)$ in Table 2. The reduction in Equation (8) gives rise to the statistical parameterization of $\theta := (\beta_0, \boldsymbol{\beta}_1, \sigma^2)$ in Table 2, and ensures the completeness, internal consistency, and testability of the probabilistic assumptions [1]–[5]. Notice that assumption [5] is missing from the specification in Table 1. In addition, the relationship between the reduction assumptions of NIID for $\{\mathbf{Z}_t, t \in \mathbb{N}\}$ and model assumptions [1]–[5] for

$\{(y_t|\mathbf{X}_t = \mathbf{x}_t), t \in \mathbb{N}\}$

$$\mathsf{N} \longrightarrow [1]\text{–}[3], \quad \mathsf{I} \longrightarrow [4], \quad \mathsf{ID} \longrightarrow [5], \tag{9}$$

plays a crucial role at all stages of *modeling* (specification , M-S testing, respecification) when any of the assumptions [1]–[5] are found invalid (Spanos, 2019).

The above recasting of a statistical model $\mathcal{M}_\theta(\mathbf{z})$ as a purely probabilistic concept is instrumental in addressing the above two foundational issues. This enables one to separate the *modeling* (specification, M-S testing, respecification) from the *inference* facet, and ensure the reliability of inference when posing the substantive questions of interest relating to the substantive model $\mathcal{M}_\varphi(\mathbf{z})$, knowing that the optimal properties of inference procedures hold; see (Spanos, 1990, 2006). The distinction between *statistical* and a *substantive* information/model enables one to separate two different forms of adequacy (Spanos, 2006)

(a) **Statistical adequacy**: $\mathcal{M}_\theta(\mathbf{z})$ adequately accounts for the chance regularities in $\mathbf{Z}_0$, or equivalently, the probabilistic assumptions comprising $\mathcal{M}_\theta(\mathbf{z})$ are valid for data $\mathbf{Z}_0$, which is solely data-oriented; do data $\mathbf{Z}_0$ satisfy assumptions [1]–[5]?
(b) **Substantive adequacy**: the extent to which $\mathcal{M}_\varphi(\mathbf{z})$ sheds adequate light on (describe, explain, predict) the phenomenon of interest. This pertains to ceteris paribus clauses, omitted variables, latent confounders, causal links, and so forth, since they all relate to substantive "information" about "how the world works."

## 4.3 | M-S testing

An effective way to apply M-S testing to the LR model (Table 2), is to use joint tests based on *auxiliary regressions* in terms of the residuals $\widehat{u}_t$ and $\widehat{u}_t^2$ to probe assumptions [2]–[5] by adding terms that could pick up different potential departures in ways they might affect the regression and skedastic functions; see Spanos (2018). Intuitively, these auxiliary regressions could be viewed as attempts to probe $\{(\widehat{u}_t, \widehat{u}_t^2)\ t = 1, 2, \ldots, n\}$ for any leftover systematic information in $\mathbf{Z}_0$ that might have been overlooked by $\mathcal{M}_\theta(\mathbf{z})$. M-S tests for the Normality assumption [1] invariably assume that assumptions [2]–[5] are valid, and thus it should be applied only when the validity of these assumptions is established; see Spanos (2018).

**Example 4.** Lai and Xing (2008), pp. 71–81, illustrate the Capital Asset Pricing Model (CAPM) using *monthly data* for the period August 2000 to October 2005 ($n = 64$). Focusing on one of their equations where: $y_t$ is excess (log) returns of Intel Corp., $x_t$ is the market excess (log) returns based on the SP500 index, where the risk-free returns are based on the 3-month Treasury bill rate. Estimation of the statistical (LR) model that nests the CAPM when the constant is zero yields

$$Y_t = \underset{(0.009)}{0.002} + \underset{(0.236)}{1.996}x_t + \widehat{u}_t, \quad R^2 = 0.536, \quad s = 0.0498, \quad n = 64, \tag{10}$$

where the standard errors are given in parentheses. When the estimated LR model in Equation (10) is taken at face value, it appears as though the signs and magnitudes of the estimated $(\beta_0, \beta_1)$ corroborate the CAPM in the sense that at $\alpha = .025$ the beta coefficient $\beta_1$ is statistically significant, but $\beta_0$ is not; the goodness-of-fit ($R^2 = 0.536$) is high enough. Such inferences

will be reliable only when assumptions [1]–[5] in Table 2 are valid for the particular data. But are they?

Departures from assumptions [2]–[5] are confirmed by the auxiliary regressions

$$\widehat{u}_t = \underset{(0.008)}{-0.011} + \underset{(0.468)}{0.595}x_t + \overbrace{\underset{(6.34)}{4.85}x_t^2}^{\neg[2]} - \overbrace{\underset{(0.046)}{0.212^*}D_2}^{\neg[5]} - \overbrace{\underset{(0.012)}{0.021^*}t_s + \underset{(0.023)}{0.065^*}t_s^2}^{\neg[5]} - \overbrace{\underset{(0.11)}{0.28}\widehat{u}_{t-1}}^{\neg[4]} + \widehat{v}_{1t}, \qquad (11)$$

where $t_s = [(2t - n - 1)/(n - 1)]$, $D_2$ is a dummy variable for $t = 2$

$$\widehat{u}_t^2 = \underset{(0.0004)}{0.002} + \overbrace{\underset{(0.0026)}{0.026^*}D_2}^{\neg[5]} - \overbrace{\underset{(0.0006)}{0.002^*}t}^{\neg[5]} - \overbrace{\underset{(0.06)}{0.061}x_t^2}^{\neg[3]} + \overbrace{\underset{(0.08)}{0.099}\widehat{u}_{t-1}^2}^{\neg[4]} + \widehat{v}_{2t}. \qquad (12)$$

The statistical significance of the added terms for possible departures from assumptions [2]–[5] indicate that assumptions [4] and [5] are invalid for the particular data; see Spanos (2019), p. 644–646. Hence, the above M-S testing results indicate clearly that no reliable inferences can be drawn on the basis of the estimated model in Equation (10).

The above auxiliary regressions can be easily modified/extended to apply to many statistical models of interest, including generalized linear models, time series, and panel data models; see Spanos (2019).

## 4.4 | Statistical versus substantive information/models

Empirical modeling across different disciplines involves the merging of *substantive* (subject matter) and *statistical* (chance regularities in data) *information*, both of which play crucial roles in shedding light (explain, describe, predict) observable phenomena of interest. The substantive information comes in the form of a theory or theories about the phenomenon of interest and could range from simple tentative conjectures to intricate *substantive* (structural) models. The substantive information has a crucial and multifaceted role to play, that begins with demarcating the crucial aspects of the phenomenon of interest and suggesting the relevant variables $\mathbf{Z}_t$ and observed data $\mathbf{Z}_0 := (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n)$ to be used for learning from $\mathbf{Z}_0$.

The *statistical information* comes in the form of chance regularity (recurring) patterns exhibited by a particular data set $\mathbf{Z}_0$. Such patterns can be revealed using a variety of graphical techniques in conjunction with preliminary data analysis; see Spanos (2019). In empirical modeling, these patterns constitute the statistical systematic information that provides the basis for selecting an appropriate statistical model $\mathcal{M}_\theta(\mathbf{z})$. The difficult conundrum in empirical modeling over the last century has been to find a way to blend both sources of information in a pertinent way that would not undermine the credibility of either.

### 4.4.1 | Reinterpreting the simultaneous equations model (SEM)

A crucial problem in macroeconometric modeling in the 1930s was the *simultaneity*—contemporaneous feedbacks—among the variables involved. This raised technical problems for

the *identification* and *estimation* of the parameters $\boldsymbol{\varphi}$ of a structural model that were initially addressed by Haavelmo (1943) and subsequently formalized and extended by the Cowles Commission in the form of the Simultaneous Equations Model (SEM) (Hood & Koopmans, 1953) based on the distinction between the structural and reduced forms

$$\text{Structural form}: \mathbf{\Gamma}^{\top}\mathbf{Y}_t = \mathbf{\Delta}^{\top}\mathbf{x}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \backsim \mathsf{N}(\mathbf{0}, \mathbf{\Omega}), \quad t \in \mathbb{N}, \tag{13}$$

$$\text{Reduced form}: \mathbf{Y}_t = \mathbf{B}^{\top}\mathbf{x}_t + \mathbf{u}_t, \quad \mathbf{u}_t \backsim \mathsf{N}(\mathbf{0}, \mathbf{\Sigma}), \quad t \in \mathbb{N}, \tag{14}$$

where $\mathbf{Y}_t$: $(m \times 1)$-endogenous and $\mathbf{X}_t$: $(k \times 1)$-exogenous variables, $E(\mathbf{Y}_t) = \mathbf{0}$ and $E(\mathbf{X}_t) = \mathbf{0}$ (for simplicity) and $\boldsymbol{\varphi} := [\mathbf{\Gamma}(\boldsymbol{\varphi}_1), \mathbf{\Delta}(\boldsymbol{\varphi}_2), \mathbf{\Omega}(\boldsymbol{\varphi}_3)]$, denotes a $(q \times 1)$ vector of the unknown structural parameters in $(\mathbf{\Gamma}, \mathbf{\Delta}, \mathbf{\Omega})$. Note that Equation (14) is derived by pre-multiplying Equation (13) with $(\mathbf{\Gamma}^{\top})^{-1}$ solving for $\mathbf{Y}_t$ but *ignoring* any parameter restrictions imposed on $\boldsymbol{\theta} := [\mathbf{B}(\boldsymbol{\theta}_1), \mathbf{\Sigma}(\boldsymbol{\theta}_2)]$, where $\boldsymbol{\theta}$ is $(p \times 1)$ vector of the unknown parameters in $(\mathbf{B}, \mathbf{\Sigma})$, $p = mk + \frac{1}{2}m(m+1)$.

Taking $\boldsymbol{\theta}$ *at face value*, the *identification problem* is "solved" by deriving conditions (rank and order) under which the implicit system of equations

$$\mathbf{B}(\boldsymbol{\theta}_1)\mathbf{\Gamma}(\boldsymbol{\varphi}_1) = \mathbf{\Delta}(\boldsymbol{\varphi}_2), \quad \Omega(\boldsymbol{\varphi}_3) = \mathbf{\Gamma}^{\top}(\boldsymbol{\varphi}_1)\mathbf{\Sigma}(\boldsymbol{\theta}_2)\mathbf{\Gamma}(\boldsymbol{\varphi}_1), \tag{15}$$

can be solved uniquely for $\boldsymbol{\varphi}$ given $\boldsymbol{\theta}$; see Greene (2018). When Equation (13) is identified using the order and rank conditions, one can use the data $\mathbf{Z}_0 := \{(\mathbf{x}_t, \mathbf{y}_t), t = 1, 2, \ldots, n\}$ to estimate $\boldsymbol{\varphi}$ in Equation (13) using modified least-squares and ML procedures to account for the simultaneity among the variables in $\mathbf{Y}_t$; see Hendry (1976).

As argued in Spanos (1986), this perspective is misleading because Equations (13 )–(15) are nothing more than a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$—the multivariate LR in Equation (14)—subject to the substantive restrictions in Equation (15) by viewing Equation (13) as a reparameterization/restriction of Equation (14) with $\boldsymbol{\theta}$ taken at face value. That is, in traditional econometrics, the identification and estimation problems are "solved" relative to a "notional" reduced form whose statistical adequacy is taken at face value. In practice, however, Equation (14 ) is often statistically misspecified, calling into question the judiciousness of the traditional "solutions" of the identification and estimation problems.

Viewed from this model-based perspective, when $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ in Equation (14), comprising assumptions analogous to [1]–[5] in Table 2 with vector $\mathbf{Y}_t$, is statistically misspecified, the sampling distributions (finite sample or asymptotic) of the traditional estimators of $\boldsymbol{\varphi}$ in Equation (13), such as IV, 2LSL, 3SLS, LIML, FIML, and so forth (Phillips, 1983), will be distorted, rendering their inference procedures unreliable, and the ensuing evidence untrustworthy; see Spanos (1990).

Spanos (1986) proposed to view the reduced form in Equation (14) as the *implicit statistical model* $(\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z}))$, in the form of a multivariate LR model (Table 2 with a vector $\mathbf{Y}_t$) underlying the substantive (structural) model in Equation (13). The latter is *parametrically nested* within the former via Equation (15) whose generic form is

$$\mathbf{G}(\boldsymbol{\theta}, \boldsymbol{\varphi}) = \mathbf{0}, \quad \boldsymbol{\varphi} \in \mathbb{R}^q, \quad \boldsymbol{\theta} \in \mathbb{R}^p, \quad p \geq q. \tag{16}$$

These constitute the overidentifying restrictions (Greene, 2018) that can be tested using the hypotheses:

$$H_0 : \mathbf{G}(\theta, \varphi) = \mathbf{0} \quad \text{vs.} \quad H_1 : \mathbf{G}(\theta, \varphi) \neq \mathbf{0}. \tag{17}$$

When the reduced form $\mathcal{M}_\vartheta(\mathbf{z})$ is statistically misspecified, however, the testing of Equation (17) is likely to be unreliable in the sense that the actual error probabilities will be different from the nominal ones; hence, the need to be based on a statistically adequate model after M-S testing and respecification.

