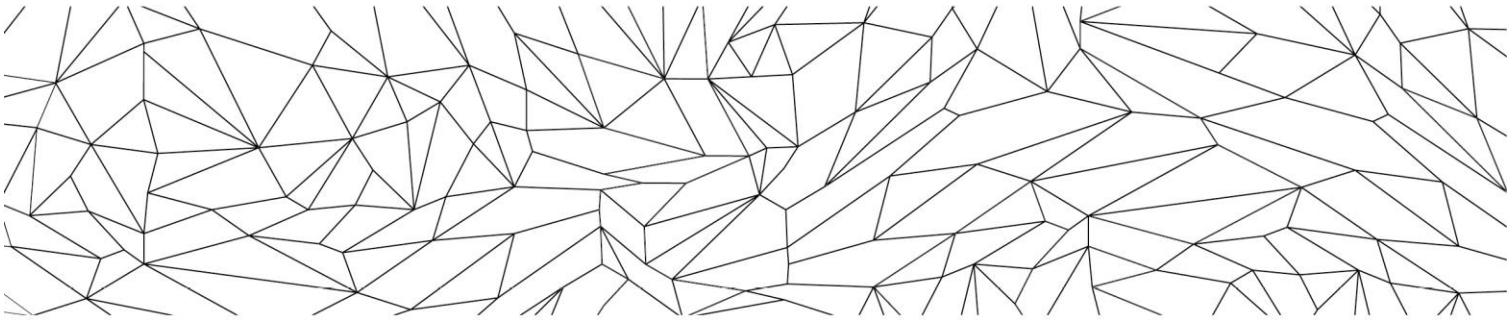
SOAS Department of Economics



Redirect the Probability Approach in Econometrics Towards PAC Learning

Duo Qin

Working paper

No. 249

March 2022

SOAS University of London

The SOAS Department of Economics Working Paper Series is published electronically by SOAS University of London.

ISSN 1753 – 5816

Suggested citation
Qin, Duo, (2022), “Redirect the Probability Approach in Econometrics towards PAC Learning”, SOAS Department of Economics Working Paper No. 249, London: SOAS University of London.

Department of Economics
SOAS University of London
Thornhaugh Street, Russell Square, London WC1H 0XG, UK
Phone: + 44 (0)20 7898 4730
Fax: 020 7898 4759
E-mail: economics@soas.ac.uk
http://www.soas.ac.uk/economics/

Redirect the Probability Approach in Econometrics Towards PAC Learning

Duo Qin[1]

Abstract

Infiltration of machine learning (ML) methods into econometrics has remained relatively slow, compared with their extensive applications in many other disciplines. The bottleneck is traced to two key factors – a communal nescience of the theoretical foundation of ML and an outdated probability foundation. The present study ventures on an overhaul of the probability approach by Haavelmo (1944) in light of ML theories of learnibility, centred upon the notion of *probably approximately correct* (PAC) learning. The study argues for a reorientation of the probability approach towards assisting decision making for model learning and selection purposes. The first part of the study is presented here.

Keywords: probability; uncertainty; machine learning; hypothesis testing; knowledge representation.

JEL classification: C10, C18, B40.

Acknowledgements: I am greatly indebted to Xi-Yu Jiao, Ruben Lee, Xue-Lin Liu, Shan Lu, Sophie van Huellen, Qing-Chao Wang, Chris Watkins, and also posthumously Olav Bjerkholt for invaluable discussion, encouragement, comments and suggestions I have received.

[1] Department of Economics, SOAS University of London. Russell Square, London WC1H 0XG, UK. Email: dq1@soas.ac.uk

‘If the inference/algorithm race is a tortoise-and-hare affair, then modern electronic computation has bred a bionic hare.’

Efron and Hastie (2016, p.4)

Machine learning (ML) is ubiquitous nowadays. Contrasts are striking when we try to compare econometrics with ML. Mathematical contents are generally simpler in ML than in econometrics textbooks; but achievements in ML already significantly outshine those in econometrics, although econometrics precedes ML as disciplines of analysing passively observed data from an open world. There must be an issue of inefficiency in econometrics, if judged by economists’ standards.

Applications and adaptions of ML methods in econometrics are definitely on the rise in recent years, so are reviews and promotions of ML toolkits, e.g. see reviews by Varian (2014), Mullainathan and Spiess (2017), Charpentier *et al* (2018), Athey and Imbens (2019) and Iskhakov *et al* (2020). Noticeably, ML is widely communicated and conceived as purely a toolbox for data analysis in the econometric circle. There is a communal blind spot of the theoretical foundation of ML. Two Chinese idioms comes to mind when facing the situation: ‘looking at the sky from the bottom of a well’, and ‘seeing trees but not the forest’.

It is obviously impossible for any discipline to gain widespread success without a sound foundation. What underlies ML algorithms are theories of learnability, centred upon the notion of probably approximately correct (PAC) learning. PAC learning lays a sound base for systematic search of empirically best possible models, and the search is done predominantly following a distribution-free strategy. Moreover, development of ML is accompanied by that of artificial intelligence (AI), and intimately related to logical representation of knowledge, reasoning and decision-making under uncertainty. The perspective and insight of PAC learning makes me realise that the root cause of inefficiency in econometrics lies at its outdated probability approach, and venture on a deconstruction of the approach.

The venture brings me to T. Haavelmo’s 1944 monograph. His eloquent rhetoric for the probability approach and subsequently the rigorous formalisation work presented in the Cowles Commission Monograph 10, see Koopmans (1950), have firmly framed econometrics in the classical statistics paradigm and facilitated its extensive development in the discipline. However, various modelling issues have emerged and accrued from practice, implicating shortcoming of the paradigm. Remarkably, the shortcoming was pinned down in Judge *et al* (1980) to the narrow focus of the paradigm – inference for ‘the case of a *known* sampling model’. They further stated, ‘The framework for a theory of statistical inference based on false models remains to be developed. This problem must be squarely faced in the future if we are to meet and cope with the problem of developing more efficient procedures for learning from finite samples of passively generated economic data’ (p.778). Now, the framework is within sight of ML theories.

An overhaul of the probability approach inevitably touches on various issues of historical and methodological nature. Nevertheless, my ultimate aim is to help free applied economists from the bondage of textbook econometrics so as to unleash their potentials in producing better synthesis of theoretical and empirical knowledge. ‘When academics encounter a new idea that doesn’t conform to their preconceptions, there’s often a sequence of three reactions: first dismiss, then reject, and finally declare it obvious.’ (Sloman and Fernbach, 2017, p.255) Hopefully, this study can speed up the arrival of the third reaction in the econometric circle. To accomplish the present formidable task, I have resorted to writing interchangeably in both English and Chinese. Here, the draft of first three chapters in both languages are presented.

若将推断与算法之争比做龟兔赛跑，那么当代的电脑算法已经育出仿生赛兔了。

Efron and Hastie (2016, p.4)

经济计量学与机器学习都是以分析从开放世界被动观测到的数据为主的学科。在机器学习不断渗透各个学科的当今，对比两个学科就不难发现，机器学习教科书的数学内容要比计量学教科书浅显容易得多，但在实际应用中却明显是后来居上。虽然计量学的形成要先于机器学习，但至今的应用业绩的确相形见绌。依经济学家的评判尺度，计量学一定存在效率低下问题。

近年来，计量应用研究中引入机器学习技术的案例与日俱增，英文文献中介绍和推荐机器学习的综述可参见 Varian (2014), Mullainathan and Spiess (2017)、 Charpentier *et al* (2018)、 Athey and Imbens (2019) 及 Iskhakov *et al* (2020)。遗憾的是，计量学界对机器学习理论基础的认识存在盲点，普遍把机器学习表述为数据分析的工具箱。坐井观天、只见树木不见森林，这两个成语恐怕是最能贴切形容这种情形的了。

无需赘言，任何学科的成功都离不开坚实的理论基础后盾。机器学习算法的后盾是其严谨的、以概率近似正确（probably approximately correct (PAC)）学习概念为核心的可学性理论。PAC 学习为如何设计算法系统学习到尽量最优经验模型奠基，并为学习过程制定了以无分布假定为主的策略。机器学习还与人工智能的发展互助互长，相得益彰。人工智能对于实现对人脑知识在充满不确定性的现实中的逻辑推理表述和决策模拟，给机器学习提供了清晰的推理关系表述规范。PAC 学习的视角和分析思路启发了我，计量学科效率欠佳的根源其实在于它已过时的概率论基础。这便推动了我的系统反思、更新和重组计量学概率方法论之旅。

本项研究的直接对象是哈维尔莫 1944 年发表的专著。专著中哈维尔莫对概率方法论的雄辩和后继美国 Cowles Commission 专著系列第 10 集对计量学的严格正规表述（见 Koopmans (1950)）不但把经济计量学稳稳定位在经典统计学框架中，而且为计量学大量拓展统计数学工具奠定基础。然而，计量学理论在应用建模方面上却不断遇到各种难题困扰。其实，Judge *et al* 早在其 1980 教科书中就把这些难题归为过窄的统计推断对象— “*已知的*抽样模型”。他们还指出：“针对有误模型做统计推断的理论框架尚待开发出来。我们若要为学习被动观测的经济数据有限样本研制出更为有效的分析步骤手段，就必须在将来的研究中正视并且解决这一问题” (第 778 页)。如今，机器学习理论业已为解决这一问题提供了成熟的理论框架。

对于计量学概率方法论的全面检修必然会涉及许多学说史和方法论方面的问题。不过，本项研究的最终服务对象还是应用经济学家。毕竟有培根的名言，“历史使人明智”，只有彻底摆脱计量学教科书框架的束缚，接受 PAC 学习理念，应用经济学家才可能充分发挥出他们综合利用理论与经验知识、有效分析现实经济问题的潜能。Sloman and Fernbach (2017, p.255) 的书中写道：“当学者面临一个与其先知预想截然不同的观点时，他们的反应过程往往呈三步：首先是不予考虑，然后是抵制拒绝，最后称该观点不过是显而易见的。” 但愿我的探索之旅能够加速计量学界尽快踏出这第三步，顿悟出大道至简的真谛。本次任务对于我来说犹如愚公移山。为了实现目标，我采用了交替中文英文的写作方式，以尽量明确分析思路。本工作论文是用两种文字草撰的前三章，希望起到抛砖引玉的作用。

1. Abstract Modelling of Reality

1.1 Introduction

A probabilistic outlook forms the epistemological base of empirical verification of theories, and probability-based statistical inference offers the essential means of the verification. These principles underlie Haavelmo's promotion of the probability approach. The principles are upheld by roughly three inter-linked claims. First, decisions of economic theory falsification can only be made in practice by means of probability thresholds. Second, empirical measurement of economic theory entails the use of probability-based tools and notions. Third, inaccuracy is inevitable in the measurement or quantification of economic variables. Ubiquitous uncertainty thus sanctions probability models as the 'bridge' between theory and data, and classical statistics as the tool provider for econometrics. Haavelmo thus outlines econometrics in three parts: hypothesis testing, model estimation and prediction, as discussed in chapters 4, 5 and 6 respectively of his monograph.

It should be noted that, prior to Haavelmo's monograph, economists had voiced their reservation and scepticism over the feasibility of probability in economic analysis. The voice is led by two publications in 1921, *A Treatise on Probability* by J.M. Keynes, and *Risk, Uncertainty, and Profit* by F.H. Knight. The scepticism extends mainly on two accounts: one is related to the historical nature of economic events, in that information of the past is never adequate for forecasting the future; the other is related to the abstract nature of many economic concepts, in that probability measures are inept in formulating various uncertain aspects of those concepts. The ensuing debates have never ceased over the century, but they appeal mainly to the history and methodology of economics communities.[1] Little collective interest is discernible from the econometrics circle. Although Keynes challenged Tinbergen's pioneering econometric modelling work with serious methodological concerns,[2] the deterrent effect of Keynes' critique was short lived. Econometrics has by now fledged so much along Haavelmo's probability approach that it forms a fundamental sub-discipline of economics nowadays. With rapidly increasing data for economic analysis, widespread application of statistical tools is inevitable, whereas it is hard to find constructive help from the anti-probability rhetoric following the Keynes tradition.

To a certain extent, the growth of econometrics is sheltered by Haavelmo's demarcation of modelling tasks between economists and econometricians. The three-part research structure of econometrics is conditioned on the presence of *a priori* theoretical models, which are complete for statistical purposes.[3] Once the task of model formulation is delegated to economists, there seems little relevance between econometrics research and debates over the feasibility of probability in the history of economic thought circle. However, when it comes to the origin of models, Haavelmo concedes that the 'construction of tentative models … is a

[1] For more details, see Bateman (1990), Davidson (1991), McCann (1994) and also the special issue of Cambridge Journal of Economics in 2021 (issue 5, vol. 45) marking the centenary of the two books, i.e. Keynes (1921) and Knight (1921).

[2] Keynes' critique of Tinbergen's macro model is published as a review of Tinbergen's 1939 book entitled 'Statistical Testing of Business Cycle Theories: A Method and its Application to Investment Activity'; see Part VI, 'The Tinbergen Debate', in Hendry and Morgan (1995) for details.