**Statistical versus structural identification**. To elbow room for securing the statistical adequacy of Equation (14) in the context of the SEM, Spanos (1990) introduced the distinction between *statistical* versus *structural identification*. The statistical identification renders $\theta := [\mathbf{B}(\theta_1), \mathbf{\Sigma}(\theta_2)]$ statistically well-defined (meaningful) by:

(i) establishing the statistical adequacy of Equation (14) using comprehensive M-S testing of assumptions [1]–[5] (Table 2) for a multivariate $\mathbf{Y}_t$, and
(ii) ensuring statistical identification: $f(\mathbf{y}_t \mid \mathbf{x}_t; \theta) = f(\mathbf{y}_t \mid \mathbf{x}_t; \psi)$, if and only if $\theta = \psi \in \Theta$.

Given (i) and (ii), the structural identification of Equation (13) relates to whether Equation (15) can be solved uniquely for $\varphi = \mathbf{H}(\theta)$. Equation (13) is *just identified* ($p = q$) or *overidentified* ($p > q$), depending on whether the mapping $\mathbf{H}(.)$ is bijective (one-to-one and onto) or surjective (many-to-one), respectively.

The distinction between statistical and substantive adequacy suggests that in the case of the SEM, a structural model (13) is *data-cogent* when:

(a) the implicit statistical model (14) is *statistically adequate*, and
(b) the *overidentifying restrictions* (17) are *data-acceptable*—do not belie $\mathbf{Z}_0$.

Conditions (a) and (b) are needed to ensure what Pearl and Mackenzie (2018), p. 85, refer to as "get something that was not in the data to begin with." The rationale for a larger $(p - q) > 0$ is that, when the restrictions in Equation (17) are valid for data $\mathbf{Z}_0$, the structural model $\mathcal{M}_\varphi(\mathbf{z})$ is less data-specific, and thus more *informative*! Under (a) and (b), the estimated *empirical model* $\mathcal{M}_{\hat{\varphi}}(\mathbf{z})$ enjoys both (i) statistical adequacy and (ii) theoretical meaningfulness, and thus it provides a better basis for prediction and policy simulations and a basis for "learning from data" about the phenomenon of interest. Securing (a) and (b), however, does not imply that $\mathcal{M}_\varphi(\mathbf{x})$ is substantively adequate. In certain disciplines, including the social sciences, there is a need to allow for provisional substantive adequacy that could be improved by gradual and cumulative betterment stemming from improving the quality of the data, and the cogency of the substantive information; see Spanos (2019).

### 4.4.2 | Substantive versus statistical models

It turns out that once the distinction between a substantive and statistical model is considerably more general than the SEM, and its reduced form in the sense that behind every substantive model

$\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$, generically specified by:

$$\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})=\{f(\mathbf{z};\boldsymbol{\varphi}),\boldsymbol{\varphi}\in\Phi\subset\mathbb{R}^{p}\},\quad \mathbf{z}\in\mathbb{R}^{n}_{Z},\quad p\leq m, \tag{18}$$

there is an implicit statistical model $\mathcal{M}_{\theta}(\mathbf{z})$ that comprises the probabilistic assumptions imposed on data $\mathbf{Z}_0$. What is not so obvious is how to bring out these assumptions explicitly, and test their validity before blending the two models by relating their parametrization $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$, and testing the over-identifying restrictions in Equation (17).

The above discussion on re-interpreting the SEM provides the answer in finding the implicit $\mathcal{M}_{\theta}(\mathbf{z})$, which can then be respecified as long as the parametric nesting of $\boldsymbol{\varphi}$ in $\boldsymbol{\theta}$, as in Equation (16), is retained.

**Example 5.** The structural CAPM for one asset, say $k$, takes the form:

$$\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z}):\ Y_t=\beta x_t+\varepsilon_t,\quad \varepsilon_t\backsim \mathsf{NIID}\big(0,\sigma^2_{\varepsilon}\big),\quad t\in\mathbb{N}, \tag{19}$$

where $\boldsymbol{\varphi}:=(\beta,\sigma^2_{\varepsilon})\in\mathbb{R}\times\mathbb{R}_+$, $Y_t:=(r_{kt}-r_{ft}),x_t:=(r_{Mt}-r_{ft}),r_{kt}$-returns of asset $k,r_{ft}$-returns of risk free asset, $r_{Mt}$-market returns. When $\mathcal{M}_{\theta}(\mathbf{z})$ is viewed as a particular parameterization of the observable stochastic process $\{\mathbf{Z}_t:=(r_{kt},r_{ft},r_{Mt}),\ t\in\mathbb{N}\}$ underlying data $\mathbf{Z}_0$, the implicit statistical model is the LR (Table 2) with:

$$\mathcal{M}_{\theta}(\mathbf{z}):\ r_{kt}=\beta_0+\beta_1x_t+\beta_2r_{Mt}+u_t,\quad u_t\backsim \mathsf{NIID}(0,\sigma^2),\quad t\in\mathbb{N}, \tag{20}$$

where $\boldsymbol{\theta}:=(\beta_0,\beta_1,\beta_2,\sigma^2)\in\mathbb{R}^3\times\mathbb{R}_+$. Hence, the restrictions in Equation (16) are:

$$\beta_0=0,\quad \beta_1+\beta_2=1. \tag{21}$$

To evaluate the empirical validity of $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$, one should estimate Equation (20), establish its statistical adequacy, and then test the validity of the substantive restrictions in Equation (21).

In conclusion, the enhanced model-based frequentist inference adds to the original Fisher perspective features [a]–[e], [f] **Substantive versus statistical**: the substantive, $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$, is nested within a statistically adequate $\mathcal{M}_{\theta}(\mathbf{z})$ using restrictions $\mathbf{G}(\boldsymbol{\theta},\boldsymbol{\varphi})=\mathbf{0},\boldsymbol{\theta}\in\Theta,\boldsymbol{\varphi}\in\Phi$.

## 5 | CURRENT GC MODELING

### 5.1 | From path diagrams to directed acyclic graphs

The initial framing of GC modeling was proposed by the geneticist (Wright, 1921, 1923, 1934) in the form of *path diagrams* among different variables to examine substantive hypotheses in phylogenetic studies. The analysis involved writing a system of equations based on the correlations among variables influencing the outcome and then solving them for the path coefficients in the model. This approach gained prominence with path diagrams for linear Structural Equation Models (SEMs) in the 1960s and 1970s; see Blalock (1964) and Duncan (1975).

Since these tentative beginnings, scientists in several disciplines, including computer science, philosophy of science, economics, and psychology, have made substantial progress in developing a formal language and an underlying coherent framework for GC models that generalize the linear path diagrams and integrate graphical models with the earlier literature on counterfactual tradition based on potential outcomes initiated by Neyman (1923, 1935) and Rubin (1974); see Morgan and Winship (2015). This more recent GC modeling literature revolves around the directed acyclic graphs (DAGs), where the dots and arrows are translated using formal rules into probabilistic models. These GC models can be estimated using data after they are particularized into structural (substantive) models related to the SEM discussed above; see Pearl (1988, 2009), Glymour et al. (1988), and Spirtes et al. (2000).

When viewed in the context of the model-based approach, GC modeling over-emphasizes the role of the structural (substantive) $(\mathcal{M}_{\varphi}(\mathbf{z}))$ model by viewing it as a curve-fitting problem that takes the validity of the statistical $(\mathcal{M}_{\theta}(\mathbf{z}))$—often implicit—at face value. This is a crucial mistake that can easily undermine the reliability of any empirical GC modeling. The distinction between $\mathcal{M}_{\theta}(\mathbf{z})$ and $\mathcal{M}_{\varphi}(\mathbf{z})$ and their parametric interrelationship $\mathbf{G}(\theta,\varphi)=\mathbf{0}, \varphi\in\mathbb{R}^q,\ \theta\in\mathbb{R}^p,\ p\geq q$, suggests most clearly that the sound quantification of $\mathcal{M}_{\varphi}(\mathbf{z})$ relies crucially on the statistical adequacy of the (often implicit) $\mathcal{M}_{\theta}(\mathbf{z})$. To borrow a widely used catchphrase from Pearl (2009), "correlation is not causation," but "causation needs to rely on trustworthy 'correlations' $(\mathcal{M}_{\theta}(\mathbf{z}))$." That is, a reliable empirical GC model $\mathcal{M}_{\varphi}(\mathbf{z})$ can only be constructed on a statistically adequate $\mathcal{M}_{\theta}(\mathbf{z})$ foundation; see Section 4. For instance, as argued in Section 3, Wright's path coefficients will be untrustworthy unless the estimated correlations among the variables involved (Fisher, 1918) are statistically adequate, that is, the NIID assumptions are valid for data $\mathbf{Z}_0$. The case against modern model-based statistics being anti-causal information put forward by Pearl (2009), pp. 339–342, Pearl and Mackenzie (2018), pp. 71–72, is primarily based Karl Pearson's (1911) descriptive statistics, curve-fitting, view of correlation and causality, and not Fisher (1922). After the 1930s, one would be hard-pressed to find any references to Pearson's statistical contributions beyond his chi-square test after it was reframed and integrated into the model-based statistics F–N–P frequentist statistics; see Spanos (2018). Moreover, one can make a strong case that the Cowles Commission formulation of the SEM in econometrics was strongly causal in nature; see Marshak (1953) and Simon (1953).

## 5.2 | Functional GC models

Interestingly, Wright's *path diagrams* can be transformed into linear equations among the relevant variables to specify a substantive model in the form of the SEM (Pearl, 2009). In that sense, the above recasting of the SEM (14)–(16) suggests that the problems with the traditional identification and estimation of the structural parameters $\varphi$ apply equally to path analysis. The concept of the implicit statistical model $\mathcal{M}_{\theta}(\mathbf{z})$ and its statistical adequacy play a crucial role in ensuring the structural causal model $\mathcal{M}_{\varphi}(\mathbf{x})$ does not belie the systematic information in the data.

GC modeling has formalized the causal dimension of substantive models by encoding the perceived causal information about "how we think the world works." A GC model is a set of connected variables, satisfying a number of conditional independence relations encoded by a graph using a Markov property, which often translates into a corresponding factorization of the associated density functions. To achieve that the literature has introduced a new language for causality that renders the causal structure transparent and provides ways to relate it to the underlying joint distribution $f(\mathbf{z}_1,\ldots,\mathbf{z}_n;\theta),\ \mathbf{z}\in\mathbb{R}_Z^{mn}$, of the sample $\mathbf{Z}:=(\mathbf{Z}_1,\ldots,\mathbf{Z}_n)$. Particularly

| Substantive (GC) model [a] $Z$ - **confounder** | Substantive (GC) model [b] $Z$ - **mediator** | Substantive (GC) model [c] $Z$ - **collider** |
|---|---|---|
| $Y_k=\beta_0+\beta_1X_k+\beta_2Z_k+\varepsilon_{1k}$, $X_k=\alpha_0+\alpha_1Z_k+\varepsilon_{2k}$, $k\in\mathbb{N}$ | $Y_k=\beta_0+\beta_1X_k+\beta_2Z_k+\varepsilon_{1k}$, $Z_k=\gamma_0+\gamma_1X_k+\varepsilon_{3k}$, $k\in\mathbb{N}$, | $Y_k=\delta_0+\delta_1X_k+\varepsilon_{4k}$, $k\in\mathbb{N}$, $Z_k=c_0+c_1Y_k+c_2X_k+\varepsilon_{5k}$, |
| Statistical model for [a] $Y_k=a_{01}+a_{11}Z_k+u_{1k}$, $X_k=a_{02}+a_{12}Z_k+u_{2k}$, | Statistical model for [b] $Y_k=b_{01}+b_{11}X_k+u_{3k}$, $Z_k=b_{02}+b_{12}X_k+u_{4k}$, | Statistical model for [c] $Y_k=b_{01}+b_{11}X_k+u_{3k}$, $Z_k=b_{02}+b_{12}X_k+u_{4k}$, |

**Diagram 1** Functional causal models

important are the conditions under which one can replace the do-calculus based on the $do(.)$ operator ("doing/intervening") with probabilistic conditioning ("seeing") in the reduction of the joint distribution of the sample; see Pearl (2009). Indeed, the new language of do-calculus has crystallized several vague notions of causal connections. A case in point is the notion a confounder for the relationship between $Y$ and $X$, which can be defined unambiguously as any factor $Z$ that renders $P(Y \mid X, Z) \neq P(Y \mid do(X), Z)$; see Pearl and Mackenzie (2018), pp. 150–157.

Of special interest in this context is a subset of graphical models based on DAGs that do not allow cycles of directed arrows; see Pearl (2009) and Spirtes et al. (2000). An important weakness of the GC modeling is that the causal information is usually treated as established knowledge instead of best-daresay conjectures whose adequacy needs to be tested against the relevant data. Foisting a DAG model, $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$, on data $\mathbf{Z}_0$, will often give rise to a statistically and substantively misspecified model because the estimation invokes more than the causal information; it implicitly imposes probabilistic assumptions on the particular data—(the implicit statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$—that might or might be valid for $\mathbf{Z}_0$.

It turns out that the answer to addressing this problem stems from the fact that a GC model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ is directly related to the SEM formulation in Equations (13) and (14). When the modeler proceeds to estimate the parameters of a SEM, such as those in diagram 1, using, say instrumental variables (IVs), he/she ignores the fact that every GC ($\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$) model has an implicit statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ whose statistical adequacy is usually taken at face value, but in practice, it needs to be established when modeling with real data.

That is, in practice, the modeling process for a GC model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ should involve several additional steps before one can declare the particular causal structure valid for a particular data; specifying a transparent causal structure, although extremely important, is only the beginning of the modeling facet. That process involves

Step 1. Unveil the implicit statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ implicit in the GC model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$.

Step 2. Estimate the unknown statistical parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and establish the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ using comprehensive M-S testing, and respecification if any of the original probabilistic assumptions turn out to be invalid.

Step 3. Use a statistically adequate $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ to address the identification and estimation of the structural parameters $\boldsymbol{\varphi} \in \boldsymbol{\Phi}$.

Step 4. Test the validity of the overidentifying restrictions stemming from Equation (14). Goodness-of-fit/prediction becomes relevant at this stage as part of evaluating the substantive adequacy of $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$.

For the GC models in diagram 2 (Spanos, 2020), this will involve

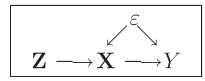


**Diagram 2** **Z** as an IV

(i) specifying explicitly the (implicit) statistical model $\mathcal{M}_\theta(\mathbf{z})$, a multivariate LR model for all three GC models,
(ii) testing the validity of their probabilistic assumptions analogous to [1]–[5] (Table 2), and after establishing the adequacy of the underlying statistical model
(iii) one needs to estimate the causal parameters and (iv) assess the validity of the relevant causal restrictions for each substantive model.

For instance, in case [a] one needs to test the validity of the restrictions needed to establish the particular causal structure by testing the hypotheses $H_0$: $\beta_1 = 0$ versus $H_1$: $\beta_1 \neq 0$; see Spanos (2005). This procedure is illustrated in Spanos (2020) using data for Simpson's paradox to bring out its empirical dimension.

It is important to emphasize that the distinction between the statistical, $\mathcal{M}_\theta(\mathbf{z})$, and structural model, $\mathcal{M}_\varphi(\mathbf{z})$, often renders possible the testing, not only of the overidentifying restrictions via Equation (17), but also certain causal conditions, such as the stability (faithfulness) in the context of $\mathcal{M}_\theta(\mathbf{z})$; see Spanos (2010a).

It is important to emphasize that DAG models can be viewed as restricted the SEM models in Equation (13) with a *recursive* structure

$$\mathbf{\Gamma}_R^\top(\boldsymbol{\varphi}_1)\mathbf{Y}_t = \mathbf{\Delta}_R^\top(\boldsymbol{\varphi}_2)\mathbf{x}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon} \backsim \mathsf{N}(\mathbf{0}, \mathbf{\Omega}_R(\boldsymbol{\varphi}_3)), \tag{22}$$

where $\mathbf{\Gamma}_R^\top$ is *lower triangular* (with $-1$ along the main diagonal) and $\mathbf{\Omega}_R$ a *diagonal matrix*. The underlying statistical model is a multivariate LR model as in Equation (14), and without any additional restrictions on $\boldsymbol{\varphi}_R := [\mathbf{\Gamma}_R^\top(\boldsymbol{\varphi}_1), \mathbf{\Delta}_R^\top(\boldsymbol{\varphi}_2), , \mathbf{\Omega}_R(\boldsymbol{\varphi}_3)]$, $\boldsymbol{\varphi}_R$ is *just identified* with respect to Equation (14). This suggests that a recursive SEM is observationally equivalent to all SEMs that share the same Equation (14), assuming the latter is statistically adequate! That is, the statistical model is a simple reparametrization of the statistical ($\boldsymbol{\theta}$) into structural ($\boldsymbol{\varphi}_R$) parameters. To restrict the equivalence class of DAG models, one needs to impose further structural restrictions such as "dropping" $(y_{it}, i = 1, \ldots, m)$ and $(x_{jt}, j = 1, \ldots, p)$ variables from specific equations. Such additional restrictions will give rise to overidentification restrictions, which can be tested; see Spanos (1986).

When modeling with observational time series data, a serious problem with DAG (recursive) models is that the observation period (weekly, monthly, quarterly, annual, etc.) rarely, if ever, coincides with the timing of the actual data GM, raising serious issues of what is being estimated. This problem could explain why the Strotz and Wold (1960) paper on recursive models did not have any noticeable impact on the subsequent development of econometric modeling, and not to any aversion for causal modeling.

## 5.3 | GC modeling and IVs

To illustrate how focusing exclusively on the structural model can be misleading in practice, consider a GC model relating to IVs in diagram 2 (Pearl, 2009), where $\mathbf{Z}$ is an $(m \times 1)$ vector

of IVs variables, $\mathbf{X}$ is a $(p\times 1)$ vector $(m\geq p)$ of explanatory variables, and $Y$ the variable to be explained.

The identification and estimation problems arise since both $\mathbf{X}$ and $Y$ are related to a third unobserved error term $\varepsilon$. When the GC model is framed in terms of a substantive (structural) model, in its simplest form is

$$\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z}):\ y_t=\boldsymbol{\alpha}^{\top}\mathbf{X}_t+\varepsilon_t,\quad \varepsilon_t\sim \mathsf{N}\big(0,\sigma^2_{\varepsilon}\big),\quad t\in\mathbb{N}, \tag{23}$$

where $E(\mathbf{X}_t)=\mathbf{0}$, $E(y_t)=0$ (to simplify the notation), $\mathbf{X}_t$ is a vector of explanatory variables that are arrogated to be correlated with the error term $\varepsilon_t$: (i) $E(\mathbf{X}_t\varepsilon_t)\neq\mathbf{0}$.

The traditional exposition (Greene, 2018) argues that the IV method addresses the non-orthogonality (i) by invoking substantive (theory-based) arguments to choose an $m\times 1$ $(m\geq p)$ vector of *IVs* $\mathbf{Z}_t$ that satisfy the restrictions

$$\text{(ii) } E(\mathbf{Z}_t\varepsilon_t)=\mathbf{0},\quad \text{(iii) } E\big(\mathbf{X}_t\mathbf{Z}_{\mathbf{t}}^{\top}\big)=\boldsymbol{\Sigma}_{23}\neq\mathbf{0},\quad \text{(iv) } E\big(\mathbf{Z}_t\mathbf{Z}_{\mathbf{t}}^{\top}\big)=\boldsymbol{\Sigma}_{33}>0. \tag{24}$$

The curve-fitting perspective on IV methods gives the impression that the choice of IVs $\mathbf{Z}_t$ has very little, if anything, to do with the probabilistic structure of the data since both methods rely on theory-based orthogonality conditions that revolve around latent error terms. This is both false and highly misleading. The same impression is also given by the causal diagram 1, which restricts the causal structure by ensuring that there is no direct arrow between $\mathbf{Z}$ and $Y$. Does this entails that there is no correlation between $\mathbf{Z}$ and $Y$? As argued below, the answer is no, for the simple reason that the restrictions (i)–(iv) in Equation (24), and diagram 1, ignore the other side of the modeling coin, which is the implicit statistical model $\mathcal{M}_{\theta}(\mathbf{z})$ underlying the IV method for identifying and estimating the unknown parameters of Equation (23)!

For simplicity, consider the case $p=m$ where one can use the sample equivalent to (ii) to derive the IV estimator (Spanos, 1986)

$$\begin{aligned}
&\mathbf{Z}^{\top}\boldsymbol{\varepsilon}=\mathbf{Z}^{\top}(\mathbf{y}-\mathbf{X}\boldsymbol{\alpha})=\mathbf{0}\rightarrow\widehat{\boldsymbol{\alpha}}_{IV}=(\mathbf{Z}^{\top}\mathbf{X})^{-1}\mathbf{Z}^{\top}\mathbf{y},\\
&\mathbf{X}=(\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_n)^{\top},\quad \mathbf{Z}=(\mathbf{Z}_1,\mathbf{Z}_2,\ldots,\mathbf{Z}_n)^{\top},\\
&\mathbf{y}=(y_1,y_2,\ldots,y_n)^{\top},\quad \boldsymbol{\varepsilon}=(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_n)^{\top}.
\end{aligned} \tag{25}$$

In their attempt to find instruments using (ii), applied econometricians have been seeking more and more ingenious and exotic IVs, including "lethal mosquitoes," "heirless maharajahs," "wind speeds," and "sea currents"; see Owen (2017). What such ingenious choices of instruments do not appreciate sufficiently is that (ii)–(iv) are not the only conditions needed for $\widehat{\boldsymbol{\alpha}}_{IV}$ in Equation (25) to make statistical sense as a consistent estimator of $\boldsymbol{\alpha}$. Given that $\widehat{\boldsymbol{\alpha}}_{IV}\overset{\mathbb{P}}{\rightarrow}\boldsymbol{\Sigma}_{23}^{-1}\boldsymbol{\sigma}_{13}$, where "$\overset{\mathbb{P}}{\rightarrow}$" reads "converges in probability," to avert the trivial but nonsensical case $\boldsymbol{\alpha}=\mathbf{0}$ one needs the additional condition:

$$\text{(v) } E(\mathbf{Z}_t y_t)=\boldsymbol{\sigma}_{13}\neq 0. \tag{26}$$

That is, $\mathbf{Z}_t$ needs to be uncorrelated with $\varepsilon_t$ but (statistically) correlated with $y_t$.

Closer scrutiny of the above theory-based choice of $\mathbf{Z}_t$ relying on the non-testable conditions (i)–(v) seems nothing short of the famous *deus ex machina* in Greek tragedies. To address this issue,

Spanos (1986) proposed a recasting of the conditions (i)–(v) in terms of the observable variables involved $\mathbf{W}_t^\top := (y_t, \mathbf{X}_t, \mathbf{Z}_t)$ by relating these conditions to the joint distribution of $\mathbf{W}_t$ using the PR:

$$f(y_t, \mathbf{x}_t, \mathbf{z}_t; \boldsymbol{\psi}) = f(y_t, \mathbf{x}_t \mid \mathbf{z}_t; \boldsymbol{\psi}_1) \cdot f(\mathbf{z}_t; \boldsymbol{\psi}_2), \quad \forall (y_t, \mathbf{x}_t, \mathbf{z}_t) \in (\mathbb{R} \times \mathbb{R}^m \times \mathbb{R}^p). \tag{27}$$

The recasting transforms (i) and (ii) into explicit substantive parameterizations of interest, analogous to the statistical ones in $\theta := (\beta_0, \boldsymbol{\beta}_1, \sigma^2)$ (Table 2), and renders conditions (iii) and (v) testable, by unveiling the implicit statistical model underlying Equation (27), based on $f(y_t, \mathbf{x}_t \mid \mathbf{z}_t; \boldsymbol{\psi}_1)$, which is a multivariate LR model:

$$\mathcal{M}_\theta(\mathbf{z}) : \mathbf{y}_t = \mathbf{B}^\top \mathbf{Z}_t + \mathbf{u}_t, \quad (\mathbf{u}_t \mid \mathbf{Z}_t = \mathbf{z}_t) \sim \mathsf{N}(\mathbf{0}, \boldsymbol{\Omega}), \quad t \in \mathbb{N}, \tag{28}$$

where $\mathbf{y}_t := (y_t : \mathbf{X}_t)^\top, \mathbf{B} := (\boldsymbol{\beta}_1 : \mathbf{B}_2)^\top, \mathbf{u}_t := (u_{1t} : \mathbf{u}_{2t})^\top$.