[3] It should be noted that conditions of when models are 'complete … for statistical purposes' were elaborated by Koopmans (1950). However, his notion of 'complete' models is geared towards identification conditions of simultaneous-equation models and does not match with that of adequacy to be discussed in the final paragraph of this section.

creative process, an art' (p.10). From hindsight, this amounts to admitting that uncertainties in model construction are too complex to be formalised by the probability approach.[4] It also implies that econometrics is predicated on a condition whose realisation is beset with uncertainties beyond scientific formalisation. Indeed, accruing evidence from decades of applied econometric research shows that *a priori* formulated models are inadequate for empirical purposes and that the probability approach is inept to deal with uncertainties from model formulation. To those economists whose research is focused on designing empirically verifiable models, they find the classic Haavelmo programme 'irrelevant to the inductive process by which theory actually evolves. … his programme had nothing to contribute to the construction of economic models' (Eichenbaum, 1995, p.1619). In fact, experiments in data-assisted model modification and specification are already part of routine research activities among applied modellers. Unfortunately, lack of established and systematic ways for such 'transgressions' has given rise to ceaseless contention. It also brings about methodological reflections on the probability foundation of econometrics.[5]

According to Hempel's deductive-nomological scheme (1965), adequacy of any scientific explanations embodies both logical adequacy and empirical adequacy. Logical adequacy of theoretical models in economics is secured generally via formal algebraic deduction. The deduction renders a sense of completeness. Empirical adequacy, on the other hand, is usually in shortage, as evident from routinised data-assisted model revision activities. Essentially, the activities reflect the lack of means to achieve empirical adequacy in the process of *a priori* model formulation.[6] In other words, there are empirically uncertain factors which are beyond what *a priori* deductive formulation of probabilistic models can tackle. In order to gain a clear understanding of what these factors are, it is necessary for us to delve carefully into the complexity of uncertainties when theoretical knowledge and empirical information are met and joined via models.

1.2 Uncertainty in Quantification of Economic Variables

Uncertainty in quantitative measures of economic variables forms the backbone of Haavelmo's probability approach. In conceiving the uncertainty as random errors, he brings in a three-type classification of economic variables: observational, true and theoretical variables. Contemplation over the three categories, however, shows the conceptualisation problematic and confounding. Characteristics of the uncertainty are far more complex than what random measurement errors can depict. Moreover, quantification of the uncertainty under many circumstances is more evasive than what can be coped with by statistical inference tools. The appeal of dealing with measurement uncertainty from a probabilistic outlook does not stand up for scrutiny.

Let us start from observational variables. They are more commonly referred to as 'observable' than 'observational' variables nowadays. Many economic variables are defined directly in measurable terms, such as income, consumption, output, cost, profit and price etc.. The directly un-observables are a minority, such as risk, utility and capacity. On examining

[4] For more discussion on the behavioural nature of uncertainty, see e.g. Bradley and Drechsler (2014), where a typology of uncertainty from agents' decision making perspective is proposed.
[5] For an account of unsettling contention on model selection issues in the history of econometrics, see Qin (2013, Ch9); as for methodological reflections on econometrics in connection to its probability foundation, see Rowley and Hamouda (1987), and also Stanley (1998).
[6] See Stigum (2003, Part III) for a formal discussion on this point.

observable variables, economists usually treat them in highly general and idealised forms. The variables fall into Haavelmo's category of true variables as the degree of generality is pushed to infinity. The abstract nature of true variables determines them having no one-to-one counterparts in available data even though they are innately observable. Take the concept of income for example. We have, from the scale perspective, national income, gross national product (GNP), gross domestic product (GDP) at the macroeconomic level, and household or personal income, final income, after-tax income, disposable income at the microeconomic level. From the time perspective, we have monthly income, annual income, regular income and irregular income. When it comes to the concept of price, measurement varieties extend much further. Scale wise, there are consumer price index, fixed capital investment price index, import and export price indices, various financial market indices at the macro level; at the micro level, the producer's prices, wholesale prices, retail prices and average price indices for single commodities, and also various wage rates for labours of different occupations, not to mention numerous equity prices listed on financial markets. Time wise, price data are available at higher frequencies than income data; there are daily prices in commodity markets and tick-by-tick stock prices in financial markets. On the whole, the framework of classical statistics is discernibly too narrow to cope with the gap and measurement uncertainty between true variables and their observable counterparts.

Fortunately, econometric research is spared from a sizeable part of the gap, due largely to a combination of two factors: data sources and research purposes. As far as the data source is concerned, most economic data are classified as *secondary* by collection. Unlike *primary* data, which are collected directly by researchers in accordance with specific purposes of studies, secondary data are collected and released mostly by governments, financial market authorities and international organisations. Expectation is generally rather low on the accuracy of observable variables of the secondary source in representing true variables (usually below what random measurement errors characterise in statistics). It is widely conceded that inexact data representation of true variables is unavoidable. Nontrivial statistical discrepancies arise even with the same observable variable when data collection and processing methods differ. A well-known example is GDP by the expenditure side, which can differ sizeably from GDP by the production side. Nevertheless, it is taken for granted that quality improvement of secondary data is the responsibility of the institutions which release them. So long as methods of data collection and calculation remain unchanged, discrepancies between true variables and their chosen observable counterparts are usually not taken into account in econometric modelling research. This is mainly justified by the practical need of policy analyses. Since what confront policy makers and analysts are various data of the secondary source, they are mostly concerned with how to make data-congruent judgements in respect to the variables of their concerns. Hence, they care only about the inferability of the given observable counterparts of the variables of interest, rather than of the true variables *per se*, i.e. variables devoid of any specific scale or time frames. To a large extent, the practical value of any modelling results is predicated upon clearly targeted circumstances within which the research is undertaken.

Now, under what circumstances do measurement errors in economic variables fall into the range of uncertainty deemed as the object of econometric investigation? This brings us to the situation where the variables of interest are unobservable directly and their measurement is difficult to acquire without resorting to modelling. The variables are referred to as latent

variables in the situation. Noticeably, our concerns over latent variables reveal the ambiguous nature in labelling variables by observability. Consider again the price variable. Although regarded usually as an observable, this variable is not always directly measurable by commonly used averaging formula at certain desired aggregative levels. Hence, prices can become latent unobservables. In microeconomics, for instance, the same consumer goods are often marketed under different brands and specifications, such as automobiles and electronic products. Different prices of the same type of goods signify qualitative heterogeneity associated with different brands and/or specifications. A price index of the type entails removal of the part of individual prices reflecting the heterogeneity. The need gives rise to the modelling method of hedonic price regression.

Obviously, it is impossible to have one-to-one correspondence between a latent variable and its indices generated by particular measurement models. Modellers would go for indices whose discrepancies from the latent variable are desirably trivial as purely random errors, whereas they would not expect *a priori* that differences between indices from different models be purely trivial random errors. Since latent variables are unobservable by definition, choice of indices can only be made through assessing whether they exhibit qualities closest to what are expected of the latent variables. This assessment process is essentially one of model choice. The need for model choice, in turn, implies that non-trivial uncertainty is unavoidable as far as whether purely theory-based models are data coherent is concerned. Can this uncertainty be adequately tackled by the probability approach? Finding answers to the question brings us back to the concluding point of the previous section, i.e. we need to focus our scrutiny on where the uncertainty arise when theoretical models are applied to data.

The above discussion leads us finally to 'theoretical' variables, the last type in Haavelmo's classification. Since models act as the media of certain theoretical postulates, model-generated indices, being postulated candidates to approximate latent variables, should thus be qualified as theoretical variables. The qualification may lead to questions over the distinction between true and theoretical variables as well as its empirical usefulness, since the former is essentially an idealised concept without one-to-one correspondence in reality. While these open questions admit no ready and easy answers, it is easy to see that the issue of model choice which arises from latent variable model research should not be treated in any different ways methodologically from that which arises from other model research areas in econometrics. Consequently, our quest for the exact condition under which measurement uncertainty in economic variables fall into the range of econometric investigation pinpoints the uncertainty of theoretically formulated models when they are confronted with data.

1.3 Theoretical Models and Economic Reality

Let us start by delving into where uncertainty lies in the example given in the previous section, i.e. measuring goods prices at the aggregate level by hedonic price model (HPM). Suppose that we have large samples of cross-section data on individual products of one goods on sale for *T+1* time periods, $(t = 0,1,\cdots,T)$, with each sample consisting of prices and major features of individual products: $\left(p_i \middle| x_{ij}; i = 1,\cdots,n; j = 1,\cdots,J\right)$. To facilitate the analysis, let us examine the HPM in its adjacent period time dummy variable specification.[7]

[7] Griliches (1961) is credited as the pioneering HPM research; see Triplett (2004) and Hill (2013) for detailed reviews of the topic.

Specifically, we pool two adjacent samples together and run the following HPM on *T* pooled samples:

$$\ln\left(p_i^{t,t+1}\right) = \beta_0 + f\left(z_{ik}^{t,t+1}, \beta_k\right) + \alpha^{t+1} D_i + \varepsilon_i^{t,t+1} \quad (1.3.1)$$

where z_{ik} $(k = 1, \cdots, K)$ are quality variables derived from features x_{ij}, e.g. $z_{ik} = x_{i1}$, or $z_{ik} = ln(x_{i1})$, or $z_{ik} = x_{i1}x_{i2}$, or $z_{ik} = x_{i1}^2$, and D_i is a time dummy with $D_i = 0$ at t and $D_i = 1$ at $t + 1$. The key parameter of interest is α^{t+1}, since a time-series hedonic price index of the goods can be easily constructed from the series, $\left\{e^{\hat{\alpha}^{t+1}}\right\}$, of the *T* regressions. Discernibly, $\hat{\alpha}^{t+1}$ is sensitive to the exact formulation of $f\left(z_{ik}^{t,t+1}, \beta_k\right)$. In practice, two symptoms are known to frequent (1.3.1): undesirable statistical properties in residual distribution and weak constancy in $\hat{\beta}_k$ across different pooled samples. Both indicate empirical inadequacy of *a priori* formulated $f\left(z_{ik}^{t,t+1}, \beta_k\right)$ when it is tried to data. Two diagnoses are often considered: First, hedonic demand with respect to certain quality features is nonlinear, e.g. what consumers are willing to pay for certain feature enhancement is not simply a linear incremental in price; and second, there is too much heterogeneity in samples to be representable by a single regression function based HPM. Treatments of the heterogeneity involve the design and selection of $z_{ik}^{t,t+1}$, and possibly either sample trimming or extension of $f\left(z_{ik}^{t,t+1}, \beta_k\right)$ into a generalised additive model. The resulting model revisions can only be based on data-assisted experiments and will inevitably affect $\hat{\alpha}^{t+1}$. During the experiments, however, the roles that probability-based statistical tools can play are limited. On the other hand, credible probability-based statistical inference on $\hat{\alpha}^{t+1}$ is predicated on the revisions being empirically successful.

Admittedly, HPM is far from a representative case, and measurement modelling of latent variables is not a topic that appeals to many modellers. However, uncertainty about empirical adequacy of *a priori* formulated models is ubiquitous, and the need for data-assisted model revisions as described above prevails in econometric practice. Let us reflect on this need in cases more representative than HPM, using, in turn, two basic types of data: cross-section and time-series samples.[8] The former is dominantly used in micro-econometric research, where the research is usually guided *a priori* by theories of partially behavioural causal postulates, say $x_1 \to y$. Denote a corresponding sample of size n by $\left(y_i \middle| x_{ij}; i = 1, \cdots, n; j = 1, \cdots, J\right)$, where $(x_{i2}, \cdots, x_{iJ})$ are other possibly relevant causes. Models similar to (1.3.1) are postulated, e.g.:

$$y_i = \beta_0 + f(x_{i1}, z_{ik}; \beta_k) + \varepsilon_i \quad (1.3.2)$$

where z_{ik} $(k = 1, \cdots, K)$ are designed attributes based on $\{x_{ij}\}$ in a similar way as in the HPM case, despite of the fact that the parameter of interest is now inside $f(x_{i1}, z_{ik}; \beta_k)$. Undesirable statistical properties of ε_i are virtually a norm, as the residuals usually take the lion's share of data variation on y_i due to rather low model fit. Heterogeneity in agents' behaviour is a common diagnosis, and treated frequently with various dummy variable methods, e.g. fixed or random effect methods. The treatment effect is, however, rather limited as far as model fit is concerned. In the circumstances, research focus is mostly limited to the

[8] Although use of panel data samples is on the rise, the corresponding statistical tools are extended from either a time-series perspective or a cross-section perspective.

credibility of the estimated parameters of interest defined *a priori* in theoretical models. This leads to routine practice of data-assisted model revisions on the selection of z_{ik}, especially those which pose the risk of omitted variable bias to the consistent estimation of the parameters. Consequently, it is taken for granted that the global generalizability of those parameters is already guaranteed by *a priori* theorisation. Little attention is thus spared on assessing parameter constancy empirically.

Parameter constancy is a major concern in macro modelling using time-series data, due to the common expectation of model prediction. The drive for forecasting accuracy has led to a dynamic extension of models traditionally deduced from a static, general equilibrium stand, such as rational expectations models. Correspondingly, models of the vector autoregression (VAR) type prevail in macro-econometric research, e.g. an open VAR:

$$Y_t = A_0 + \sum_{i=1}^{l} A_i Y_{t-i} + \sum_{j=0}^{l} B_j X_{t-j} + \boldsymbol{\varepsilon}_t \tag{1.3.3}$$

where Y_t is a vector of modelled variables, X_t a vector of other relevant variables, A_i and B_j are parameter matrices.[9] VAR models usually fit data much better than what micro models yield using cross-section samples, and they thus result in more satisfactory statistical properties of $\boldsymbol{\varepsilon}_t$. Two factors contribute mainly to the contrasting outcome: time-series samples are relatively small and there exhibits significant inertia in modelled variables in the samples. Nevertheless, data-assisted model selection is indispensable, at least in two related aspects: lag length, l, and composition of X_t. VAR models are notorious for over-parameterisation, because of not merely small samples, but the sensitivity to the chosen l of the static state equilibrium relations embedded in VARs. Parameters in these relations, commonly referred to as long-run parameters and treated widely with focal theoretical interest, are derivative of the lag structure of the chosen VARs. As for the second aspect, institutional changes are of common occurrence over certain time periods. Needless to say, institutional and other sample specific features are disregarded in standard theoretical models, but they are not ignorable when it comes to forecasting. More often than not, unexpected regime shocks are identified to cause systematic forecasting failures. The failures imply loss of empirical constancy in certain parameters of the models concerned. It is a well-known fact that long-run parameters are particularly susceptible to 'regime shocks'.