Note that the testable assumptions for the model in (28) are multivariate versions of [1]–[5] in Table 2. The link between statistical model in (28) and structural model in (23) stems from a reparameterization/restriction of $f(y_t, \mathbf{x}_t \mid \mathbf{z}_t; \boldsymbol{\psi}_1)$ in (28) based on $f(y_t \mid \mathbf{x}_t, \mathbf{z}_t; \boldsymbol{\varphi}_1)$ (Spanos, 2007b)

$$y_t = \boldsymbol{\alpha}_0^\top \mathbf{x}_t + \boldsymbol{\gamma}_0^\top \mathbf{z}_t + \upsilon_t \underset{\text{s.t. (iii)–(v)}}{\overset{\gamma_0 = \mathbf{0}}{\longrightarrow}} y_t = \boldsymbol{\alpha}^\top \mathbf{X}_t + \varepsilon_t, \tag{29}$$

where "s.t." denotes "subject to." This reparameterization/restriction also explains why the absence of a direct arrow between $\mathbf{Z}$ and $Y$ is relevant only for the structural and not the statistical model. The causal diagram 1 indicates that $\mathbf{Z}$ has no separate effect on $Y$, except through $X$.

This recasting shows that beyond the structural model, (28) is the implicit statistical model underlying the appropriate choice of $\mathbf{Z}_t$, with the conditions (i) $E(\mathbf{X}_t \varepsilon_t) \neq \mathbf{0}$ and (ii) $E(\mathbf{Z}_t \varepsilon_t) = \mathbf{0}$ determining the parameterization of $(\boldsymbol{\alpha}, \sigma_\varepsilon^2)$. Indeed, the sampling distributions of the IV estimators of $(\boldsymbol{\alpha}, \sigma_\varepsilon^2)$, and tests of the overidentifying restrictions (Greene, 2018), depend crucially on the validity of the probabilistic assumptions [1]–[5] for data $\mathbf{W}_0$. Also, when the statistical adequacy of model in (28) is secured, conditions (iii)–(v) are rendered testable in its context. Owen (2017) applies the above procedure to empirical examples in several published papers relying on IV methods and shows that the model in (28) is invariably misspecified, and condition (v) is usually invalid, calling into question the appropriateness of the chosen $\mathbf{Z}_t$ and the trustworthiness of the ensuing evidence.

When the statistical model in (28) is statistically misspecified, the traditional story-telling justification has no merit. One needs to respecify the model in (28) to account for the unaccounted statistical information, say adding trends and lags when [4] and [5] are found wanting. Any additional terms needed to secure the statistical adequacy of the respecified model became part of the new set of instruments; see Spanos (2007b).

## 6 | NONPARAMETRIC STATISTICS

As mentioned in the introduction, parametric and nonparametric modeling differ in one important respect: instead of a direct distribution assumption (Normal, Poisson, etc.), a nonparametric statistical model $\mathcal{M}_{\mathcal{F}}(\mathbf{x})$ is specified in terms of a broader family of distributions $\mathcal{F}$ defined in

terms of *indirect* and *non-testable* indirect distribution assumptions such as: (a) the existence of moments up to order $p \geq 1$, (b) smoothness restrictions on the *unknown* density function $f(x)$, $x \in \mathbb{R}_X$ (symmetry, differentiability, unimodality, boundedness, and continuity of derivatives of $f(x)$ up to order $m > 1$); see Thompson and Tapia (1990). The generic statistical models for the two approaches are specified as follows

$$\begin{array}{ll} \text{Parametric}: & \mathcal{M}_\theta(\mathbf{x}) = \{f(\mathbf{x};\theta), \theta \in \Theta \subset \mathbb{R}^m\}, \quad \mathbf{x} \in \mathbb{R}_X^n, \quad n > m, \\ \text{Nonparametric}: & \mathcal{M}_{\mathcal{F}}(\mathbf{x}) = \{f(\mathbf{x};\boldsymbol{\psi}_n), f \in \mathcal{F}\}, \quad \mathbf{x} \in \mathbb{R}_X^n, \end{array} \tag{30}$$

where $\boldsymbol{\psi}_n$ indicates that the parameters can depend on $n$; see Wasserman (2006).

The period of nonparametric inference driven by kernel smoothing began in the late 1970s with the emphasis placed on estimating the density and related functions, building on Kolmogorov (1933) result that for an IID sample $\mathbf{X} := (X_1, X_2, \ldots, X_n)$ the empirical cumulative distribution function (ecdf) $\widehat{F}_n(x)$ is a good estimator of the cdf $F(x)$, for $n$ large enough. Attempts to derive consistent estimators of the density function $f(x)$, $x \in \mathbb{R}$, led to (a) the kernel smoothing and related techniques that include regression and other nonparametric models, and (b) *series estimators*, $\widehat{f}(x) = \sum_{i=0}^{m} \beta_i \phi_i(x_k)$, where $\{\phi_i(x_k),\ i = 1, 2, \ldots, m\}$ are (known) polynomials, usually orthogonal; see Silverman (1986).

To illustrate the dangers of indirect distribution assumptions, Bahadur and Savage (1956) explored replacing the Normality assumption in Equation (2) with a broader *family* of distributions $\mathcal{F}$ whose first two moments exist. The question they posed was: "Is there a reasonably reliable and precise test for the hypotheses, $H_0$: $\mu \leq 0$, versus $H_1$: $\mu > 0$, in the context of the family $\mathcal{F}$, analogous to the t-test? They show that such a $t$-type test is biased and inconsistent (p. 1115). The intuition underlying this result is that the existence of the first two moments provides insufficient information concerning the tails of the unknown distribution $f(x)$ to establish an informative threshold $c_\alpha$, and thus one needs to put further restrictions on $\mathcal{F}$ to get a consistent estimator, test, CI; consistency is a minimal property.

This example is a warning to practitioners striving to weaken the model assumptions and proceed to rely on limit theorems (as $n \to \infty$) for their inferences, ignoring the fact that the reliability and precision of inference depend solely on the approximate validity of the probabilistic assumptions imposed (implicitly or explicitly) on the particular data $\mathbf{x}_0$ and nothing else. As argued by Le Cam (1986), p. xiv

"... limit theorems 'as $n$ tends to infinity' are logically devoid of content about what happens at any particular $n$. All they can do is suggest certain approaches whose performance must then be checked on the case at hand. Unfortunately, the approximation bounds we could get are too often too crude and cumbersome to be of any practical use."

This calls into question the reliability of nonparametric inference that relies solely on consistent and asymptotically normal (CAN) estimators and the associated testing and prediction procedures, especially when the dependence and heterogeneity assumptions invoked by the limit theorems have not been validated.

**Nonparametric regression**.

A typical example is a regression curve

$$y_k = m(x_k) + \varepsilon_k, \tag{31}$$

where $m(x_k) = E(y_k \mid X_k = x_k)$ is an *unknown function*, assuming that the error term $\varepsilon_k$ is white-noise. By definition

$$m(x) = E(y \mid X = x) = \int_{y \in \mathbb{R}_Y} y f(y \mid x) dy = \int_{y \in \mathbb{R}_Y} y \frac{f(y,x)}{f(x)} dy. \tag{32}$$

Using estimators for $f(y,x)$ and $f(x)$ based on kernel ($\mathbb{K}(z)$) smoothing with bandwidth $h$, and assuming that (i) $\mathbb{K}(z) \geq 0, z \in \mathbb{R}$

$$\widehat{f}(y,x) = \frac{1}{nh^2} \sum_{i=1}^{n} \mathbb{K}_1\Big(\frac{x - x_i}{h}\Big)\mathbb{K}_0\Big(\frac{y - y_i}{h}\Big), \quad \widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} \mathbb{K}_1\Big(\frac{x - x_i}{h}\Big), \tag{33}$$

Adding the restriction (ii) $\int \mathbb{K}_0(y)dy = 1 \longrightarrow \int \widehat{f}(y,x)dy = \frac{1}{nh} \sum_{i=1}^{n} \mathbb{K}_1(\frac{x-x_i}{h})$.

Adding the restriction (iii) $\int y\mathbb{K}(y)dy = 0$ implies that

$$\int y\widehat{f}(y,x)dy = \frac{1}{nh^2} \sum_{i=1}^{n} \mathbb{K}_1\Big(\frac{x - x_i}{h}\Big) \int y\mathbb{K}_1\Big(\frac{x - x_i}{h}\Big)dy = \frac{1}{nh} \sum_{i=1}^{n} y_i\mathbb{K}_1\Big(\frac{x - x_i}{h}\Big), \tag{34}$$

since $\int y\mathbb{K}_1(\frac{y-y_i}{h})dy = y_i$, yielding the *Nadaraya–Watson* kernel estimator

$$\widehat{m}(\mathbf{x}) = \left[\sum_{i=1}^{n} \mathbb{K}_1\Big(\frac{x - x_i}{h}\Big)y_i / \sum_{i=1}^{n} \mathbb{K}_1\Big(\frac{x - x_i}{h}\Big)\right] = \sum_{i=1}^{n} w_i y_i. \tag{35}$$

Assuming that $\{(x_k, y_k), k \in \mathbb{N}\}$ is an IID process, and the following mathematical assumptions (a) the support of $X$, $\mathbb{R}_X := \{x : f(x) > 0\}$, $x \in \mathbb{R}$, is a proper subset of $\mathbb{R}$, (b) $x$ is an interior point of $\mathbb{R}_X$, (c) $f^{(i)}(x) = \frac{d^i f(.)}{dx^i}, i = 0, 1, 2, 3$, are uniformly continuous on $\mathbb{R}_X$, (d) in addition to properties (i)–(iii) $\mathbb{K}(.)$ satisfies the restriction $\kappa_2(x) = \int x^2\mathbb{K}(x)dx > 0$, (e) $n \to \infty, h_1 \to 0$, and $nh_1 \to \infty$, it can be shown that

$$(\widehat{m}_n(\mathbf{x}) - m(x)) \underset{n\to\infty}{\backsim} \mathsf{N}(\text{bias}(\widehat{m}(\mathbf{x})), V_\infty(\widehat{m}(\mathbf{x}))). \tag{36}$$

This can be used to define the asymptotic risk (AMSE):

$$R(\widehat{m}, m) = \frac{h^4}{4}\left(\int x^2\mathbb{K}_1(x)dx\right)^2 \left[\int \left(\frac{d^2m(.)}{dx^2} + 2\frac{dm(x)}{dx}\left[\frac{df(x)}{dx}/f(x)\right]\right)^2 dx\right.$$
$$\left. + \frac{\sigma^2}{mh}\left(\int \mathbb{K}_1^2(x)dx\right)\int [1/f(x)]dx + o(nh) + o(h^4)\right. \tag{37}$$

where $R(\widehat{m}, m) = E(\frac{1}{n}\sum_{i=1}^{n}[\widehat{m}(x_i) - m(x_i)]^2)$, the first term in Equation (37) is the $[\text{bias}(\widehat{m}(x))]^2$, the second is the $V_\infty(\widehat{m}(x))$, and "$c_n = o(nh)$" denotes that $(c_n/nh) \underset{n\to\infty}{\to} 0$; see Wasserman (2006).

The fact that $R(\widehat{m}(x), m(x))$ depends on the unknown functions $m(x)$, $f(x)$ we seek to learn about, and their derivatives, renders such an approximation result highly questionable for inference purposes. In this sense, Equation (37) is much worse than traditional asymptotic results since

its bounds involve the very functions we want to learn about. This, combined with the fact that most of the invoked assumptions (a)–(e) in Equation (37) are non-testable, call into question the reliability and merit of such asymptotic results; see Wasserman (2006).

In summary, the main features of the nonparametric approach are:

[a] **Underlying framework**: the underlying probabilistic framework is defined on a Kolmogorov probability space $(S, \mathfrak{F}, \mathbb{P}(.))$, combined with additional mathematical structure imposed on the relevant curves to be fitted. The inference revolves around the concept of a prespecified nonparametric statistical model $\mathcal{M}_{\mathcal{F}}(\mathbf{x})$.
[b] **Inductive premises**: the probabilistic assumptions comprising $\mathcal{M}_{\mathcal{F}}(\mathbf{x})$ come from the two categories, Dependence and Heterogeneity (often IID), but the direct distribution assumption is replaced by certain mathematical restrictions on the density function or other curves of interest. $\mathcal{M}_{\mathcal{F}}(\mathbf{x})$ is viewed as a substantive model.
[c] **Model choice**: the appropriate $\mathcal{M}_{\mathcal{F}}(\mathbf{x})$ is chosen on goodness-of-fit grounds.
[d] **Quantification**: the estimation of $\mathcal{M}_{\mathcal{F}}(\mathbf{x})$ relies on curve-fitting methods that are justified in terms of particular loss functions based on information other than $\mathbf{x}_0$.
[e] **Inductive inference**: the interpretation of probability is frequentist and the effectiveness (optimality) of inference procedures is framed in terms of asymptotic properties for estimators, predictors, and tests.
[f] **Substantive versus statistical**: $\mathcal{M}_{\mathcal{F}}(\mathbf{x})$ is (implicitly) viewed as a substantive model which is foisted on data $\mathbf{x}_0$, ignoring the validity inductive premises comprising the implicit statistical model.

What are the consequences of replacing $f(\mathbf{x};\boldsymbol{\theta})$ with $f(\mathbf{x};\boldsymbol{\psi}_n),\ f\in\mathcal{F}$? The most important consequence is that likelihood-based inference procedures are replaced by loss function-based procedures driven by mathematical approximation theory and goodness-of-fit measures. This is often touted as a great advantage of the nonparametric inference. For instance, Dickhaus (2018) argues that: "Of course, the advantage of considering $\mathcal{F}$ is that the issue of model misspecification, which is often problematic in parametric models, is avoided." (p. 13). This claim is questionable on several additional grounds.

First, the least problematic assumption in a statistical model $\mathcal{M}_{\theta}(\mathbf{x})$ is the distribution, since it is trivial to test and respecify; see Spanos (2019), ch. 5. The dependence and heterogeneity assumptions are a lot more difficult to test, and respecify $\mathcal{M}_{\theta}(\mathbf{x})$, or $\mathcal{M}_{\mathcal{F}}(\mathbf{x})$ when invalid. The above-quoted claim ignores the fact that nonparametric statistical models always impose *dependence* and *heterogeneity* assumptions (often IID), which are highly restrictive for observational data.

Second, the "indirect" and non-testable distribution assumptions invariably contribute substantially to the imprecision of inference since asymptotic bounds blunt the reliability/precision of inference; see Le Cam (1986) quotation in Section 6.

**Example 6.** Let $X \backsim \mathsf{U}(-\theta,\theta)$, denoting a uniform distribution:

$$f(x;\theta)=\frac{1}{2\theta},\quad E(X)=0,\quad Var(X)=\frac{\theta^2}{3},\quad -\theta\leq x<\theta, \tag{38}$$

and one is interested in evaluating the tail area, $\{|\,X-E(X)\,|>(1.645)\sqrt{Var(X)}\}$, as one would do in testing or constructing CIs. Comparing the upper bound given by Chebyshev's inequality

$\mathbb{P}(|\ X\ | \geq 1.645\sqrt{Var(X)}) \leq \frac{1}{(1.645)^2} = .37$, with $\mathbb{P}(|\ X\ | \geq .949748\theta)) = .05$, shows that the bound is highly inaccurate.

Third, weaker premises do not imply less vulnerability to statistical misspecification just because of $\mathcal{M}_\theta(\mathbf{x}) \subset \mathcal{M}_\mathcal{F}(\mathbf{x})$. What renders $\mathcal{M}_\theta(\mathbf{x})$ coveted is that statistical adequacy in conjunction with M-S testing can guide the search for $\mathcal{M}^*(\mathbf{x}), \mathbf{x} \in \mathbb{R}^n_X$, but $\mathcal{M}_\mathcal{F}(\mathbf{x})$ makes M-S testing impracticable; see Spanos (2018). Instead, nonparametric inference presumes that $\mathcal{M}_\mathcal{F}(\mathbf{x})$ includes $\mathcal{M}^*(\mathbf{x})$, which is unreasonably optimistic since the invoked probabilistic assumptions include restrictive dependence and heterogeneity restrictions; often IID.

Fourth, in nonparametric inference, the moments of the underlying distribution $f(x),\ x \in \mathbb{R}_X$, are invariably estimated using the corresponding sample moments: $\frac{1}{n}\sum_{i=1}^n X_i^r,\ r = 1, 2, ...$, and $\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2, r = 2, 3, ....$ Although reasonable for some distributions, it is not appropriate for others.

**Example 7.** In the case of the uniform distribution $X \backsim \mathsf{U}(\theta - \frac{1}{2}, \theta + \frac{1}{2})$, $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ is not a good estimator of $\theta$ when compared to the ML estimator $\widehat{X}_{ML} = \frac{1}{2}(X_{[n]} + X_{[n]})$, since $Var(\overline{X}) = n^{-1} > Var(\widehat{X}_{ML}) = [2(n+1)(n+2)]^{-1}$; see Spanos (2011).

Fifth, contrary to the conventional wisdom, excellent goodness-of-fit/prediction relative to a particular loss function is neither necessary nor sufficient for the statistical adequacy of a parametric $\mathcal{M}_\theta(\mathbf{x})$ or a nonparametric $\mathcal{M}_\mathcal{F}(\mathbf{x})$ model; Spanos (2007a). Intuitively, small residuals are not equivalent to non-systematic (e.g., white-noise) residuals.

In summary, the crucial difference between parametric and nonparametric inference is that the easily testable distribution assumption of $\mathcal{M}_\theta(\mathbf{x})$ is replaced with untestable mathematical restrictions at a high price: the finite sample optimal inferences stemming from likelihood-based procedures, whose appropriateness can be established by M-S testing, are jettisoned in favor of asymptotically justified (as $n \to \infty$) inferential procedures (i) whose reliability is impossible to establish, and (ii) their precision has been undermined using weaker probabilistic assumptions; see Spanos (2001, 2018). The same comments apply to nonparametric inference based on local and orthogonal polynomials since the choice of the best-fitted curve relies on risk functions with similar problems as Equation (37); see Wasserman (2006).

What about using *bootstrapping* and other resampling techniques (Efron & Hastie, 2016) to avoid the perils of "as $n \to \infty$" asymptotics? The catch is that the bootstrap is a lot more vulnerable to statistical misspecification than other procedures because resampling the original data $\mathbf{x}_0$ overutilizes the IID assumptions. Hence, when the IID assumptions are invalid for data $\mathbf{x}_0$, the bootstrap statistics are highly unreliable; see Spanos (2019), ch. 10. Extending the simple bootstrap to more sophisticated resampling and sub-sampling methods rarely works in practice, despite their good "asymptotic" (!) properties; see Lahiri (2003). The problem is that reliable resampling requires one to generate data realizations that constitute faithful replicas (in terms of the chance regularities) of the original data $\mathbf{x}_0$. That, however, requires one to establish the chance regularities exhibited by the original data $\mathbf{x}_0$ by applying thorough M-S testing to test the validity of all the probabilistic assumptions comprising $\mathcal{M}_\mathcal{F}(\mathbf{x})$, and then ensure that the replicated data $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$ enjoy the same chance regularities using M-S testing; the very objective resampling methods seek to sidestep.

## 7 | BIG DATA AND DS

DS includes ML, SLT, pattern recognition, data mining, and certain reformulated statistical techniques with the emphasis placed on large data sets, computation, and predictive learning. These include supervised learning, unsupervised learning, online learning, reinforcement learning, and deep learning; see Murphy (2012).

DS shares many similarities with nonparametric inference in terms of its setup and curve-fitting, but it differs from the latter in three respects. First, DS methods rely on huge sample sizes that can be divided into training, validation, and test subsets that are sufficiently large. Second, the additional mathematical structure employed in curve-fitting (mathematical approximation) takes place in the context of normed linear spaces, with sufficient mathematical structure to ensure the existence and uniqueness of the fitted curve; see Luenberger (1969). Third, the fitted curve is used primarily for prediction purposes by extrapolation.

This section attempts to place DS in the broader context of "learning from data" with a view to bring out certain foundational issues that distinguish it from the model-based statistical induction. The first foundational issue is that the perspective dominating ML is primarily one of the *mathematical approximation theory* based on *functional analysis* (Luenberger, 1969; Powell, 1981) where data modeling is dominated by linear algebra-based optimization (Aggarwal, 2020), with probability theory being used as an add-on to avail the derivation of asymptotic inference procedures.

### 7.1 | Big Data and ML

Broadly speaking "predictive learning," as understood today in DS, began at the intersection of neuroscience and computer science in the 1980s with some apparent success of neural networks and related methods modeled loosely on the human brain. Neural networks consist of several layers of multiple simple processing nodes that are densely interconnected aiming to capture patterns in large data sets using linear combinations of inputs to define several hidden layers and then using local polynomials (splines, logistic, etc.) of linear combinations of these hidden layers giving rise to outputs; see Hastie et al. (2017) and Murphy (2012). The early promise of computation-intensive model-free modeling, and the enthusiasm of the apparent success of neural nets, subsided when it was realized that neural nets constitute a quintessential example of curve-fitting with overparametrized nonlinear models that have serious limitations for predictive learning. This raised the problem of overfitting and various attempts to tame it led to several techniques known as *regularization*: constrains that shrink the coefficient estimates towards zero to countercheck the overfitting.

This solution to the overfitting problem encouraged the research in ML in the 1990s to view their flexible nonlinear curve-fitting in the context of the traditional statistical framework with one important difference. The traditional techniques of least-squares and likelihood-based optimizations based on the statistical model $\mathcal{M}_\theta(\mathbf{z})$ are replaced by minimizations based on loss functions with additional regularization constraints defining the predictive risk function; see Clarke et al. (2009).

## 7.2 | ML and inference: a brief summary

The literature on ML and the statistical Learning Theory (SLT) has grown exponentially since the 1990s, and any attempt to provide an overview will be a hopeless task; see Aggarwal (2020), Iske (2018), and Murphy (2012). Hence, the discussion in this section focuses on a generic problem that brings out the main features of DS (ML, SLT, etc.).

DS frames the problem of statistical inference in terms of how a machine can "learn from data," and framed in terms of *learner's input* [a domain set $\mathcal{X}$, a label set $\mathcal{Y}$, and training data $\mathcal{X} \times \mathcal{Y}$: $\mathbf{z}_i := (x_i, y_i), i = 1, 2, \ldots, n$, with an unknown distribution behind the data], and *learner's output*: $h(.) : \mathcal{X} \to \mathcal{Y}$. The learning algorithm is all about choosing $h(.)$ to approximate "closely" the true relationship $y_i = f(x_i),\ i \in \mathbb{N}$, by minimizing the distance $\| h(x) - f(x) \|$; see Shalev-Shwartz and Ben-David (2014).

A case can be made that DS modeling has taken empirical modeling back to the Pearson data-driven curve-fitting by replacing the Pearson family $\mathcal{F}_P$ with a broader set of mathematically smooth functions $\mathcal{G}$ that goes beyond frequency curves and include all functional dependence relationships. These families of unknown functions $\mathcal{G}$ are embedded in a normed linear space with sufficient mathematical structure to ensure completeness, compact subsets, and convex norms. The classical literature on mathematical approximation theory from Euler to Chebyshev and Bernstein (Karl-Georg, 2006) found a natural home in functional analysis, which unified the classical theory and made available powerful tools for obtaining new results, such as the open mapping theorem, the Banach-Steinhaus theorem, and the Hahn-Banach theorem; see Kreyszig (1978). For instance, Fisher's discriminant analysis has been reformulated as an application of the Hahn–Banach theorem to derive the optimal separating hyperplanes; see Luenberger (1969) and Hastie et al. (2017).

**Example 8.** The normed linear space $(C[a, b], \| \ . \ \|_p)$ of all real-valued continuous functions $f(x)$ defined on $[a, b] \subset \mathbb{R}$, with the $p$-norm (Deutsch, 2001):

$$\|f\|_p = \left( \int_a^b |f(x)|^p dx \right)^{\frac{1}{p}}, \quad \text{or} \quad \|f\|_p = \left( \sum_{i=0}^{n} |f(x_i)|^p \right)^{\frac{1}{p}}, \quad p = 1, 2, \infty. \tag{39}$$

In such a context, the approximation problem is transformed into an optimization using vector space methods.

The approximation problem is often viewed in the context of a complete inner product vector (linear) space $(C[a, b], \| \ . \ \|_2)$ of real-valued functions $h(X)$, defined on $[a, b] \subset \mathbb{R}$, with the 2-norm, also known as a **Hilbert space** of square-integrable functions $(E(|X_t|^2))$. Intuitively, a Hilbert space extends the methods of vector algebra and calculus from the Euclidean space, that revolves around $\mathbb{R}^n$, to spaces (sets of elements with certain mathematical structure) to any finite or infinite number of dimensions. A Hilbert space is a vector (linear) space $(\mathbf{V}, < ., . >)$ whose geometric structure is based on an inner product $< \mathbf{v}_1, \mathbf{v}_2 >$, where $\mathbf{v}_1, \mathbf{v}_2$ are elements of $\mathbf{V}$, an operation that allows lengths and angles to be defined to render (geometric) optimization possible. Hilbert spaces are complete in terms of their metric in the sense that they include the limits of their sequences to render differential and integral calculus possible.