Evidently, uncertainty of empirical adequacy in *a priori* formulated models is an unignorable reality. Data-assisted model revisions are not only inevitable but also prerequisite to empirical verification of economic theories by means of statistical inference. The revisions commonly extend into two intertwined dimensions: (a) Revisions with respect to possibly omitted input variables in *a priori* formulated models. Theoretically deduced models, no matter how algebraically rich for inference, are bound to omit certain causes, especially vernacular features of the economy under investigation, which are unignorable in passively observed data. The uncertainty of whether, and if yes, how these causes affect the theoretical relations of interest has to be empirically examined. (b) Revisions with respect to input variable forms. For many hypothetical causal relations, their appropriate formulation entails adequate representation of the feature of scale dependency of the variables involved. For instance, flow variables, such as income and consumption, and stock variables, such as

[9] VAR models are systematically promoted by Sargent and Sims (1977), e.g. see Qin (2013, Ch 3) for a more detailed historical account.

savings and inventory, usually lack the statistically separable property desired of regressors. The lack is a major contributor to model-form uncertainty, leaving the hypothetical causal relations prone to omitted variable bias. Consequently, careful input feature design is a prerequisite, as shown in experiments with lag lengths, short-run versus long-run specifications in the dynamic model research. In addition to the revisions, modellers should be aware of the uncertainty pertaining to inference boundaries. Unlike experimentally collected samples, 'populations' corresponding to passively observed samples in an open world are not known for certain. Practical use of econometric models is thus predicated on *a posteriori* clarification of this uncertainty, e.g. via identification of specific real world situations under which models retain empirical regularity and thus generalizability. The apparatus of statistical inference is not designed to tackle all these uncertain aspects, since its framework is built on the premise of absence of these aspects.[10] On reflection, Haavelmo's concession of the situation as an 'art' is indeed a precise sum-up.

Despite of the sum-up, Haavelmo devotes his entire subsequent chapter to the fundamental properties of fruitfully formulated models. In his view, 'constancy', 'simplicity' and 'autonomy' are the basics expected of any causal relations depicted in models (see his Ch.2). In a world where model construction is delegated out of econometrics *per se*, these basics have been taken as guaranteed by rigorous mathematical deduction in *a priori* model formulation. However, that assumption has been widely falsified in practice, and econometrics has been infested with problems from models which lack those properties.[11] It turns out, from the rise of machine learning theory, that the properties should be utilised as key operational criteria of statistical learning. These criteria facilitate the transformation of a substantial part of the 'art' process into a scientifically operational process. The next chapter explains in more details basic statistical machine learning principles for model construction, with emphasis on notions such as probably approximately correct (PAC) learning, structural risk minimisation and Occam's Razor. In particular, it argues that fruitful model construction has to come from a fusion of prior knowledge and data information through inductive experiments which are targeted at model generalisability, and that the need for making various decisions during the construction leads to uses of mathematical tools well beyond the realm of classical statistics.[12]

Adoption of a machine-learning-theory-based modelling approach will extend and thus reshape the landscape of econometric research. Hence, it is imperative to overhaul the existing probability-based paradigm. The ensuing chapters are an endeavour towards this goal. Chapter 3 examines the effective roles of probability theory in econometric research. The examination reveals that empirically viable econometric models actually dispense with the widely claimed probabilistic foundation, i.e. joint distributions of all the relevant variables. Theoretically, formal representations of economically causal claims are mostly logic based and distribution-free. Correspondingly, econometric models built from and to be used in an open world environment are predominantly discriminative rather than generative. The

[10] It is interesting to find from Kardaun *et al* (2003) how these uncertain aspects are regarded as 'enigma' and 'cryptic issues' by statisticians.

[11] For a historical account of methodological discussions on inadequacy of *a priori* model formulation, see Qin (2013, Ch 9); for discussions on models and theories versus reality from broader methodological perspectives, see e.g. Mäki (2002), and Rodrik (2015).

[12] For reviews and discussions on methodological differences between statistical machine learning and classical statistics, see e.g. Breiman (2001) and Efron (2011).

primary role of probability is therefore to assist the process where those models are discriminatively learnt through synthesis of prior knowledge and data information. The subsequent chapters further delve into this assistant role from three aspects in line with Haavelmo's monograph, namely hypothesis testing, parameter estimation, and prediction.

Chapter 4 exposes the oversimplistic and over-restrictive nature of the hypothesis-testing framework for the task of testing and verifying economic theories against data. The exposure manifests how pivotal it is for model formulation to follow the machine learning approach, how diagnostic function of hypothesis testing takes precedence over its inferential function under the approach, and where limitation of probability-based diagnostic tests lie during the learning process. Chapter 5 extends the discussion onto parameter estimation. Conceptive defects of the estimation-centred mainstream position are highlighted. The primary function of estimation to assist the design and learning of meaningful and estimable parameters of interest is analysed. It is only after models with interpretable parameters are successfully learnt by the criterion of structural risk minimisation that the issue of parameter inference comes onto the research agenda. Basic problems concerning parameter inference post model selection are briefly discussed. Chapter 6 turns to model prediction. By reiterating the pivotal role of generalisability during the model learning process, it appeals for an imperative rebase of the probability approach in econometrics towards decision-making assistance following the PAC learning methodology.

2. Learnability of Economic Relations

Prior to the discussion of the probability approach, Haavelmo devotes his Chapter 2 on the fundamental question of concern to economists, namely 'whether we might have any hope at all of constructing rational models that will contribute anything to our understanding of real economic life' (p11). Noticeably, much of his vision anticipates the principles of machine learning, whereas it has been largely left out of the formalisation of econometrics. His emphasis on 'the degree of permanence of economic laws' (p12) that models should exhibit corresponds to the primary drive for generalisability in machine learning. His description of the passive condition of modellers and of the need to have '*design of actual experiments*' (p13) is effectively the basic setting of statistical machine learning.

The task of this chapter is to revamp Haavelmo's principal vision in light of the theory of generalisation in machine learning. It is argued in Section 2.1 that the various difficulties presented in Haavelmo's Chapter 2 demonstrate that model formulation is nothing but a statistical learning issue. The challenge of fulfilling Haavelmo's requirement of 'reversibility of economic relations' (p17) is now met by learning algorithm which targets at reverse engineering data generation mechanisms. In Section 2.2, the gist of basic machine learning theory is presented. Targeting originally at uniform learning issues, the concept of learnability is formulated around the notion of probably approximately correct (PAC) learning. Empirical risk minimisation is used as the basic optimisation criterion and the bias-complexity tradeoff as the basic model selection principle. The feasibility of the theory in econometrics is also discussed. Extension of the PAC learning theory with respect to non-uniform learning issues is described in section 2.3. Here, the criterion of structural risk minimisation is introduced, along with the notions of Occam's Razor, model consistency and

stability, as well as the equivalence between bias-complexity tradeoff and fitting-stability tradeoff. Haavelmo's desired criteria of 'simplicity' and 'autonomy'[13] have finally been formalised into an operational framework. A summary of the typology of machine learning issues is given in Section 2.4. How the typology can enlighten us on model formulation issues is also briefly discussed there.

It should be noted that one key concern of Haavelmo's is the reversibility of simultaneous-equations models (SEM). However, naïvely proposed static SEMs have long been abandoned in econometric research because of their serious misrepresentation of complex dynamic interactions of economic variables. In any multi-equation dynamic models, how to formulate single equations to minimise uncertainty and maximise generalisability remains the most elementary learning task. Hence, this task forms the focus of the present chapter. Problems and consequences of treating the absence of reversibility of static SEMs as an estimation issue are discussed in Chapter 5.

2.1 Construction of Economic Relations is a Learning Issue

Formulation of economic hypotheses into economic relations for empirical verification evokes certain basic properties on those relations. Three basic properties have been identified by Haavelmo and extensively discussed in his Chapter 2: constancy or permanence, passive observability with respect to available data (i.e. reversibility), and a high degree of 'autonomy' or structural invariance in an open and ever-changing world.

While fully convinced of the possibility of having relational models with these properties, Haavelmo is aware of the challenge of constructing such models. As far as formulation of individual relations is concerned, the difficulties that he discusses at length can be summed up as follows: (1) Prior reasoning is usually based on the premise of rational economic behaviour, which is usually characterised by simple rules of optimality. However, these rules are often too simple and over-sweeping in reality. (2) Choice of a theoretical perspective amounts to narrowing down the focus on specific causal relations and omit other possibly causal factors. The omission is covered by the *ceteris paribus* condition, a condition which is untenable in empirical analyses where data are passively observed from an open world. (3) When the causal relations of prior interest involve multiple input variables which are highly correlated with each other, it is very hard to disentangle the individual effects of these variables using passively observed data. (4) Policy or regime changes are normal occurrence, as agents often find themselves facing new 'milieu' (p.20). Hence, statistical population uncertainty and shifts are unavoidable with passively observed data from an open world.

What underlies all these difficulties is the inevitable uncertainty in *a priori* formulated models, a point already discussed in the previous chapter. Moreover, the desired model properties that Haavelmo put forward are not innate of *a priori* formulated models and entail empirical investigation and evaluation. Hence, Haavelmo's discussion essentially implies indispensability of empirical adequacy in model formulation. This indispensability is phrased, in Haavelmo's parlance, as the necessity of having some 'design of experiments' in mind as 'an essential appendix to any quantitative theory'. This is backed by a quote of B. Russell,

[13] The notion of autonomy was originally put forward by Frisch, e.g. see Qin (2014) for a historical account.

"the actual procedure of science consists of an alternation of observation, hypothesis, experiment, and theory" (p14). On hindsight, his 'design of experiments' foreshadows the approach of computer-aided statistical learning, or machine learning in short. The machine learning approach is developed explicitly for analysing unknown data mechanism in an open world. Here, it is worthwhile noting that this situation is described as *theoryless* by Valiant (2013, Ch.1) in his insightful exposition of PAC learning theory, as differentiated from *theoryful* ones where there already is good scientific knowledge. For *theoryless* problems, the formal mathematical deductive route, which works wonder for *theoryful* matters, such as in physics, fails generally. Our cognition has been dealing with problems in *theoryless* situations mainly by means of common-sense knowledge, and pivotal in that knowledge formation is the role of learning. The goal of machine learning is to mimic that human learning process.

A central task of machine learning is model construction in *theoryless* situations, a task sometimes referred to as 'function estimation', cf. Goodfellow et al (2016, Ch.5). During the learning process, what researchers desire as the necessary model properties are turned into model selection criteria such that computers can help researchers find relations which are not only data-congruent but also best serve their purposes. In other words, the computer age has made it possible to execute various designed experiments essential for the learning task. Let us now set up this task using the framework of machine learning. Denote the causal relation of *a priori* interest by $X \to y$, where X is an input variable set of postulated economic causes of the output variable, y. By economics convention, certain optimal rules under the premise of rational behaviour are imposed to facilitate algebraic derivation of the following functional relation:

$$y = h_p(x_1, x_2, \cdots, x_k; \kappa), \tag{2.1.1}$$

where the causal relation becomes measurable via the parameter set, κ. Since the optimal rules usually imply convexity, that results in linear h_p with respect to κ. However, it is impossible to ensure h_p being data-congruent, especially in view of those prevalent difficulties Haavelmo discussed. Now, denote the data-congruent and ideal/correct functional relation as:

$$y = f(x_1, x_2, \cdots, x_k, z_1, \cdots, z_m; \beta) + \varepsilon \tag{2.1.2}$$

where $(z_1, \cdots, z_m)$ is a variable set disregarded by (2.1.1), $\kappa \subset \beta$ and ε represents ignorable noises/errors. Clearly, the target function f is unknown *a priori* and not completely knowable *a posteriori* due to the presence of ε. It is thus the researchers' task to search for the best possible approximation, $h_d \approx f$. The search entails designing learning algorithm which should incorporate the rules of what constitute a credible relation together with the criteria of what data-congruent relations are. Specifically, the latter should include parameter constancy with respect to κ and structural invariance with respect to h_d, in view of Haavelmo's three properties. Discernibly, his requirement for 'the reversibility of economic relations' is already built into the setting of the learning task, $h_d \approx f$, whereas prevalence of the difficulties listed by Haavelmo amounts to $h_p \not\approx h_d$ in general.

The above setting effectively assumes the feasibility of circumventing, by means of empirical solutions, the impasse of finding data-congruent relations through analytical

solutions. Is this route of inductive learning reliable and credible? An affirmative answer can be found in the computational learning theory.

2.2 Reversibility Embodied in PAC Learnability

Theories of computer-aided statistical learning are well developed and widely introduced in machine learning textbooks,[14] as a result of concentrated research mostly during the second half of the 20th century. This section offers only a very short summary of the gist of the theories.