**Geometric structure**.

A Hilbert space $(\mathbf{V}, < ., . >)$ is rich in geometric structure since the inner product $< ., . >$ (think $Cov(X_t, X_s) = E[(X_t - \mu)(X_s - \mu)]$) can be used to define orthogonality, $< \mathbf{v}_1, \mathbf{v}_2 >= 0 \Longrightarrow$

$\mathbf{v}_1 \perp \mathbf{v}_2$, [think $Cov(\varepsilon_t, \varepsilon_s) = 0,\ t \neq s$] and implied norm $\|\mathbf{v}_1\| = \sqrt{< \mathbf{v}_1, \mathbf{v}_1 >}$ (length/magnitude) [think $\sqrt{Var(X_t)}$] and a metric $d(\mathbf{v}_1, \mathbf{v}_2) = \|\mathbf{v}_1 - \mathbf{v}_2\|$ (distance) [think $E(X_t - E(X_t \mid \sigma(X_{t-1}))^2$]. A very important result in the context of a Hilbert space is

**Orthogonal projection theorem**.

For any $\mathbf{x} \in \mathbf{V}$, and $\mathbf{V}_1$ a closed proper subset of a Hilbert space $\mathbf{V}$, $\mathbf{x}$ can be uniquely expressed in the form of

$$\mathbf{x} = \hat{\mathbf{x}} + \mathbf{u}, \quad \text{where } \hat{\mathbf{x}} \in \mathbf{V}_1 \quad \text{and} \quad \mathbf{u} \text{ is orthogonal to } \mathbf{V}_1, \tag{40}$$

denoted by $\mathbf{u} \perp \mathbf{V}_1$, ensuring that:$\|\mathbf{x} - \hat{\mathbf{x}}\| \leq \|\mathbf{x} - \mathbf{v}_1\|$ for all $\mathbf{v}_1 \in \mathbf{V}_1$, is minimized if and only if $\mathbf{v}_1 = \hat{\mathbf{x}}$; where $\hat{\mathbf{x}}$ denotes the orthogonal projection of $\mathbf{x}$ on $\mathbf{V}_1$.

**Regression**.

A typical example of a SLT for a regression model:

$$y = g(\mathbf{x}; \boldsymbol{\psi}_n) + \upsilon_t, \tag{41}$$

with data $\mathbf{z} := (\mathbf{x}, y)$ would specify a risk functional of the form:

$$R(f^*, g) = E(L(y, g(\mathbf{x}; \boldsymbol{\psi}_n))) = \int_{\mathbf{z}} L(y, g(\mathbf{x}; \boldsymbol{\psi}_n)) f^*(\mathbf{z}) d\mathbf{z}. \tag{42}$$

$L(y, g(\mathbf{x}; \boldsymbol{\psi}_n))$ denotes a loss function, $g(\mathbf{z}; \boldsymbol{\psi}_n) \in \mathcal{G}$, where $\mathcal{G}$ is a family of smooth enough mathematical functions whose parameters could vary with $n$, and $f^*(\mathbf{z})$ is the joint distribution of the random vector $\mathbf{Z}$, which is *unknown*. It is important to note that the notion of a statistical model implicit in SLT is

$$\mathcal{M}_{\mathcal{F}}(\mathbf{z}) = \{f(\mathbf{z}; \boldsymbol{\psi}_n), f \in \mathcal{F}\}, \quad \mathbf{z} \in \mathbb{R}^n_Z, \tag{43}$$

where the number of parameters $\boldsymbol{\psi}_n$ to be estimated depends on $n$; see Murphy (2012).

**Risk functions**.

Commonly used loss functions are

$$\text{(a) } L(y, g(\mathbf{x}; \boldsymbol{\psi}_n)) = (y - g(\mathbf{x}; \boldsymbol{\psi}_n))^2, \quad \text{(b) } L(y, g(\mathbf{x}; \boldsymbol{\psi}_n)) = |y - g(\mathbf{x}; \boldsymbol{\psi}_n)|, \tag{44}$$

whose expected value defines a risk function $R(f^*, g(\mathbf{z}; \boldsymbol{\psi}_n)) = E[L(y, g(\mathbf{x}; \boldsymbol{\psi}_n))]$. Since $R(f^*, g(\mathbf{z}; \boldsymbol{\psi}_n))$ is not observable, one could minimize the corresponding empirical risk function:

$$\widehat{g}_m(\mathbf{z}; \boldsymbol{\psi}_n) = \arg\min_{g \in \mathcal{G}} \widehat{R}(f^*, g), \quad \text{where } \widehat{R}(f^*, g) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, g(\mathbf{x}_i; \boldsymbol{\psi}_n)), \tag{45}$$

which, under the IID assumptions, provides a consistent estimator of $R(f^*, g(\mathbf{z}; \boldsymbol{\psi}_n))$. Since the minimization is over different regression functions $g \in \mathcal{G}$, to ensure that $\widehat{g}(\mathbf{z}; \boldsymbol{\psi}_n)$ is a consistent estimator of the unknown $g(.)$, one needs a LLN that holds for all functions in the class $\mathcal{G}$:

$$\lim_{n \to \infty} \mathbb{P}(\sup_{g \in \mathcal{G}} |\widehat{R}(f^*, g(\mathbf{z}; \boldsymbol{\psi}_n)) - R(f^*, g(\mathbf{z}; \boldsymbol{\psi}_n))| > \varepsilon) = 0, \tag{46}$$

known as the ULLN, which holds when the vector stochastic process $\{\mathbf{Z}_t, t \in \mathbb{N}\}$ is IID; see Vapnik and Chervonenkis (1991) and Schölkopf et al. (2013).

**PAC learnable**.

The primary aim of DS modeling is to learn about $f^*(\mathbf{z})$ by selecting a function $g(\mathbf{z};\boldsymbol{\psi}_n)$ from the class $\mathcal{G}$, which takes the form of *probably approximately correct* (PAC) learning that concerns how $f^*(\mathbf{z})$ relates to $g(\mathbf{z};\boldsymbol{\psi}_n)$ in $\mathcal{G}$. The concept of PAC learning was initially proposed by Valiant (1984) and defined as: "the process of learning from examples, where the number of computational steps is polynomially bounded and the errors are polynomially controlled." (Valiant, 2013), p. 187. That is, $\mathcal{G}$ is said to be PAC-learnable if there is a polynomial-time algorithm that can identify a function $g(\mathbf{z};\boldsymbol{\psi}_n)$ in $\mathcal{G}$ that is PAC learnable. The learning amounts to constructing an upper bound relating to the difference between $\widehat{R}(f^*, g(\mathbf{z};\boldsymbol{\psi}_n))$ and $R(f^*, g(\mathbf{z};\boldsymbol{\psi}_n))$ for any probability distribution $f^*$ and training data $\mathbf{Z}_0$ of sample size $N$, and the dimensionality of the class of functions $\mathcal{F}$ (Murphy, 2012)

$$\mathbb{P}(\max_{g\in\mathcal{G}}|\widehat{R}(f^*, g(\mathbf{z};\boldsymbol{\psi}_n)) - R(f^*, g(\mathbf{z};\boldsymbol{\psi}_n))| > \epsilon) \leq 2\,\dim(\mathcal{G})\exp(-2N\epsilon^2). \tag{47}$$

When the parameters $\boldsymbol{\psi}_n$ are real-valued, the relevant dimension is the V-C.

The PAC learning has a computational dimension that relates to learning by computation being completed in a finite number of steps—polynomially bounded ($n^k$), and a statistical dimension that relates to how probability is integrated into the learning process; see Valiant (2013). Unfortunately, the emphasis is currently placed mostly on the former at the neglect of the latter, which is treated as an afterthought by invariably assuming an IID probabilistic structure for the data, and no tests for their validity are performed as part of the learning process, rendering PAC learning is highly vulnerable to statistical misspecification. As argued by Valiant (2013): "When first introduced [PAC], the corresponding class was simply called learnable,... The 'probably' acknowledges misfortune errors, and the 'approximately' rarity errors." (p. 71). According to Valiant (2013), pp. 65–66: "... while neither of these two sources of error can be totally eliminated, both can be controlled." That is, their probability of occurrence can be made arbitrarily small by increasing the sample size $N$. Regrettably, this claim neglects the fact that "misfortune errors," include departures from IID, such as the presence of dependence and heterogeneity, that can easily increase their probability of occurrence as $N$ increases; see Spanos and McGuirk (2001).

**Regularization**.

Depending on the capacity (complexity) of the class of functions $\mathcal{G}$, Equation (45) will often give rise to over-fitting. To counter that one needs to use a constrained *regularized* risk function:

$$R_r(f^*, g(\mathbf{z};\boldsymbol{\psi}_n)) = R(f^*, g(\mathbf{z};\boldsymbol{\psi}_n)) + \lambda C(g(\mathbf{z};\boldsymbol{\psi}_n)). \tag{48}$$

where $C(g(\mathbf{z};\boldsymbol{\psi}_n))$ is often related to the V-C dimension of the class $\mathcal{G}$, which is measured in terms of computational complexity; see Cherkassky and Mulier (2007) and Schölkopf et al. (2013).

As mentioned above, DS methods rely on huge sample sizes that can be divided into training, validation, and test subsets, which are sufficiently large in size. The idea is that the "training" data set is used to select $\widehat{g}(\mathbf{x}_i;\boldsymbol{\psi}_n)$, and the "validation" data set is used to evaluate its predictive ability and fine-tune the selected model, such as choosing the parameter $\lambda$ for the regularization of $\widehat{g}(\mathbf{x}_i;\boldsymbol{\psi}_n)$, or employ Akaike-type model selection procedures, or cross-validation and resampling procedures, to modify $\widehat{g}(\mathbf{x}_i;\boldsymbol{\psi}_n)$ into $\widehat{g}_m(\mathbf{x}_i;\boldsymbol{\psi}_n)$. The "test data" is used to provide a final "unbiased" evaluation of the modified $\widehat{g}_m(\mathbf{x}_i;\boldsymbol{\psi}_n)$ on the training data set. Moreover, having such

large data sets encourages practitioners to use asymptotic results, such as the ULLN in Equation (46), and CAN estimators, and related tests based on IID samples, but is that confidence justified? The short answer is that the training/validation/testing split can improve the selected models on prediction grounds, but would not secure the reliability of inference when the selection of models is guided by goodness-of-fit/prediction.

## 7.3 | DS methods: a brief appraisal

In summary, the main features of the DS approach are the following:

[a] **Underlying framework**: the underlying probabilistic framework is defined on a Kolmogorov probability space $(S, \mathfrak{F}, \mathbb{P}(.))$, combined with the additional mathematical structure of a normed space $L_p(S, \mathfrak{F}, \mathbb{P}(.))$ in the context of which the stochastic process $\{\mathbf{Z}_t, t \in \mathbb{N}\}$ is defined. The inference revolves around a family of smooth enough mathematical functions $\mathcal{G}$.
[b] **Inductive premises**: the probabilistic assumptions underlying $\{\mathbf{Z}_t, t \in \mathbb{N}\}$ come from the two categories, Dependence and Heterogeneity, (often IID), but the direct distribution [D] assumption is replaced by certain mathematical restrictions on $\mathcal{G}$. The curves $g(\mathbf{x}_t; \boldsymbol{\psi}_n)$ in $\mathcal{G}$ are viewed as substantive models.
[c] **Model choice**: the appropriate $\widehat{g}_m(\mathbf{x}_t; \boldsymbol{\psi}_n)$ in $\mathcal{G}$ is chosen on goodness-of-fit/prediction grounds using Akaike-type information criteria.
[d] **Quantification**: the estimation of $\widehat{g}_m(\mathbf{x}_t; \boldsymbol{\psi}_n)$ in $\mathcal{G}$ relies on curve-fitting methods based on minimization subject to restrictions based on particular loss functions based on information other than the data.
[e] **Inductive inference**: the interpretation of probability can be both frequentist and Bayesian and the effectiveness (optimality) of inference procedures is framed in terms of asymptotic properties for estimators, predictors, and tests.
[f] **Substantive versus statistical**: $\mathcal{G}$ is broadly viewed as a family of substantive models when $g_m(\mathbf{x}_t; \boldsymbol{\psi}_n)$ is foisted on $\mathbf{Z}_0$, ignoring the validity of the implicit $\mathcal{M}_\theta(\mathbf{z})$.

Appraising the methods and procedures of DS in terms of learning from data about phenomena of interest, one can make a strong case that: (i) DS is a mathematically more sophisticated form of Karl Pearson's data-driven descriptive statistics and the nonparametric curve-fitting, (ii) guided by impromptu probabilistic assumptions (often IID) to justify the invoked mathematical approximation theory based on norm-minimization subject to restrictions (prediction risk), and (iii) evaluated using goodness-of-fit/prediction measures and empirical risk functions based on arbitrary loss functions. The only significant differences between nonparametric statistics and DS methods is (a) the mathematically richer framework and (b) the huge data sets enabling practitioners to provide additional goodness-of-fit/prediction safeguards, in the form of cross-validation and related procedures, and enhance the precision of their inferences, or so it is hoped for. As such, DS inherits all the problems and weaknesses associated with nonparametric statistics and Karl Pearson's approach.

Additional problems for DS methods stem from its combining two potentially incompatible forms of convergence, mathematical approximation convergence results (as $m \to \infty$), where $m$ denotes the degree of the approximation polynomial, with probabilistic limit theorems (as $n \to \infty$). Also, the various *alternation theorems* (Powell, 1981) that characterize the optimum $g_m(x_k; \boldsymbol{\alpha})$ in terms of alternating signs for the approximation error, often give rise to statistically systematic

residuals $\widehat{\varepsilon}_t = y_k - g_m(x_k;\widehat{\boldsymbol{\alpha}})$ (non-white noise) that render $y_k = g_m(x_k;\widehat{\boldsymbol{\alpha}})$ statistically misspecified; see Spanos (2010b).

When DS methods are viewed from the model-based ($\mathcal{M}_{\theta}(\mathbf{z})$) perspective, several weaknesses are raised.

First, they rely exclusively on goodness-of-fit/prediction, and do not provide: (i) testable probabilistic assumptions to ensure the validity of the premises for inductive inference and secure the reliability of inference, or/and (ii) ascertainable error probabilities to assess the effectiveness of inference procedures for the particular data $\mathbf{Z}_0$. However, excellent goodness-of-fit/prediction does not ensure the reliability of inference or the trustworthiness of the ensuing evidence (Spanos, 2007a).

Second, the claim that DS methods are purely data-driven and "model-free" are misleading since one always begins with certain questions of interest that presuppose "some" substantive information, however vague, in selecting the relevant data $\mathbf{Z}_0$. Related to that is the fact that DS searching relies primarily on curve-fitting "pattern" recognition in data using ML algorithms that are coded in terms of a particular "language." When such algorithms use traditional data "statistics" as in Equation (5) to search for patterns, they can easily lead one astray when the probabilistic assumptions (IID) implicit in Equation (5) are invalid, as in Equation (46). Also, the search for patterns associated with functional forms relating to the curve-fitting as indicated clearly by Equations (42)–(48), is very different from the *recurring chance regularity patterns* that can be "accounted for" by probabilistic assumptions. Indeed, as argued in Spanos (2019), the latter can provide information about the former, but not the other way around. Attempts to go beyond IID data in DS are hampered by the plethora of different forms of dependence and heterogeneity, and thus the result is invariably a hit or miss strategy guided by goodness-of-fit/prediction; see Sugiyama and Kawanabe (2012). In contrast, the search for a statistically adequate $\mathcal{M}_{\theta}(\mathbf{z})$ in model-based statistics is guided by statistical adequacy using M-S testing.

Third, using measures of goodness-of-prediction, such as $\sum_{i=n+1}^{N}(y_i - \widehat{g}(\mathbf{x}_i;\boldsymbol{\psi}_n))^2$, for data beyond the training set $\mathbf{z}_i := (\mathbf{x}_i, y_i)$ $i = n+1, n+2, \ldots, N$, is analogous to goodness-of-fit evaluations, that have little value if the selected model $y_i = \widehat{g}(\mathbf{x}_i;\boldsymbol{\psi}_n)$ is statistically misspecified. At the very least, DS practitioners should perform more systematic, but basic M-S testing, to evaluate whether the residuals $\widehat{\varepsilon}_i = y_i - \widehat{g}(\mathbf{x}_i;\boldsymbol{\psi}_n), i = 1, 2, \ldots, n$, are *non-systematic* in a statistical sense (white-noise) or not. Basic M-S testing requires only the residuals $\widehat{\varepsilon}_i$ and fitted values $\widehat{y}_i = \widehat{g}(\mathbf{x}_i;\boldsymbol{\psi}_n), i = 1, 2, \ldots, n$, both of which are readily available in DS; see Spanos (2019).

Fourth, it is well known that more data can give rise to more reliable and precise inferences, but that comes with a price: the probabilistic assumptions imposed on one's data are approximately valid for the particular data. The literature on Big Data and DS gives the impression that the last caveat about model validation can be dispensed with when one (i) uses very large sample size $n$ data and (ii) views asymptotic results as approximately valid since $n$ is very large. Both claims (i) and (ii) are false because the precision based on a large $n$ is illusory when the underlying probabilistic assumptions are invalid for data $\mathbf{Z}_0$. This is particularly pernicious for departures from homogeneity assumptions, such as ID and stationarity, which will render both the reliability and precision worse as $n$ increases; see Spanos and McGuirk (2001). Also, no limit theorem "as $n \to \infty$" holds for any $n < \infty$; see the quotation from Le Cam (1986) in Section 7. Almost all techniques employed in DS rely on mathematical approximation bounds, which can be too crude in practice, and their precision depends on increasing the degree of the polynomial approximation $m$. Indeed, there is a conflict between $n$ and $m$ and the approximation errors are usually systematic from a statistical perspective (not white-noise); see Spanos (2010b).

Fifth, it can be shown that the choice of an inner product (loss function) is directly related to the underlying distribution assumption which defines the relevant distance (norm) via the log-likelihood function.

**Example 9.** The normal distribution for $\varepsilon_k = y_k - g_m(\mathbf{x}_k;\boldsymbol{\psi}_n)$, via the log-likelihood

$$\ln L(\sigma^2;\mathbf{z}) = \text{const.} - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{k=1}^{n}\varepsilon_k^2. \tag{49}$$

has an inherent connection to the 2-norm $\|\,.\,\|_2$ that induces the *inner product*

$$\langle y_k - g_m(x_k;\boldsymbol{\alpha})\rangle_2 = \sum_{k=1}^{n}(y_k - g_m(\mathbf{x}_k;\boldsymbol{\psi}_n))^2. \tag{50}$$

Also, it is no accident that a natural base set for $L_2(-\infty,\infty)$ is provided by the Hermite polynomials whose weight function, $w(x) = e^{-x^2}$, for orthogonality is a rescaled standard normal density $\phi(x)$; see Timan (1994).

In contrast, the 1-norm $\|\,.\,\|_1$ induces the *inner product* (Powell, 1981)

$$\langle y_k - g_m(x_k;\boldsymbol{\alpha})\rangle_1 = \sum_{k=1}^{n}|y_k - g_m(\mathbf{x}_k;\boldsymbol{\psi}_n)|, \tag{51}$$

which is related to the *Laplace* distribution, $f(\varepsilon_k) = (\frac{1}{2\sigma})e^{(-|\varepsilon_k|/\sigma)}, \varepsilon_k \in \mathbb{R}$; see Bloomfield and Steiger (1980).

Similarly, the $\infty$-norm $\|\,.\,\|_\infty$, induces the *inner product*

$$\langle y_k - g_m(x_k;\boldsymbol{\alpha})\rangle_\infty = \sup_{\varepsilon_k\in[-\sigma,\sigma]}|y_k - g_m(\mathbf{x}_k;\boldsymbol{\psi}_n)|, \tag{52}$$

which relates to the *uniform* distribution, $f(\varepsilon_k) = \frac{1}{2\sigma}, \varepsilon_k \in [-\sigma,\sigma]$. Note also 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 Timan (1994).

This connection between the norm and the associated distribution suggests that the choice of a loss function should not be based on convenience or mathematical expediency, since the chance regularities in the data could be used to select the appropriate distribution (and likelihood function) on statistical adequacy grounds; see Spanos (2010b).

Curiously, DS tends to view model-based statistics relying on the likelihood function $L(\boldsymbol{\theta};\mathbf{Z}_0)$, as a special case of their empirical risk minimization just because $-\ln L(\boldsymbol{\theta};\mathbf{Z}_0), \boldsymbol{\theta}\in\Theta$, can be viewed as another loss function; see Cherkassky and Mulier (2007), p. 31. This claim misses the crucial point that $\ln L(\boldsymbol{\theta};\mathbf{Z}_0)$ is based on testable probabilistic assumptions comprising $\mathcal{M}_\theta(\mathbf{z})$, as opposed to arbitrary loss functions based on information other than data $\mathbf{Z}_0$; see Spanos (2017).

Last, in DS, the estimated parameters of $g_m(\mathbf{x}_i;\boldsymbol{\psi}_n)$ in $\mathcal{G}$ do not usually have any statistical or substantive interpretation because the focus is primarily on goodness-of-prediction. Indeed, the regularization term, $\lambda C(g(\mathbf{x}_i;\boldsymbol{\psi}_n))$, imposes arbitrary restrictions on these parameters to fine-tune the prediction error, obviates any possibility for any interpretation of these parameters. Hence, as

argued by Pearl and Mackenzie (2018), p. 351, the fitted models in DS are of little to no value for explanation purposes.

In light of the above-mentioned DS weaknesses, how does the literature justify their modeling and inference procedures? As argued by Cherkassky and Mulier (2007):

"The conceptual approach used by SLT is different from classical statistics in that SLT adopts a goal of system imitation rather than system identification." (p. 100).

This is a very weak argument because it ignores the fact that prediction could be unreliable, and it begs the question what is a "good model" on "system imitation" grounds? Goodness-of-fit/prediction measures are reliable when their invoked probabilistic assumptions are valid for the particular data, rendering prediction errors "non-systematic" (e.g., white-noise); no just small relative to a particular loss function for a particular prediction period; see Spanos (2007a). For instance, Akaike-type model selection procedures, whose goodness-of-fit/prediction measure is $-2\ln L(\widehat{\boldsymbol{\theta}};\mathbf{Z}_0)$, which is erroneous when the underlying statistical model is misspecified; see Spanos (2010b). Statistical adequacy is needed to secure the reliability of any form of statistical inference, including Pearson-type descriptive statistics as well as "prediction", irrespective of whether the stated objective of statistical modeling is "system imitation" or "system identification."

On a positive note, DS techniques can be very useful in cases where:

(i) the data $\mathbf{Z}_0$ constitute a realization of a random (IID) sample,
(ii) the practitioner is dealing with a large number of variables with meager substantive information to guide the modeling, and
(iii) a short horizon prediction is the sole objective.

## 8 | SUMMARY AND CONCLUSIONS

### 8.1 | A summary of the main argument

In light of the enhanced model-based frequentist inference, one can make a case that as a result of the features (i)–(iv), in Section 1, shared by Pearson's descriptive statistics, nonparametric statistics, DS, and GC modeling, have three crucial weaknesses

(a) The (implicit) statistical model $\mathcal{M}(\mathbf{z})$ comprising the probabilistic assumptions imposed on the observable process $\{\mathbf{Z}_t,\ t\in\mathbb{N}\}$ underlying data $\mathbf{Z}_0$ is often nebulous, veiled, and often overparametrized. This renders the task to secure the statistical adequacy of $\mathcal{M}(\mathbf{z})$ impractical.
(b) The indirect distribution assumptions imposed on the data weaken the effectiveness of the ensuing inferences by foregoing finite-sample inference for much less accurate asymptotic approximations; see Le Cam (1986) quotation in Section 6.
(c) Excellent goodness-of-fit/prediction is neither necessary nor sufficient for the statistical adequacy of a nonparametric $\mathcal{M}_{\mathcal{F}}(\mathbf{z})$ model or the DS residuals $\widehat{\varepsilon}_k = y_k - g_m(\mathbf{x}_k;\widehat{\boldsymbol{\psi}}_n)$; see Spanos (2007a)

In light of (a)–(c), the shared curve-fitting perspective, based on features (i)–(iv) in Section 1, calls into question the data-driven (DS) versus theory-driven (GC) modeling as a false dilemma since the difference in fitting a frequentist curve in $\mathcal{F}_P$, or a nonparametric family $\mathcal{F}$ of smooth density functions or a polynomial of degree $p$ or a structural GC model ($\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$) is much less

important when the trustworthiness of empirical evidence is a primary objective. In all four competing approaches, one is foisting a family of mathematical curves onto $\mathbf{Z}_0$, and uses goodness-of-fit/prediction to evaluate their appropriateness.