Following the setup in the previous section, we define the learning task as seeking the best approximation of the unknown function $f: \mathcal{X} \to \mathcal{y}$ with the help of available data set, $\mathcal{D}$, e.g. see Friedman (1994, 1997). A hypothesis class, $\mathcal{H}$, i.e. a certain class of possible functions is proposed with respect to prior knowledge. To choose the best possible $h_{\mathcal{D}} \approx f$ from $\mathcal{H}$ using $\mathcal{D}$, certain data optimisation criteria are needed. One obvious criterion is to minimise empirical noises/errors, which is known also as empirical risk minimisation (ERM). By using a certain loss function, ℓ, ERM is realised at $h_{\mathcal{D}} = argmin_{h \in \mathcal{H}} E_{in}$, where E_{in} denotes the within-sample error term, i.e. $E_{in} = \mathrm{E}[\ell(h)]$.

Since generalisability of $h_{\mathcal{D}}$ is the ultimate goal, we need to go beyond ERM and examine the properties of the out-of-sample error, E_{out}, with respect to E_{in}. The model, $h_{\mathcal{D}}$, is said to be PAC, i.e. probably approximately correct, if, for a given accuracy or performance level, ϵ, also referred to as the accuracy parameter, the following probability is tolerably small:

$$P[|E_{out} - E_{in}| > \epsilon] \leq \delta, \tag{2.2.1}$$

where the threshold, δ, is commonly referred to as the confidence parameter, in that we are confident of $h_{\mathcal{D}}$ being PAC learnable at the level of $1 - \delta$. Noticeably, the concept of PAC learnability is defined by the two approximation parameters, $\epsilon, \delta \in (0,1)$, without any distributional assumptions on $\mathcal{D}$. Hence, the framework is often referred to as a distribution-free learning one.[15]

In order to pin down the circumstances under which PAC learnability of $h_{\mathcal{D}}$ is feasible with high confidence, we need to examine the size of the generalization error, $|E_{out} - E_{in}|$, as well as its properties. The examination results in formalised theorems on the conditions of convergence of E_{out} using the ERM rule, and especially on the generalisation bounds implied in (2.2.1): $E_{out} \leq E_{in} + \epsilon$, such as bounds defined by Hoeffding inequality, VC (Vapnik-Chervonenkis) dimension and Rademacher complexity. Roughly, these bounds can be summarised as:

[14] PAC theory of learnability was proposed by Viliant (1984). Vapnik (1999, 2003) provides an authoritative framework and overview of statistical machine learning theory. For a detailed and formal account of the theory of machine learning, see Shalev-Shwartz and Ben-David (2014, Part I); explanations of the theory at far less technical levels can be found in Abu-Mostafa et al (2012, Chs 1 & 2) and Russell and Norvig (2016, Part V).

[15] More precisely, this distribution-free PAC learning theory is described as *agonistic* PAC learnability, cf. Shalev-Shwartz and Ben-David (2014, Ch. 4). Consequently, PAC learning should be 'viewed as substantially assumption-free and not as imposing substantive constraints' from a cognitive perspective, Valiant (2008, p420). It is also worthwhile noting that PAC learning is regarded as 'rational approximations to concepts' (Doyle, 1992, p399) from the lens of economic principle of rationality.

(2.2.2) $$E_{out} \leq E_{in} + \mathrm{O}\left(\sqrt{\frac{d}{N} lnN}\right),$$

where N denotes the size of $\mathcal{D}$, d is a dimensionality parameter of $\mathcal{H}$, and the big-O is a standard asymptotic notation for the limiting property of the term. Essentially, the argument of the term demonstrates a close relationship between the complexity of $\mathcal{H}$ and the size of $\mathcal{D}$. The more complex $\mathcal{H}$ is, the larger data samples are needed to keep the argument, i.e. the bound, non-increasing. For given $\mathcal{H}$ and chosen values of $\epsilon, \delta \in (0,1)$, the resulting N by (2.2.2) is referred to as 'sample complexity'.

Since $h_{\mathcal{D}}$ is an approximation of the assumed true, but unknown f, its choice is bound to erring in a passively observed open world, as reflected by $\delta > 0$. This can be further examined through a conceptual decomposition of the expectation of E_{out} in respect to f:

(2.2.3) $$\mathbb{E}[E_{out}] = [\mathbb{E}(h_{\mathcal{D}}) - f]^2 + \mathbb{E}\left[\left(h_{\mathcal{D}} - \mathbb{E}(h_{\mathcal{D}})\right)^2\right] = \epsilon_{app} + \epsilon_{est}$$

The above decomposition reveals the essence of the generalisation goal: making trade-off decisions between approximation error, ϵ_{app}, and estimation error, ϵ_{est}. Since ϵ_{app} measures the deviation of $\overline{h_{\mathcal{D}}}$ from f, this term is widely regarded as model approximation *bias* induced by the choice of $\mathcal{H}$. This bias can be reduced by expanding $\mathcal{H}$, i.e. increasing d. ϵ_{est} measures empirical risk once $h_{\mathcal{D}}$ is determined; it usually increases as d rises and decreases as N rises. It should be noted that the unknown nature of f determines ϵ_{app} as a purely theoretical notion. In practice, the generalisation goal amounts to choosing, via monitoring ϵ_{est}, $h_{\mathcal{D}}$ where the two opposite tendencies strike a balance. This tradeoff is widely referred to as bias-variance or bias-complexity tradeoff in the machine learning literature.

Execution of the tradeoff examination for the model selection purpose entails the partition of available $\mathcal{D}$ into two parts: a training set and a test set, i.e. $\mathcal{D} = \mathcal{D}_{train} \cup \mathcal{D}_{validation}$ and $\mathcal{D}_{train} \cap \mathcal{D}_{validation} = \emptyset$.[16] The partition enables derivation of E_{in} from $\mathcal{D}_{train}$, as well as simulation or approximation of E_{out} by $E_{validation}$ from $\mathcal{D}_{validation}$. Both error terms are computed through execution of algorithm, $\mathcal{A}$, which is programmed to combine $\mathcal{H}$ and ℓ. The design of $\mathcal{A}$ plays a vital role in machine learning, and $\mathcal{A}$ is sometimes referred to as *perceptron* learning algorithm (PLA), following the byname of neural networks as 'perceptrons' in artificial intelligence (AI).

It is worthwhile noting that the reckoning of ϵ_{app} as bias and the focus on bias-complexity tradeoff help shed light on where the probability-approach-based econometrics has gone astray epistemologically. The focal attention on estimation bias in econometrics, as well as wide conviction of its presence, is predicated on the untenable assumption of no approximation bias in any algebraically deduced economic models.

But is the paradigm of PAC learning theory indeed applicable for econometric research? Our previous discussion has already argued for the inevitability of *a posteriori* approximation of the *a priori* unknownable f. What remains to be examined are: (a) How compatible the three components, namely $\mathcal{H}$, ℓ and $\mathcal{D}$, in PAC learning theorems are with their counterparts

[16] In machine learning textbooks, data are split into three subsets for large samples: training, validation and test subsets. The validation subset is for model selection, whereas the test subset is for model prediction assessment.

in econometrics; and (b) if compatible, what limitations the learning theory has when applied to econometric modelling. A brief discussion of the two issues is offered below.

Almost all positive theoretical models in economics are deduced on the basis of certain optimality rules of rational economic behaviour, such as maximising utility or profit, or minimising risk or cost. These rules effectively enable economic relations to be representable by a convex hypothesis class, as depicted by (2.1.2). It is upon this class type of $\mathcal{H}$ that the foundations of PAC learning theory are built. Within the convex class, a family of linear relations, known as linear predictors in machine learning, is used for two types of targets: (i) when $\mathcal{y}$ is continuous, and (ii) when $\mathcal{y}$ takes discrete, especially binary, values to represent classification issues. Consequently, the approximation criterion of ERM leads to two common types of ℓ: regression loss based on a quadratic or squared error loss function and classification loss based on a logistic function. It is in the latter context that PAC learning theory is originally developed, before its extension to the former. In applied econometrics, both model types are also extensively used, even though differences exist in the objectives of optimisation underlying the loss functions. The original PAC learning theory assumes randomly sampled data, corresponding to the i.i.d. (independently and identically distribute) assumption of classical statistics. Recently, the theory has been extended to the situation of stochastic time-series samples, where the assumption of randomly sequenced sampling method is removed, e.g. see Zimin and Lampert (2017), Dawid and Tewari (2020). On the whole, a fairly high degree of commonality is discernible.

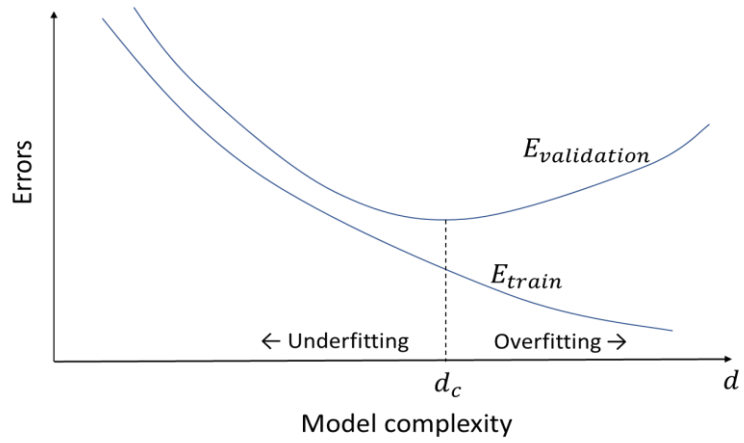
Available data sample sizes can be a constraint to the application of PAC learning theory to econometric modelling. But this is generally a weak constraint in view of what data have been used in the existing literature. Most of the hypothesis classes and associated loss functions in that literature are in routine use in machine learning, as pointed out above. Adoption of the machine learning approach should actually unleash research potential of empirical issues which are more complex than what is possible under the current econometric approach. Nevertheless, we ought to be aware that the range of theoryless questions that PAC learnability can tackle faces severe computational limitations, an issue commonly referred to as computational complexity, e.g. see Valiant (2013, Chs.3 & 5), and also Shalev-Shwartz and Ben-David (2014, Ch. 8). A key deciding factor here is how much simplification is feasible in the formulation of representations of the complex issues of interest. In fact, explicit pursuit for simplicity is fundamental in machine learning. This is particularly evident from the introduction of structural risk minimization (SRM) into learning theories, in order to overcome the over-restrictive condition of uniform convergence of empirical risks implied in ERM, upon which PAC learnability is essentially formulated. Remarkably, theories of nonuniform learnability are formulated under conditions, which coincide with ‘simplicity’ and also ‘autonomy’, two desired criteria of model formulation that Haavelmo has discussed at length in his Chapter 2.

2.3 Simplicity and Autonomy: Occam’s Razor and Stability

ERM can become an ill-posed induction criterion for the task of generalisation when only finite samples from in an open-world environment are available, because uniform convergence as $N \to \infty$ is a non-dependable state, see Mukherjee *et al* (2006). Under the

circumstance, a learner can only try and make the best possible decision on the bias-complexity tradeoff shown in (2.2.3) through experiments using available $\mathcal{D}_{train}$ versus $\mathcal{D}_{validation}$. The outcome is illustrated via the learning curve shown in Figure 2.1. While E_{train}, which represents errors of a fitted model, is decreasing with the rise of model complexity measured by d, $E_{validation}$, which represents model prediction errors, is not. Movements of $E_{validation}$ are d dependent. It will start increasing once d surpasses a certain threshold point, d_c. Hence, models with $d < d_c$ underfit the data whereas models with $d > d_c$ overfit the data. In the event of having finite sample information, models which are selected under ERM without undergoing the tradeoff experiments may not be the best generalisable ones. That implies that we can exploit the tradeoff decision rule to make up for avoiding the uniform convergence requirement under ERM. In other words, we can extend the rule of maximising model fit by combining it with the rule of minimising model predictive errors. The latter implies going for the simplest possible models with respect to the prediction purposes. This principle of model parsimony is widely referred to as Occam's Razor, attributing to Occam's philosophical idea that the simplest possible explanation is usually the best one. On reflection, Occam's Razor effectively reflects and articulates the gist of Haavelmo's discussion on simplicity for model formulation.