## 8.2 | Conclusions

The discussion has sketched four paradigm shifts in statistical modeling and inference since the 1920s to bring out and shed light on their similarities and differences as they influence their effectiveness (optimality, reliability) in "l earning from data about phenomena of interest." The main argument is that to learn from data one needs trustworthy evidence stemming from statistically adequate inductive premises; excellent goodness-of-fit/prediction is no substitute. A strong case can be made that the key weakness of the curve-fitting perspective is that it does not provide: (i) testable probabilistic assumptions that ensure statistical adequacy, or (ii) ascertainable error probabilities to assess the effectiveness of the inference procedures. Instead, (i)* they rely on goodness-of-fit/prediction measures, and (ii)* asymptotic inferential results whose finite sample (any $n < \infty$) statistical reliability is unknown. That is, parametric statistical models $[\mathcal{M}_\theta(\mathbf{z})]$ based on strong probabilistic assumptions that are validated vis-a-vis the data $\mathbf{z}_0$, provide the best way to learn from data because they secure the effectiveness (reliability and precision) of inference and the trustworthiness of the ensuring evidence.

The enhanced F–N–P model-based approach, which includes the statistical versus substantive and modeling versus inference demarcations (Spanos, 2006), can accommodate the recent developments in GC modeling and inference in a way that (a) brings out the substantive assumptions comprising the causal structure in $\mathcal{M}_\varphi(\mathbf{z})$, (b) allows for securing the statistical adequacy of $\mathcal{M}_\theta(\mathbf{z})$, and (c) evaluating the cogency of the causal structure. Also, when integrated properly, the computational dimension of PAC learning procedures can enhance the modeling facet of the model-based approach, which in turn can bring out explicitly the probabilistic (statistical) aspects of PAC learning and render them testable vis-a-vis data $\mathbf{Z}_0$. These relate to both the presence and the learnability of regularities, and their appropriateness beyond the IID assumptions.

### ACKNOWLEDGEMENTS

Thanks are due to two anonymous reviewers whose constructive criticisms and many insightful suggestions improved the original version of the paper greatly.

### ORCID

*Aris Spanos* https://orcid.org/0000-0002-9229-424X

### REFERENCES

Aggarwal, C. C. (2020). *Linear algebra and optimization for machine learning*. Springer.

Bahadur, R. R., & Savage, L. J. (1956). The nonexistence of certain statistical procedures in nonparametric problems. *The Annals of Mathematical Statistics*, *27*, 1115–1122.

Billingsley, P. (1995). *Probability and measure* (4th ed.). Wiley.

Blalock, H. M. (1964). *Causal inferences in nonexperimental research*. UNC Press.

Bloomfield, P., & Steiger, W. (1980). Least absolute deviations curve-fitting. *SIAM Journal on Scientific Computing*, *1*(2), 290–301.

Cherkassky, V., & Mulier, F. M. (2007). *Learning from data: Concepts, theory, and methods*. Wiley.

Clarke, B., Fokoue, E., & Zhang, H. F. (2009). *Principles and theory for data mining and machine learning*. Springer.

Deutsch, F. (2001). *Best approximation in inner product spaces*. Springer.
Dickhaus, T. (2018). *Theory of nonparametric tests*. Springer.
Doob, J. L. (1953). *Stochastic processes*. Wiley.
Duncan, O. D. (1975). *Introduction to structural equation models*. Academic Press.
Edgeworth, F. Y. (1884). A priori probabilities. *Philosophical Magazine*, *18*, 204–210.
Efron, B., & Hastie, T. (2016). *Computer age statistical inference: Algorithms, evidence and data science*. Cambridge University Press.
Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, *52*(2), 399–433.
Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A*, *222*, 309–368.
Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver and Boyd.
Fisher, R. A. (1935). *The design of experiments*. Oliver and Boyd.
Glymour, C., Scheines, R., Spirtes, P., & Kelly, K. (1988). *Discovering causal structure: Artificial intelligence, philosophy of science, and statistical modeling*. Academic Press.
Gossett, W. (1908). The probable error of the mean. *Biometrika*, *6*, 1–25.
Greene, W. H. (2018). *Econometric analysis* (8th ed.). Prentice Hall.
Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.
Haavelmo,. (1943). The statistical implications of a system of simultaneous equations. *Econometrica*, *11*, 1–12.
Hendry, D. F. (1976). The structure of simultaneous equations estimators. *Journal of Econometrics*, *4*, 51–88.
Hood, W. C., & Koopmans, T. C. (eds.) (1953). *Studies in econometric method*, Cowles Commission Monograph, No. 14, Wiley.
Iske, A. (2018). *Approximation theory and algorithms for data analysis*. Springer.
Karl-Georg, S. (2006). *The history of approximation theory from Euler to Bernstein*. Birkhauser.
Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. MIT Press.
Kolmogorov, A. N. (1933). *Foundations of the theory of probability* (2nd English ed.). Chelsea Publishing Co.
Kreyszig, Erwin. (1978). *Introductory Functional Analysis with Applications*. NY: Wiley.
Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). University of Chicago Press.
Lahiri, S. N. (2003). *Resampling methods for dependent data*. Springer.
Lai, T. L., & Xing, H. (2008). *Statistical models and methods for financial markets*. Springer.
Le Cam, L. (1986). *Asymptotic methods in statistical decision theory*. Springer.
Lehmann, E. L. (1975). *Nonparametrics: Statistical methods based on ranks*. Holden-Day.
Lehmann, E. L. (1990). Model specification: The views of Fisher and Neyman, and later developments. *Statistical Science*, *5*, 160–168.
Luenberger, D. G. (1969). *Optimization by Vector space methods*. Wiley.
Marshak, J. (1953). In W. C. Hood, & T. C. Koopmans (Eds.), *Economic measurement for policy and prediction* (pp. 1–26).
Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. The University of Chicago Press.
Mills, F. C. (1924). *Statistical methods*. Henry Holt and Co.
Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference*. Cambridge University Press.
Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.
Neyman, J. S. (1923). On the application of probability theory to agricultural experiments. Essay on principles. *Annals of Agricultural Sciences*, *10*, 1–51.
Neyman, J. (1935). Statistical problems in agricultural experimentation. *Supplement to the Journal of the Royal Statistical Society*, *2*(2), 107–180.
Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Statistical Society of London, A*, *236*, 333–380.
Neyman, J. (1939/ 1952). *Lectures and conferences on mathematical statistics and probability* (2nd ed.). U.S. Department of Agriculture.
Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society, A*, *231*, 289–337.
Ord, J. K. (1972). *Families of frequency distributions*. Griffin.

Owen, P. D. (2017). Evaluating ingenious instruments for fundamental determinants of long-run economic growth and development. *Econometrics*, *5*(3), 1–38.
Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. Kaufman.
Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.
Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.
Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, *50*, 157–175.
Pearson, K. (1911). *The grammar of science*. (3rd ed.). Adam & Charles Black.
Pearson, K. (1920). The fundamental problem of practical statistics. *Biometrika*, *XIII*, 1–16.
Phillips, P. C. B. (1983). Exact small sample theory in the simultaneous equations model. In Z. Griliches, J. J. Intriligator, M. D. Intriligator, R. F. Engle, E. E. Leamer, & D. McFadden (Eds.), *Handbook of econometrics* (Vol. 1, pp. 449–516). Elsevier.
Powell, M. J. D. (1981). *Approximation theory and methods*. Cambridge University Press.
Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688–701.
Schölkopf, B., Luo, Z., & Vovk, V. (2013). *Empirical inference: Festschrift in honor of Vladimir N. Vapnik*. Springer.
Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.
Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Chapman & Hall.
Silvey, S. D. (1975). *Statistical inference*. Chapman & Hall.
Simon, H. A. (1953). In W. C. Hood, & T. C. Koopmans (Eds.), *Causal ordering and identifiability* (pp. 49–74).
Spanos, A. (1986). *Statistical foundations of econometric modelling*. Cambridge University Press.
Spanos, A. (1990). The simultaneous equations model revisited: Statistical adequacy and identification. *Journal of Econometrics*, *44*, 87–108.
Spanos, A. (2001). Parametric versus non-parametric inference: statistical models and simplicity. In A. Zellner, H. A. Keuzenkamp, & M. McAleer (Eds.), *Simplicity, inference and modelling* (pp. 181–206). Cambridge University Press.
Spanos, A. (2005). Structural equation modeling, causal inference and statistical adequacy. In P. Hajek, L. Valdes-Villanueva, & D. Westerstahl (Eds.), *Logic, methodology and philosophy of science* (pp. 639–661). King's College.
Spanos, A. (2006). Where do statistical models come from? Revisiting the problem of specification. In J. Rojo (Ed.), *Optimality: The second Erich L. Lehmann Symposium*, Lecture Notes-Monograph Series (Vol. 49, pp. 98–119). Institute of Mathematical Statistics.
Spanos, A. (2007a). Curve-fitting, the reliability of inductive inference and the error-statistical approach. *Philosophy of Science*, *74*, 1046–1066.
Spanos, A. (2007b). The instrumental variables method revisited: On the nature and choice of optimal instruments. In G. D. A. Phillips, & E. Tzavalis (Eds.), *Refinement of econometric estimation and test procedures* (pp. 34–59). Cambridge University Press.
Spanos, A. (2010a). Graphical causal modeling and error statistics. In D. G. Mayo, & A. Spanos (Eds.), *Error and inference: Recent exchanges on experimental reasoning, reliability and the objectivity and rationality of science* (pp. 364–375). Cambridge University Press.
Spanos, A. (2010b). Akaike-type criteria and the reliability of inference: Model selection vs. Statistical model specification. *Journal of Econometrics*, *158*, 204–220.
Spanos, A. (2010c). Statistical adequacy and the trustworthiness of empirical evidence: Statistical vs. substantive information. *Economic Modelling*, *27*, 1436–1452.
Spanos, A. (2011). Revisiting the welch uniform model: A case for conditional inference? *Advances and Applications in Statistical Science*, *5*, 33–52.
Spanos, A. (2013). A frequentist interpretation of probability for model-based inductive inference. *Synthese*, *190*, 1555–1585.
Spanos, A. (2017b). Why the Decision-Theoretic Perspective Misrepresents Frequentist Inference, chapter 1, pp. 3–28, *Advances in Statistical Methodologies and Their Applications to Real Problems*, (editor) T. Hokimoto, InTechOpen.
Spanos, A. (2018). Mis-specification testing in retrospect. *Journal of Economic Surveys*, *32*(2), 541–577.

Spanos, A. (2019). *Introduction to probability theory and statistical inference: Empirical modeling with observational data* (2nd ed.). Cambridge University Press.
Spanos, A. (2020). Yule-simpson's paradox: The probabilistic vs. the empirical conundrum. *Statistical Methods & Applications*, *30*, 605–635. https://doi.org/10.1007/s10260-020-00536-4.
Spanos, A., & McGuirk, A. (2001). The model specification problem from a probabilistic reduction perspective. *Journal of the American Agricultural Association*, *83*, 1168–1176.
Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search*. MIT Press.
Strotz, R. H., & Wold, H. O. (1960). Recursive vs. nonrecursive systems: An attempt at synthesis. *Econometrica*, *28*, 417–427.
Sugiyama, M., & Kawanabe, M. (2012). *Machine learning in non-stationary environments*. MIT Press Cambridge.
Thompson, J. R., & Tapia, R. A. (1990). *Nonparametric function estimation, modeling, and simulation*. SIAM.
Timan, A. F. (1994). Theory of approximation of functions of a real variable, Dover Publications.
Valiant, L. (1984). A theory of the learnable. *Communications of the ACM*, *27*(11), 1134–1142.
Valiant, L. (2013). *Probably approximately correct*. Basic Books.
Vapnik, V. N. (2000). *The nature of statistical learning theory*. Springer.
Vapnik, V. N. (2006). *Estimation of dependences based on empirical data* (2nd ed.). Springer.
Vapnik, V. N., & Chervonenkis, A. (1991). The necessary and sufficient conditions for consistency of the method of empirical risk minimization. *Pattern Recognition and Image Analysis*, *1*(3), 284–305.
Wasserman, L. (2006). *All of nonparametric statistics*. Springer.
Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, *20*, 557–580.
Wright, S. (1923). The theory of path coefficients: A reply to Niles's criticism. *Genetics*, *8*(3), 239–255.
Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics*, *5*(3), 161–215.
Yule, G.U. (1916). *An introduction to the theory of statistics* (3rd ed.). Griffin.

**How to cite this article:** Spanos A. Statistical modeling and inference in the era of Data Science and Graphical Causal modeling. *Journal of Economic Surveys*. 2021;1–37. https://doi.org/10.1111/joes.12483