Figure 2.1 Diagram of Learning Curve



Consequently, the criterion of SRM is proposed to replace ERM for nonuniform learnability. SRM can be represented as follows, see Abu-Mostafa *et al* (2012, p178). Define a hierarchy or nested sequence of hypothesis classes as a 'structure': $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \cdots \subseteq \mathcal{H}_d \subseteq \cdots$; select one hypothesis from each $\mathcal{H}_i$ under ERM, i.e. $h_{i,\mathcal{D}} = argmin_{h\in\mathcal{H}_i} E_{i,in}$; make the final selection among the selected $\{h_{i,\mathcal{D}}\}$ by imposing, on the E_{in} minimisation, a penalty rule, Λ, for model complexity:

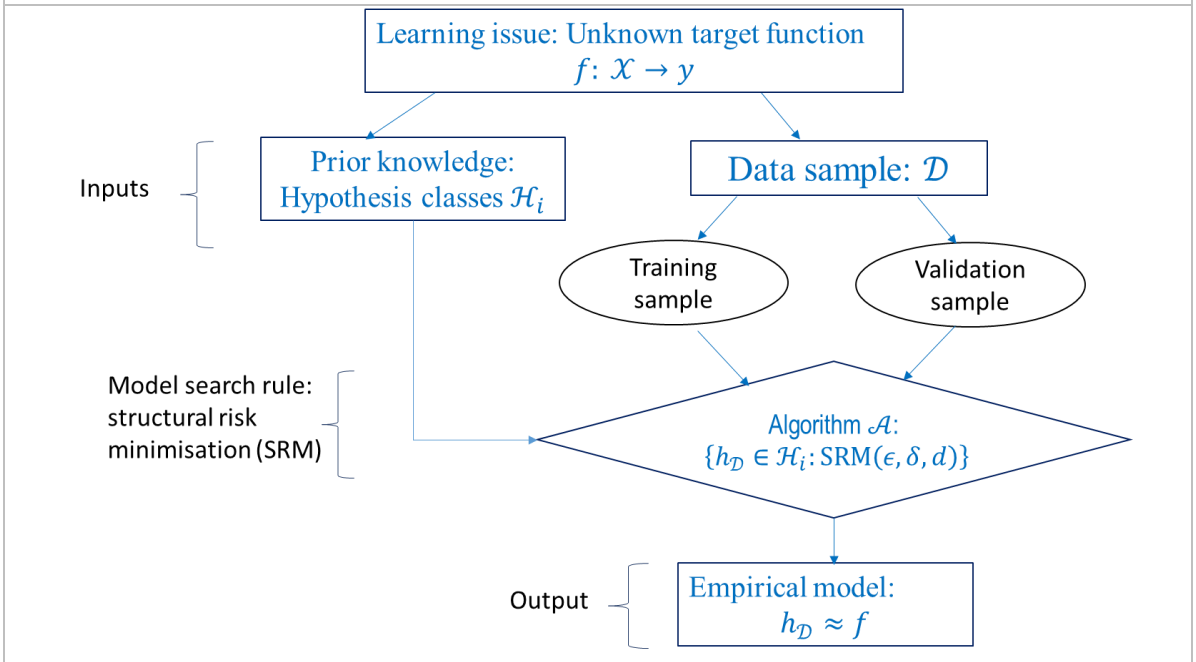
(2.3.1) $$h^*_{i,\mathcal{D}} = argmin_{1,2,\cdots}\left[E_{i,in}\left(h_{i,\mathcal{D}}\right) + \Lambda(\mathcal{H}_i)\right]$$

The resulting $h^*_{i,\mathcal{D}}$ is proved to embody the principle of Occam's Razor, in that its selection is at the best bias-complexity tradeoff point, d_c, in Figure 2.1. The selection thus fences off the

overfitting risk. The last step in the SRM process amounts to searching, among all competing hypotheses, for the one with the smallest generalisation bound. [17]

The pivotal role of predictive error minimisation in model selection leads to growing scrutiny on the tradeoff properties by means of resampling techniques. From the perspective of varying sample sizes in the division between $\mathcal{D}_{train}$ and $\mathcal{D}_{validation}$, the minimisation implies that $h^*_{i,\mathcal{D}}$ should be consistent in their predictive performance. In other words, the expectation of the predictive errors should show convergence with changing test sample sizes. This consistency is formalised as a stability property measurable from $E_{validation}$, and shown to be a generic condition for generalisability or learnability, e.g. see Poggio et al (2004), and Shalve-Shwartz et al (2010). The condition of stability plays a particularly important role in learning algorithm designs. It facilitates the operational search for the best possible bias-variance tradeoff by matching it to the best fitting-stability tradeoff, see Shalev-Shwartz and Ben-David (2014, Ch13). Discernibly, theorisation of 'stable rules do not overfit' (p.173, ibid.) provides us with a parsimonious formalisation of Haavelmo's advocacy for autonomy as well as simplicity in model formulation.

Figure 2.2 Basic Setup of Nonuniform Learning Problems



Adaptation from Figure 1.2 in Abu-Mostafa *et al* (2012).

The basic setup of the above machine learning paradigm is illustrated in Figure 2.2. It is worth noting here that the paradigm refutes what is popularly perceived of machine learning in the econometric community: Machine learning is black-box based and solely data-driven

[17] Noticeably, the SRM-based learning theory crystalises, to a great extent, the main theme of a 'progressive research strategy' proposed by Hendry and Richard (1982) in econometric modelling. In particular, Occam's Razor is embodied in the parsimonious encompassing principle, e.g. Hendry (1995, Ch14), although the principle is not focally and exclusively targeted at the generalisation capacity of models via the test stage subsequent of sample partition.

towards predictive fit, in sharp contrast to causal-explanation centred econometrics. The perception stems probably from the observation that data analysis techniques form the core of machine learning textbooks. But that is because the initial formulation of $\mathcal{H}_i$ entails domain knowledge not because $\mathcal{H}_i$ is dispensable. The indispensable role of $\mathcal{H}_i$ is clearly specified in various learnability theorems. The fruitfulness of learning is embodied in sample complexity and tightness of the generalisation bounds. Both measures are predicated on the choice of $\mathcal{H}_i$. In particular, the chosen hypothesis space ought to be large enough for solutions to the research issues of interest, whereas small enough for generalisation checks within available data samples. Both underfitting and overfitting should be avoided.

Furthermore, two fundamental notions in econometrics – bias and consistency, are conceived and conveyed in a strikingly different context. Specifically, bias is an innate quality of any models from the perspective of PAC learnability. All one can do for fruitful model selection is to go for the best possible bias-complexity tradeoff. Correspondingly, design of consistent algorithms becomes a prerequisite for the learning process. Both notions clearly transcend the narrow context of parameter estimation, prompting us to reassess their methodological basis in econometrics.

Obviously, SRM-based model selection involves multiple decision-making tasks. Their implementation demands for tools well beyond what the toolbox of classical statistics can offer. In particular, the design of $\mathcal{A}$ demands careful consideration of the interplay of computational and statistical tools. Furthermore, it entails improvements and renovations of the processes whereby prior knowledge and data information are infused. To embrace and synthesize the paradigm will be game-changing in econometrics.

2.4 Statistical Learning Paradigm and Formulation of Econometric Tasks

Theorisation of PAC-based learnability is originally built on what is classified as supervised learning problems. The domain of supervised learning comprises data samples where variables are sorted into inputs and outputs/responses, and the goal of modelling is explicitly targeted at predicting the latter group. Variables of this group are referred to as *labelled* (correct) targets. Their sample counterparts are treated as representative examples capable of teaching or guiding the learning of ways in which inputs would yield the targets systematically. Two commonly used model types are (i) classification or logistic regression models for predicting class labels, and (ii) regression models for predicting numerical labels.

In contrast to supervised learning, there are no labelled output examples to guide pattern or structure recognition in the domain of unsupervised learning. All the variables are treated as inputs. The conventional statistical tool of principal components analysis is categorised into this domain, whereas clustering analysis occupies a more centred stage nowadays. The lack of leverage of labelled examples makes it more challenging to fulfil unsupervised learning tasks than supervised ones. A major obstacle is how to choose and formulate adequate and also meaningful criteria under which both model selection and evaluation can be performed. Prior bias is inevitable during the choice and the formulation, and it is embodied in the model outputs.

An extension of the two categories is known as semi-supervised learning, e.g. see Chapelle *et al* (2006), van Engelen and Hoos (2020). This type of learning deals mainly with the situation where only a limited number of labelled examples is available. The examples act as partial guidance for the search of interesting structures from abundantly large amount of unlabelled data. Alternatively, learning tasks with unlabelled data can become semi-supervised ones when there are available prior constraints, which are utilisable as partial supervisors of the tasks. Semi-supervised learning relates intimately with a type of learning alternative to induction: transductive inference, i.e. learning general rules from specific examples for the purpose of inferring about pertinent specific classes.[18] In other words, learning by transduction aims at inferring from-specific-to-specific issues rather than from-specific-to-general (or from-sample-to-population) issues. A popular toolkit of transductive learning is cluster-based nearest neighbour classifier algorithms, e.g. see Gammerman *et al* (1998).

The domain of supervised learning can clearly accommodate a large amount of econometric modelling tasks, tasks which are led by hypothetically formulated causal models and executed with data samples containing labelled output variables corresponding to those models. In fact, regression and classification models have enjoyed prevalent application in econometrics. Often, economic theories impose more conditions than prediction of the labelled output variables. For example, certain causal relations between input and target/response variables should be negatively proportional, and others incrementally positive. However, such conditions will not alter the nature of supervised learning as they can be regarded as additional supervising constraints. Hence, the learnability theory provides us with a promising way out of the knotty problem of uncertainty in model formulation discussed in Chapter 1.3. It shows us what conditions are necessary concerning hypotheses and data in order for us to turn the art of model formulation into science. It also shows us how the framework of inductive learning differs from that of classical statistics. From the former perspective, the learning tasks of parametric inference in statistics belong to deductive learning.

The classification of different learning issues in machine learning further prompts us to reflect on possible pitfalls in the formulation of different econometric issues. Two cases are worth noting here. The first is the century-long endeavour to construct leading indicators in business cycle research. The task remains to be treated as an unsupervised learning one essentially, despite various technical enrichments on principal component analysis, the toolkit originally used.[19] Unfortunately, progress remains painfully slow if judged by the usefulness of model-generated leading indicators in macro forecasting. The statistical learning paradigm cautions us of a formulation problem. Since macro forecasting is the goal of indicator construction, the goal should be incorporated in the model formulation. Unsatisfactory indicators produced by models without such incorporation should be seen as signs of a model formulation error for omitting supervised information. The second case is the formulation of

[18] Transductive inference is pioneered by Vapnik during the mid-1970s; it embodies Vapnik's principle – when solving a given problem, one should try to avoid a more general problem as an intermediate step; for summary reviews, see Gammerman *et al* (1998) and also Chapters 24 and 25 of Chapelle *et al* (2006).
[19] For a comprehensive survey, see Marcellino (2006).

a type of models commonly known as selection models in microeconometrics (also popularly referred to as the tobit models in econometrics textbooks). What this model tries to deal with is exactly the semi-supervised learning situation, i.e. where only a limited number of labelled examples is available. From the viewpoint of the ordinary regression model approach, however, this data feature is conceived as a censored or truncated data problem and treated as an estimation issue. On reflection, the primary underlying economic issue here is nothing but one of transductive learning, be it habitual consumption or labour supply patterns of the specific group with labelled examples. More detailed discussions on formulation issues of these two cases are in Chapter 4.

Detrimental consequences of poorly conceived model formulations are too obvious to stress. A quote from "Deconstructing Statistical Questions" by D. Hand (1994) suffices. Model formulation is 'a higher level issue' than that of 'application of statistical tools to identify structure and patterns in data', because it 'determines what the questions are in the first place and which tools should be used' (p317). Correspondingly, formulation errors are referred to as type III errors among the statisticians' community, i.e. errors of seeking right answers to wrongly formulated research questions.

In fact, extensive studies of the 'higher level issue' can be found in the AI research, usually under the headings of 'knowledge representation' and 'machine reasoning', e.g. see Russell and Norvig (2016, Part III). The studies delve deep into the question of how human knowledge can be translated into computable media faithfully and efficiently to facilitate machine learning. Additional to reversibility and parsimony, two other characteristics are noted to be most fundamental in knowledge representation: ontological commitments and epistemological commitments (Chapter 8, ibid). Essentially, any knowledge representation conveys a set of ontological commitments, i.e. postulated natures of reality. While the set anchors the perspective and framing of the research, its choice does not involve data structures. From the epistemic angle, on the other hand, any representation comprises a set of inferences theorising reality in a fragmentary way. This is referred to as 'a fragmentary theory of intelligent reasoning' by Davis *et al* (1993). The choice of the inference set is intimately related to how the representation is to be engineered computationally. Here, a probabilistic causal reasoning approach is reckoned as a natural response to the inevitable uncertainty implied in the fragmentary attribute. But it is not the only approach, and this is evident from the *distribution-free* framework of PAC learnability. While uncertainty is explicitly accounted for in the very notion of PAC learning, theorems of generalisation gear machine learning systematically away from making any prior assumptions about the underlying distribution of data or variables in the hypothetical model classes. The framework accords with and is reflective of the ontological commitments of AI knowledge representation. Adoption of this framework into econometrics will challenge the fundamental position of the probability approach. What functions has the approach played in econometrics and how its role should be revised in light of the learnability theory? A close examination of these questions is the task of the next chapter.

3. Basic Functions of Probability in Econometrics

Conceptualisation of all observable economic variables as jointly distributed random variables lies at the very foundation of Haavelmo's promotion for the probability approach. This cognition, however, is predicated on an implicit assumption: The underlying economic milieu of the variables concerned be closed, adequately simple to resemble the experimental setting of classical statistics. This is a *critical* assumption in econometrics, since systematic adaption of statistical inference methods would have been impossible without a communal consent to Haavelmo's vision.

'Unrealistic assumptions are OK; unrealistic critical assumptions are not OK', states in Rodrik's 'Ten Commandments for Economists' (2015, p.116). The above assumption is clearly far from realistic in general, but is it realistic enough for econometric research purposes? It is already argued in Chapter 1 that practical usefulness of probabilistic notions for uncertainty representation is rather limited. They are not useful in analysing variables without any predetermined modelling tasks. It is further argued in Chapter 2 that, as far as those purposes are concerned, uncertainty in model formulation in an open world setting is too great to sustain the *a priori* algebraically deductive route, and that the formulation should therefore be tackled as a learning issue following the machine learning approach. The approach bypasses the above assumption, effectively suggesting it premature to treat empirical modelling tasks as statistical inference ones before relatively credible models are learnt. To assess and ensure the feasibility and necessity of the machine learning approach, this chapter delves into the fundamental issue of how probabilistic notions have functioned and should better function in econometric research.

Section 3.1 starts the discussion by assessing the successes and failures of econometric modelling research into joint, marginal and conditional distribution-based types of models guided by the factorisation principle of the basic chain rule of probability. The assessment reveals clearly a lack of empirical support for the factorisation principle. Since conditional models remain virtually the only empirically viable type, section 3.2 reflects on how conditional models are generally formulated. It transpires from the reflection that theoretical derivation of causal relations, which form the core of those conditional models, is largely probability free. Probability is invoked only after the derivation, when random shocks or error terms are appended to those relations. In other words, those conditional models are by nature *discriminative*, rather than *generative*, by the classification of machine learning. Section 3.3 maintains, by resorting to knowledge representation theories in AI, that economic theories are innately derived from logical reasoning instead of probabilistic reasoning. Hence, conditional econometric model research concords with the distribution-free stance in machine learning. Because of the unignorable and uncertain gap between logic-based relations and reality, it is indispensable to synthesise formalised prior knowledge and data examples via inductive learning. It is during the learning process that probability comes in. It plays primarily a discriminative role – assisting decision making during model selection. Acknowledgement of this role will undoubtedly demand a major shift of position in econometrics, away from the currently statistical inference centred position. However, the shift is imperative in view of widespread inefficiency in econometric practice, in noticeably sharp contrast with the rapid achievements in other applied fields through utilising AI and machine learning methods.

3.1 Stochastic Model Characterisation of Observable Variables as Target Variables

The question under scrutiny in this section is how the assumed stochastic characterisation of all observable economic variables has fared in econometric practice, particularly with respect to whether the characterisation is adequately realistic. Since the characterisation is obviously an oversimplistic representation of the uncertain features of economic variables, and since the range of those features is much wider than what is considered in econometric practice, as already pointed out in Chapter 1, the scrutiny is restricted on variables which are set as modelling targets. Furthermore, the question of whether the characterisation is realistic enough for the modelling task is approached through assessment of how empirically fruitful the characterisation is in shaping broad research strategies and directions of econometrics.

One of the most basic rules of probability theory is the chain rule – any joint distribution of random variables can be factorised into products of conditional and marginal distributions. For instance:

(3.1.1) $$P(x,y) = P(y|x)P(x) = P(x|y)P(y).$$

In the context of modelling targets as jointly distributed random variables, (3.1.1) implies that we should be able to model them either jointly, or at their factorised level, and that factor decomposition makes the second route look more scientifically fundamental than the first. To a great extent, econometrics has evolved around these implications. The groundwork built on SEMs in macroeconometrics is sustained by the conviction of the primacy of modelling variables with direct reflection of the joint-distribution character.[20] The legacy of this groundwork is best capsuled by the notion of 'endogeneity bias', thanks to Haavelmo's derivation of the OLS bias in the context of a bivariate static SEM. The conviction is so prevalent that academic concern over endogeneity bias ranks top when models fall into the conditional type. This is particularly noticeable in microeconometrics. The concern *de facto* acknowledges the incompleteness of conditional models in capturing the assumed underlying joint distribution. On the other hand, macroeconometrics has largely evolved along the factorisation route since the mid-1970s in the wake of the 1973 oil crisis. This is evident from the active pursuit of data generation processes (DGP) of modelled variables, particularly from the classification of variables by their individual stochastic time-series features, and also the ensuing widespread acceptance of the fundamental role of such classification prior to model dynamic specification choices.

In contrast to impressive technical advances, empirical gains have remained meagre. Applied endeavours are generally abortive when the assumed underlying distributions of the target variables are either joint or marginal. Comparatively satisfactory results are achievable only from models where the assumed underlying distributions are apparently of the conditional type. Econometrics appears to be stuck in the impasse of 'incomplete' conditional models if judged by empirical success.

Let us look into this impasse using models using time-series data. Consider first the type of models based on the marginal distribution assumption, as they should be the most elementary from the factorisation perspective. This model type is popularly used in characterising macro variables, especially their temporal persistence observable from time-

[20] For a historical account, e.g. see Qin (1993).

series data. An established way of categorisation of the persistence is (weak) stationarity, defined in terms of the sample-size invariance of the first two moments of the variable under concern. This property is commonly examined via the characteristic roots of autoregressive (AR) models. Take the simplest 1st-order AR for example:

(3.1.2) $$y_t = \rho y_{t-1} + u_t,$$

where y_t is referred to as stationary if the estimated parameter $|\rho| < 1$, and as nonstationary with a unit root if $\rho = 1$. The theory of cointegration analysis is predicated on the presence of unit roots. Evidence of unit-root presence has been reported widely in the literature. However, the evidence seldom stands up for assessments of generalisation. Unit-root estimates are highly sample and/or frequency dependent. Attempts to represent marginal distributions in more general ways than the AR model type have resulted in various more complex univariate models. Nevertheless, it is virtually impossible for these models to avert out-of-sample shifts in estimated moments of single variables. Moreover, contrary to the practice in controlled statistical experiments, where highly irregular observations are usually disregarded as accidental outliers, 'anomalous' observations causing nonconstancy in the moments in univariate econometric models often convey useful, if not important, information. They are seen primarily as useful indications of regime shifts rather than accidental noises. In short, what the model type like (3.1.2) serves us is a summary description of common dynamic symptoms of economic time series. It exhibits outcome rather than the underlying data generative mechanism *per se*. Here, economists' explanation is simple: economic variables are fundamentally interdependent. Empirical pursuits along the univariate modelling route are susceptible to go conceptually astray.

Indeed in practice, unit-root check-ups are used predominantly as a preliminary step for conditional modelling research. Take the popular autoregressive distributed-lag (ARDL) model for example. Supposing that the following ARDL(1, 1) model is found to be empirically adequate:

(3.1.3) $$y_t = \alpha y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + \varepsilon_t$$

Note that $|\alpha| \ll 1$ is a key prerequisite for the empirical adequacy of (3.1.3). But if (3.1.3) is empirically adequate, (3.1.2) cannot be so at the same time, because the latter must be underfit due to omission of x_t and x_{t-1}, and ρ is bound to suffer from omitted variable bias (OVB). Specifically, the bias is an upward one if $\hat{\rho} \approx 1$ is estimated from (3.1.2). According to the chain rule, x_t should obviously be the target of the marginal model to pair with (3.1.3). Hence, the view of y_t being fundamentally driven by a unit-root DGP is cognitively flawed. In practice, however, a marginal univariate model of x_t is neither necessary nor plausible in general. From the viewpoint of modelling y_t, there is no need to learn about the distribution of x_t. From the viewpoint modelling x_t, the marginal modelling route is clearly inadequate, as pointed out above. Modelling strategies guided by the factorisation rule of (3.1.1) misfit economic reality.

Let us now take a look at empirical results following the joint distribution modelling route. Currently, the most representative model type of this route is a vector auto-regressive (VAR) model, seen commonly as a dynamic extension of SEMs. However, any empirically

operational multi-equation models are innately of the conditional type.[21] The target variable set of a VAR model is conditioned upon all the lagged variables. Let us review the practical performance of VARs through a simple VAR, which is extended from (3.1.3):

$$(3.1.4)\ \begin{pmatrix} y_t \\ x_t \end{pmatrix} = \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix} \begin{pmatrix} y_{t-1} \\ x_{t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix},\ \text{assuming: } \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix} \sim IIN \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix},\ \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} \right]$$

Preference of (3.1.4) over (3.1.3) in the macro profession reflects a strong faith in the symmetric factorisation, i.e. $P(y|x)$ or $P(x|y)$ in (3.1.1). However, maintenance of simultaneity in VAR models is but skin-deep. Choices of variables to be included in VARs are made in accordance to particular modelling purposes, which imply inevitably certain directionally causal concerns. For example, in a four-variable VAR model of GDP, inflation, unemployment and interest rate, the variable choice is biased towards explaining the real economy in that the model does not give interest rate as equal causal focus as GDP. Consequently, individual equations within a VAR never yield similar or comparable levels of statistical performance. When estimated VARs are used for impulse response analysis, imposition of specific sequential variable orders is required to initiate the analysis. The imposition violates the apparent simultaneity of VARs. Moreover, VARs like (3.1.4) are not considered generally as 'structural' models. Attempts to introduce 'structural' elements lead to explicitly asymmetric VARs, e.g. the following VAR with a simultaneous element:

$$(3.1.5)\ \begin{pmatrix} 1 & \beta_{12} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} y_t \\ x_t \end{pmatrix} = \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix} \begin{pmatrix} y_{t-1} \\ x_{t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix},\ \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix} \sim IIN \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix},\ \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \right]$$

Finally, the lack of symmetry is also noticeable from two other facets. First, certain parameter estimates are bound to be statistically insignificant in an assumed symmetrically structured VAR, indicating redundancy of the corresponding input variables, and thus asymmetricity in model structure; second, improvement of model fit is usually achievable by extending a close VAR to an open VAR by introducing additional control variables, which effectively switch the model basis from the joint to the conditional distribution.

In sum, any postulated statistically operational models have to be causally asymmetric, e.g. see Cox (1992). Indeed, what have prevailed in applied econometric research are models of the conditional type, whether they are single-equation or multi-equation models, if assessed by their practical relevance and statistical performativity. This poses a perplexity: If joint distribution is the most basic representation of the fundamentally stochastic mechanism of an economic issue of concern, why are empirically viable models predominantly of the conditional type? In fact, it is in the context of a conditional model that Haavelmo first introduces probability distribution formally. The introduction effectively extends (2.1.1) by a randomly distributed error term, *s*, 'a random variable having a certain probability distribution' (Haavelmo, 1944, p51):[22]

$$(3.1.6)\qquad y = h_p(x_1, x_2, \cdots, x_n; \kappa) + s.$$

[21] Note that any estimable SEMs have to satisfy conditions of identifiability. These conditions effectively turn symmetric SEMs into asymmetric ones; for a historical account of the debates over simultaneity and conditioning during the formalisation of econometrics, see e.g. Qin (1993, Chs 4 & 6).

[22] In order to keep notations consistent across different chapters in the present monograph, Haavelmo's original notation (see equation (11.2) in Ch3 of his monograph) has been modified here.

Nowadays, (3.1.6) is the workhorse of applied econometrics, where the key role of s is to enable and justify the use of confidence interval based statistical estimates for κ. Definition of all variables as randomly distributed ones is merely a formality. Interestingly, Haavelmo has not further extended probabilistic characterisation to the variables either, after his introduction of s. Instead, his attention is turned to what he sees as a fundamental important issue: the method of splitting y between the systematic part, h_p, and the residual part, s. He devotes one section of his monograph on this issue (section 12).

It is already pointed out in Chapter 2 that the issue of splitting is essentially a PAC learning one beyond the scope of classical statistics. From that perspective, we should be able to incorporate (3.1.6) into the distribution-free approach of machine learning. This possibility offers us an attractive way out of the perplexity raised earlier. But it also confronts us with the following question: How indispensable is probabilistic reasoning in tackling the modelling uncertainty in econometric practice? Or where does the indispensability exactly lie?

3.2 Role of Probability in Economic/Econometric Modelling

Let us first reflect on the economic modelling side. Specifically, how do economists usually deal with uncertainty in their formulation of h_p? At the outset, a great deal of uncertainty is filtered out by the theoretical focus on the expected rational behaviour of representative agents. Causal relations are deduced by rules of optimisation, e.g. utility/profit maximation or cost/risk minimisation. The rules are deemed fundamental as they capture the key economic motives of decision-making behaviour. Since none of the rules are stochastic, mathematics used for the derivation of h_p involves mainly calculus and constrained optimisation apparatus. The assumption of *ceteris paribus* is used to cover up for the partiality of the rules, as embodied by the limited set of input variables. It is only after the derivation that stochastic uncertainty is considered. This is done by appending *a priori* functionally fixed h_p with a random error term, so that the resulting model becomes equivalent with the targeted models in classical statistics. Haavelmo's argument of the variables in h_p being stochastic is to solely justify the equivalence. Uncertainty in (3.1.6) is only allowed from two outlets: estimates of κ in h_p and s, whose essential role is to accommodate the estimation task. Hence, probabilistic reasoning plays virtually no role on the economic modelling side.

Much of economists' endeavours to narrow down the gap between theoretical models and reality are concentrated on relaxing the idealistic setting of the expected rational behaviour of representative agents. The relaxation effectively targets at identifying specific situations where empirical verification of the optimisation rules is expected to become sharper than without such identification.[23] Accordingly, rule-based causal relations are deduced subject to the specifics of situations under concern. These specifics occur mostly in two areas: dynamics and demand-supply side interactions. Specifics concerning the interactions are usually formalised as constraints on parameter ranges and signs. Although probabilistic notions of variables are adopted normally as increasingly dynamically rich models are proposed, derivations of the causal mechanism of interest remain fundamentally deterministic. Take cointegration analysis for example. The theoretical interest is focused on static long-run

[23] Gilboa *et al* (2014) categorise this type of reasoning as case-based reasoning as opposed to rule-based ones. However, their categorisation fails to acknowledge that choices of those cases are still rule based at the outset.

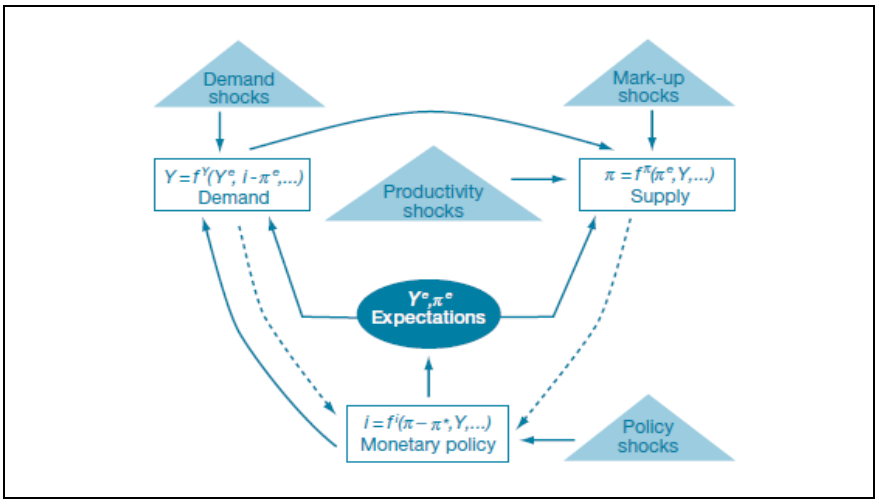
equilibrium solutions. The solutions can be adequately presented by differential/difference equations whose orders are conditioned upon hypothesised trajectories of target variables within the realm of mathematical system theory. In general, the resulting dynamic relations are produced without much concern over stochastic factors.

A case where stochastic factors find an explicit way into theoretical model formulation is the dynamic stochastic general equilibrium (DSGE) model.[24] Arguably, DSGE models are the most theory-rich type of multi-equation models in macroeconomic practice. The hallmark of DSGE models is the adoption of microeconomic notions of optimisation behaviours of rational agents for originally macro structural relations. In terms of individual equation formulation, the commonly accepted form is a linear (log-linear) function, as approximation of certain Euler equations resulting from rule-based optimisation. The idea of target variables being generated stochastically is then realised through the addition of exogeneous shock variables. They are named after the economic types of relations, such as 'demand shocks', 'mark-up shocks' illustrated in Figure 3.1. Such a *structural* veneer plus the latent status of these shocks enables modellers to experiment with various dynamic formations of the shocks to narrow the gap between target variables and those microeconomic rule-based relations. The formations are generally assumed to follow certain AR processes. For example, an AR(1) can be assumed of a supply-side mark-up shock variable explaining π in Figure 3.1:

(3.2.1) $$s_{\pi t} = \rho s_{\pi t-1} + \varepsilon_{\pi t}, \quad 0 < \rho < 1$$

where $\varepsilon_{\pi t}$ is a computer simulated white-noise error term. It should be noted that the use of latent shock variables makes target variables live up to their stochastic label, making the core, *a priori* functionally fixed relations apparently equivalent to models in classical statistics.

Figure 3.1 Basic Structure of DSGE models



Source: Sbordone *et al* (2010)

From the standpoint of multi-equation models, DSGE models can be written as restricted VARs, though their parametric structure is more complexed than that of VAR models. Similar to VARs, DSGE models are mainly used for policy analyses by means of impulse

[24] DSGE models are well described and surveyed in the literature. A good coverage of their conceptualisation can be found from Canova (2009), Sbordone *et al* (2010), Fernández-Villaverde *et al* (2016), Christiano *et al* (2018).

response analysis. Nevertheless, complexity of the parametric structure makes statistical estimation virtually infeasible, especially for those DSGE models disaggregated at relatively high levels. Hence, calibration of parameters is commonly used. A key criterion for calibration is to match the predicted 'macro moments' of target variables to those observed in macro data as closely as possible. The criterion is often termed as identification of 'macro moments' in short, cf. Nakamura and Steinsson (2018). These moments are usually estimated by univariate time-series marginal models discussed in Section 3.1, such as (3.1.2) for *Y* in Figure 3.1.

Unfortunately, the way that probabilistic notions are used in DSGE models is does not stand up for scrutiny. There are at least two discernible conceptual flaws: ungeneralisable estimates of macro moments and over-restrictive dynamics due to the imposed shock structure. The first flaw is already explained in the last section, i.e. how mis-specified univariate AR models are in general. Consequently, parameter calibration based on estimates from such a model type cannot be optimal, as judged by the criterion of ERM. As for the second flaw, the assumed AR processes of shock variables are not innocuous for the dynamic structure of DSGE models. Historically, the assumption goes back to the Cochrane-Orcutt procedure (see Section 5.1 for a more detailed demonstration). The procedure offers an expedient way to tackle residual autocorrelation, a symptom widely found from dynamically inadequately formulated relations. However, the procedure is found to impose a common factor restriction on the dynamics of the original relation, and the restriction is often rejected by data.[25] In a DSGE model, common factor restrictions can lead to substantial bias in its long-run equilibrium solutions. The dynamics of rule-based relations are commonly derived under the adaptive expectation assumption, which corresponds to a partial adjustment model where lagged effects of driving variables are omitted. The omission is frequently shown to be data incongruent, whereas the incongruence can be glossed over by the added shock variables. These defects help explain why increased model sizes in DSGE models, accompanied by the enlarged data information usage, have not resulted in clearcut precision gains in model prediction performance, as compared to VAR models, e.g. see Gürkaynak *et al* (2013).

After all, DSGE models remain conceptually faithful to (3.1.6) as far as individual equation formulation is concerned. All the micro-theory-based relations are of the conditional type. Structural attribution of latent shock variables does not alter the appendant nature of probabilistic reasoning. Observable marginal models serve only for macro descriptive purposes. Haavelmo's vision of economic variables being fundamentally randomly generated remains a loose analogy.

Will the vision come true if economists can carry out empirical studies under certain controlled experiments akin to the required setting of classical statistics? This question brings us to the case of programme evaluation models (PEMs). PEMs are built for the purpose of evaluating the average treatment effect of policy programmes. A key prerequisite of the evaluation is to design and collect data from randomised control trials (RCTs). Controlled randomisation is what classical statistical inference predicated on to ensure sample data

[25] For textbook exposition of the model type evolved from the Cochrane-Orcutt procedure, see e.g. Hendry (1995, Ch7); for a historical account, see Qin (2003, Chs 4 & 7). What should be noted here is the fact that the common factor restriction not only imposes identical short-run and long-run parameters, but also maintains the long-run equilibrium parameters being the same as what is dynamically postulated *a priori*.

having tractable distributions. However, no matter how meticulously RCTs are designed, uncertainty in PEM formulation remains a non-ignorable issue.[26]

Denoting the policy programme under concern by T, we can write a PEM as:

$$y = \alpha T + f_z(z_1, z_2, \cdots, z_m; \beta) + \varepsilon \tag{3.2.2}$$

where α represents the average treatment effect, the key parameter of interest of PEMs, $\{z_j\}$ is a set of other control variables, some of which are likely to be correlated with T. What RCTs try to achieve is to make $E(y|T)$ tractable through sample selection. But $f_z(\cdot)$ is unknown and still needs to be learnt empirically, since its *a priori* uncertainty is beyond the controls. Obviously, estimates of α are predicated on the result of this learning process. Hence, (3.2.2) can only be regarded as a hypothesis class. Moreover, the task of learning the best possible approximation of $f_z(\cdot)$ is a distribution-free one as it does not require prior knowledge of $P_j(z)$. PEM research under RCTs still cannot fully materialise Haavelmo's vision.

In machine learning textbooks, the approach of empirically resolving the decision issue of splitting y in (2.1.2) by filtering out ε is referred to as a *discriminative* approach, as opposed to the *generative* approach. It is only under the latter that parametric density estimation is the focal learning task, e.g. see Jebara (2004) and also Shalev-Shwartz and Ben-David (2014, Ch24). Clearly, classical statistics follows the generative approach. It starts inference from *a priori* known probabilistic models. By taking those models as maintained hypotheses, it becomes justifiable to impose certain ideal noise distributions, such as zero-mean normal distribution, on the residuals. In the case where maintainable models are *a priori* unknown, the immediate learning task becomes searching for model approximations. When the search is fruitful, say by following the setup of Figure 2.2 in Chapter 2, the resulting $h_{\mathcal{D}} \approx f$ leads to the distribution of ε approaching to the normal distribution of random noises. This makes the setting of $h_{\mathcal{D}}$ apparently compatible to a conditional probabilistic model, even though the learning of $h_{\mathcal{D}}$ is executed in a *distribution-free* manner.[27] In other words, statistically learnt models can be isomorphous to stochastic models representing expected conditional distributions.

3.3 Discriminative Probability in Synthesis of Logical Reasoning and PAC Learning

How should we proceed with econometric research if probabilistic reasoning falls short of representing and tackling major uncertain factors in economics? Promising answers can be found from AI researches. In particular, there have been rich AI studies on the issue of formal representation of uncertainty in an introspective manner, with careful codification of information types and characteristics, as well as differentiation of various learning tasks, e.g. see Dubois and Prade (2009), Costa *et al* (2018), also Russell and Norvig (2016, Part III). It transpires that many, if not most, of the cognitive tasks we face in an open and theoryless environment are not directly translatable into probabilistic models. The intelligent reasoning that we usually employ for decision making falls predominantly into the type of modal propositional logic. Knowledge based on this type of logic is epistemically fragmentary in an

[26] For textbook exposition, see e.g. Camero and Trivedi (2005, Chs 25-26); see Deaton and Cartwright (2018) for a critical review of RCTs.

[27] Remarkably, the distribution-free approach is effectively what H. Wold vehemently advocated nearly half a century ago under the label of a 'soft modelling' approach (1975, 1980).

open world. However, it is a cognitive pitfall to confuse our logical reasoning with probabilistic reasoning, because of the fragmentary nature of the former. In the case of conditional model representation, the target variables involved are in fact *uncertain epistemic* variables rather than simply statistically random ones. The corresponding modelling tasks are referred to as learning based on 'genuine' conditioning in Dubois and Prade (2009, Section 6), so that they can be differentiated from tasks based on probabilistic conditioning, such as Bayesian inverse probability reasoning.

From the AI perspective, economists cope with uncertainty representation mainly by the tenets of utility theory, or economic rationality, cf. Doyle (1992). Since economic reasoning falls fundamentally into the modal propositional logic type, its formal representation can be based upon calculus. Causal relations, such as h_p of (3.1.6), are thus formulated without resorting to probability distributions. In other words, the ontological commitment of h_p determines that this conditional relation is not of the generative model type. Correspondingly, the learning goal of (3.1.6) does not fit with that of classical statistics. Noticeably, the AI perspective is affirmative of the scepticism about using notions of probability in uncertainty representation among the history and methodology of economics communities, as mentioned at the beginning of Chapter 1.

It is already argued at length in the previous chapters that, in an open and theoryless environment, the gap between h_p and models targeted by classical statistics is too wide to ignore, and that construction of data-congruent models is a learning issue. To facilitate this learning task, reorientation of research strategies is required from both the economic and the econometric sides. Conventionally, economists are expected to produce *a priori* parametrically interpretable models implicitly relied on their inductive logical reasoning power. Successes and advances of AI-guided machine learning calls for a critical review of this process. Various AI approaches of knowledge representation should be carefully studied and adapted, such as knowledge based inductive learning and inductive logic programming (ILP) system, e.g. see Russell and Norvig (2016, Ch19). These approaches are developed with the aim of making common-sense knowledge representation better suited for machine learning.[28] They are also expected to facilitate knowledge transfer and enhancement 'from machine learning to machine reasoning' (Bottou, 2013).

The core of AI knowledge representations is logical formulation. Logically quantified relational rules are expressive, interpretable, and also easy to concatenate and transfer between systems. But those rules suffer from *brittleness* in a theoryless and complex world, due mainly to their implied global existence or generality over ill-specified or unspecified situations, e.g. see Valiant (2013, Ch7). In other words, 'mathematical logic is an attractive language of description because it has clear semantics and sound proof procedures. However, as a basis for large programmed systems it leads to brittleness because, in practice, consistent usage of the various predicate names throughout a system cannot be guaranteed' (Valiant 2000, p231). Learning from noisy data is a key to resolving the problem. By evaluating and selecting relational rules against data examples, the learning process can robustify logical reasoning against context-specific situation requirements, e.g. see Valiant (2000, 2008). To a

[28] It is interesting to quote T. Sargent here, 'Economics is organized common sense', the graduation address at UC Berkeley in 2007.

great extent, this promotion of 'knowledge infusion', i.e. synthesising logical reasoning and inductive learning, highlights and reinforces the gist of SRM (see section 2.3 of Chapter 2).

It is during the learning process that probability plays a fundamental role – controlling inductive risks in a discriminative manner, as explicitly specified via two parameters of PAC learning theorems: δ and ϵ (see Chapter 2). These probabilistic measures enable practitioners to guard against unguaranteed generality of logical rules while containing inductive errors of empirical learning via experiments to try and optimise the bias-complexity tradeoff. This learning process is labelled as 'the diagnostic paradigm' in Seeger (2006). It is towards this paradigm that the probability approach of econometrics needs to reorient its direction. The need for directional reforms is not just a purely methodological matter. It is arguably the only way to systematically rectify widely discernible inefficiency in applied econometric research. Let us extend this point by taking a brief review of the inefficiency in microeconometric and macroeconometric practices.

Model underfit is pervasive in microeconometric practice, and its severity is on the rise as data samples grow rapidly. The situation reflects lucidly how brittle logical reasoning based microeconomic theories are in an open and complex world. In terms of model specification, the following complications are widely recognised: heterogeneous behaviour of micro agents, unignorable causal factors which are omitted in theoretical relations, uncertain interactions among input variables, and also possible nonlinear effects concerning the input variables of causal interest. Since these complications are essentially issues of function estimation, it is no wonder that parametric-centred estimation techniques taught in microeconometrics textbooks have rather limited capacity in tackling them. Unsurprisingly, models which are empirically trained by machine learning methods can easily outperform those by conventional econometric methods, e.g. see Bajari *et al* (2015). After all, systematic underfit indicates systematically formulated oversimplistic models. In dealing with the problem, machine learning methods advocate for explicit consideration of flexible function approximations. For convex learning tasks, a leading hypothetical model class is the generalised additive model (GAM), cf Hastie *et al* (2009, Chs 5 and 9).[29] For example, a two causal variable GAM can be written as:

$$(3.3.1)\ y = \alpha + f_1\left[\sum_{j=1}^{k} b_{1j}(x_1, x_2)\right] + f_2\left[\sum_{j=1}^{k} b_{2j}(x_1, x_2)\right] + \cdots + f_p\left[\sum_{j=1}^{k} b_{pj}(x_1, x_2)\right] + \varepsilon$$

where $f_i\left[\sum_{j=1}^{k} b_{ij}(x_1, x_2)\right]$ (each f_i may differ from others in terms of its components) is a linear function in basis representations, denoted as $b_{ij}(x_1, x_2)$. These input representations can be the observed variables themselves, their polynomial terms, their ratio and/or product, or certain kernel functions. Design of these representations is commonly referred to as feature design. Feature design is vitally important for theory verification. In order to keep parametric interpretability at the individual input level and also ensure generalisability, careful formulation of all relevant causal rules based on prior knowledge is required during the design. In particular, the formulation should aim at disentangling observed inputs into separate factors of variation. Meanwhile, separately formulated f_i facilitates differentiated representations by strata, and is achievable empirically by means of various decision tree-based classification techniques, which are referred to as semi-parametric techniques in

[29] GAM is not unknown in the econometric circle, e.g. see Cameron and Trivedi (2005, Ch9). Sadly, what is amiss is due understanding and recognition of its statistical learning foundation.

econometrics textbooks. Clearly, p, the number of f_i to be learnt, is sensitive to feature design. Inadequately designed features tend to result in rises in p when function learning is clearly a complexed issuem under large data samples. Hence, the two tasks, tree-based classification and basis function design, have to be carried out iteratively, allowing for a close interplay of data and prior domain knowledge. Therefore, feature design issues are often discussed under the broad topic of 'feature learning', cf. Shalev-Shwartz and Ben-David (2014, Ch25), and studied in detail under 'representation learning' or 'feature representation learning', cf. Goodfellow *et al* (2016). It is only by approaching these learning tasks from the perspective of PAC learning that model underfit can be systematically circumvented.

In macroeconometric practice, a telling case is long-run equilibrium analysis. There are two approaches to the analysis: cointegration analysis and the LSE general-to-specific dynamic specification approach. The former is mainstream and estimation centred whereas the latter akin very much to the spirit of machine learning. Abundant empirical findings show that models by the latter approach are more generalisable and accurate than those by the former. Despite its mathematical elegance, cointegration analysis appears innately susceptible to underfit or overfit in practice. Now, the machine learning perspective offers us a clear explanation to the contrast. Let us start from an error-correction model (ECM), the model class from which cointegration analysis stems originally.[30]

Suppose (3.1.3), an ARDL(1,1) model, is found data-congruent. To make it economically interpretable, we transform it into a simple bivariate ECM:

$$\Delta y_t = \beta_0 \Delta x_t + (\alpha - 1)\left[y - \frac{\beta_0+\beta_1}{1-\alpha} x\right]_{t-1} + \varepsilon_t, \;\; (\alpha - 1) < 0 \tag{3.3.2}$$

$$= \beta_0 \Delta x_t + \gamma [y - \kappa x]_{t-1} + \varepsilon_t, \qquad \kappa = \frac{\beta_0+\beta_1}{1-\alpha}$$

where β_0 represents the effect of short-run shock in x_t, κ represents the effect of the long-run equilibrium relation, $y = \kappa x$, and γ ($\gamma = (\alpha - 1) < 0$) a negative feedback coefficient of the error-correction mechanism around the deviations $[y - \kappa x]_{t-1}$. Input transformation from (3.1.3) into long-run versus short-run features clearly makes (3.3.2) parametrically much more expressive than (3.1.3), because (3.3.2) disentangles as much as possible separate factors of variation in observed data. When it comes to long-run equilibrium analysis, two points are worth noting. First, κ, being a function of all the parameters of (3.1.3), is predicated on the lag-length search of a general ARDL(p_1, p_2). Any estimate of κ based on *a priori* imposed lag lengths can compromise its accuracy. Second, when there are more than one causal variable, disentangle long-run effects of individual causal variables is often difficult to achieve by statistical estimation alone, calibration experiments are thus needed. The experiments involve monitoring estimates of γ as well as parameters of relevant short-run features.

Contegration analysis, on the other hand, is narrowly focused on deducing statistically optimal estimators for κ. It relies implicitly on the assumption of *a priori* known dynamic models. Explicitly, it assumes that the variables involved are all generically non-stationary,

[30] For a brief account of how cointegration analysis has emerged from empirical ECM research and the related LSE dynamic specification approach, see Qin (2013, Ch4); for more introduction to the LSE approach, see Gilbert (1986).

i.e. unit-root processes. The Engle-Granger two-step procedure derives its popularity from the proof that the OLS of k from a static regression:

(3.3.3) $$y_t = kx_t + e_t, \Rightarrow e_{t-1} = [y - kx]_{t-1}$$

is a consistent estimator of κ. A more elaborate estimator for κ, provided in the Johansen procedure, is based on VAR models such as (3.1.4). Both procedures impose homogeneity of the lag structure on the assumed individual nonstationary processes. The imposition cannot be tenable in general since the underlying marginal and joint modelling routes are empirically unreliable, as already discussed in section 3.1. Static models are widely known to underfit data whereas VARs overfit with respect to finite macroeconomic time-series samples. Consequently, estimated κ based on these models usually fall short of optimal extraction of data information. In contrast, the LSE approach gains its advantage not only from the fact that (3.3.2) following the conditional modelling strategy, but also from its model reduction principles, cf. Hendry (1995, Ch9). Specifically, ARDL($p_1, p_2, \cdots, p_n$) models and their transformation into ECMs are not dependant on the unit-root assumption of individual variables, nor on them having identical lag lengths. Since the general-to-specific reduction strategy aims at selection of parsimonious ECMs which best capture the dynamic features of specific data samples, systematic model underfit or overfit has been basically ruled out. It is thus not surprising that the resulting estimated κ outperform those by cointegration analysis.

The LSE approach is widely known for its promotion of model selection diagnostic tests and procedures. Despite its DGP-advocating veneer, the approach shares many key features with the discriminative modelling approach in machine learning. In regard with the empirical achievements of the LSE approach as well as predominant absence of generically probabilistic-rule-based economic models, it is imperative for us to shed the illusive conception of an economy comprising fundamentally jointly distributed variables and to pursue explicitly a distribution-free modelling methodology. Correspondingly, probability should be duly recognised as primarily a discriminative measure – assisting decision making during model selection, rather than used as a rhetoric for endorsing directly performing classical statistical inferences on any *a priori* formulated models.

1. 现实经济的模型抽象

1.1 引子

概率思维是用数据验证理论的认识论基础，基于概率论的统计推断方法是实现验证的必要手段，这便是哈维尔莫力推的概率论方法之大纲。该大纲大致建立在以下三个层面的论断上（详见原著第一章）。首先，经济理论的经验证伪决策，需要由概率尺度来划界；其次，经济理论的经验测度需要用概率手段来实现；另外，经济变量的数据量化本身也存在着不可忽略的不确定性，因此需要把它们视为随机变量。正是由于不确定性的无所不在，随机模型便理所当然地被作为连接理论与数据的“桥梁”，而基于概率论的经典统计学则被视为经济计量学研究的系统化框架。具体地，哈维尔莫将学科的研究构划分为三个部分：理论假说检验、模型参数估计和模型预测，见他原著的第四、五、六章。

有必要指出，在哈维尔莫的大纲之前，经济学家就概率理念是否可充分概述现实中的不确定性问题是持保留和怀疑态度的，最经典的论著有 1921 年发表的两本书：凯恩斯 (J.M. Keynes) 的《概率论专论》和奈特 (F.H. Knight) 的《风险、不确定性与利润》。主要质疑点有二：其一，未来经济的不确定性，这不是依历史数据可充分预测的；其二，量化测度抽象的经济理论概念时的不确定性，这不是概率尺度能够精确构述的。然而，百年以来，虽然有关概率理念之局限性的讨论延绵不休，但是讨论范围基本徘徊于经济学思想史及方法论学界[1]，未能在经济计量学界引起多少关注。即使丁伯根 (Tinbergen) 开创性宏观计量模型研究一问世就受到凯恩斯对模型方法的严重质疑[2]，其威慑作用对于经济计量学的发展来说也是转瞬即逝。经济计量学作为当今经济学中一个基础分支，其主干是依哈维尔莫的概率论大纲而发展成型的。面对分析大量经济数据的实际需求，统计分析工具的广泛应用势在必行，而沿袭凯恩斯质疑的反概率说辩却愈发显得偏泛无济。

还需指出，哈维尔莫为经济计量学研究划界确立了一个前提条件 (见原著第 10 页)：试验性理论模型是先验给定的，而且模型设定能够满足统计实证研究的需要[3]。这一前提条件相当于为理论经济学与经济计量学的研究范围划界，它也将经济思想史学界有关概率理念之局限性的争论屏蔽于经济计量学之外。不过，当提及模型来源的问题时，哈维尔莫坦言：“建模是个创作过程，是一门艺术” (第 10 页)。这相当于认可了建模过程中的不确定因素，并非概率论公理化可构述[4]。这也意味着，计量学研究的前提条件本身，具有尚无系统方法构述的不确定特征。的确，几十年来计量模型研究的大量实践结果表明，先验设定的理论模型通常无法充分满足统计实证研究的需要，存在着哈维尔莫的概率方法大纲所不能涵盖的不确定性。鉴此，对于以建模研究为首任的经济学家来说，哈维尔莫的大纲是既无关又无助的，如参见 Eichenbaum (1995)。事实上，弥补先验建模中的不确定性已经成为应用模型研究者无可脱懈的责任，这种学科研究

[1] 有关这方面更为详细的讨论可参见 Bateman (1990)、 Davidson (1991)、 McCann (1994)以及 Cambridge Journal of Economics 2021 为纪念凯恩斯和奈特专著百年出的专辑：第 45 卷第 5 期。

[2] 凯恩斯的质疑是以对丁伯根 1939 年发表的《商业周期理论的统计检验》的书评形式发表的。Hendry and Morgan (1995) 中的第六部分：“丁伯根之争” 概括了当时有关的方法论讨论原文。

[3] 这里，Koopmans (1950) 一文对模型在什么条件下能够满足实证研究需要做了细致讨论。但是，他总结的条件是以联立方程模型的识别性为重心的，与本节最后一段讨论的有关模型充分性的条件截然不同。

[4] 有关不确定性的行为溯源讨论，可参见 Bradley and Drechsler (2014)。该文从人决策行为角度，对不确定性做了分类。

的“越界”不仅在计量学界内就频频引发有关如何系统处理建模选择问题的争论，而且也不断激发经济方法论研究者对计量学之概率论基础的反思[5]。

从科学哲学的角度看，理论之科学解释力的充分性需要反映在逻辑与经验两个层面，参见 Hempel (1965)。作为理论载体的经济模型，其逻辑层面的充分性是依靠数学演绎推理来实现的。应用计量模型研究者需要广泛参与模型选择和修订的事实，也揭示了模型解释力不足的主要原因，即先验模型设定步骤本身缺失实现经验层面上充分性的手段[6]。也就是说，先验设定模型的不确定性，主要源于建模者无法将经验信息充分考虑到建模过程中来，因为这种不确定性的量化并不能笼统地通过概率公理化手段来演绎实现。为了弄清量化这种不确定性的难点所在，我们有必要对利用模型手段将理论信息与数据信息对接时的不确定因素做缜密剖析。

1.2 量化经济变量过程中的不确定性

经济变量的测度不确定性是哈维尔莫概率论方法的最基础论据之一。为了引入随机测度误差的理念来解释测度不确定性，并且证明该理念的正确和必要性，他对经济变量做了三种划分：可测变量、真实变量、理论变量。然而，依他的分类仔细推敲量化变量中的不确定性与概率论方法的关系，我们就能发现，前者之特征或形式要比基于概率分布形式的随机误差概念更加隐晦复杂，在不少场合下，并不是能用概率统计推断手段来有效构述和处理的。量化变量中的不确定性使得概率方法成为系统构述变量测度随机特征的必用之法，这一逻辑诉求其实是皮相之谈。

让我们从可测经济变量入手。以可测性划分，大多数的经济变量的定义直接对应于量化测度，如收入、消费、产出、成本、利润及价格等等，无直接测度的变量属于相对少数，如风险、效用、能力等等。不过，经济学所关注的可测变量大都隐含一定程度的广义性。哈维尔莫把广义性抽象至极的变量称为“真实变量”。因此在经济统计数据中，真实变量一般没有一对一的量化对应物。以收入概念为例：从规模层面看，宏观上有国民总收入、国内生产总值 (GNP)、国民生产总值 (GDP)，微观上有家庭或个人总收入、最终收入、税后收入、可支配收入等等；从时间层面上则有月收入、年收入、固定收入及非固定收入等等。价格概念的对应指标则更为繁杂多样。从规模层面看，宏观上有消费者价格指数、固定资产投资价格指数、进口和出口价格指数、金融市场股票和债券总指数等等，微观上有单个商品的生产者供给价格、销售批发和零售价格以及平均价格指数，劳动力市场中有各种工资指数，金融市场中有各种有价证券的价格等等；从时间层面看，价格数据也比收入数据更丰富密集，微观价格上来自商品市场的有日价，而来自金融市场的有逐笔交易价。显而易见，抽象的真实变量与其相对应的各种现有统计数据指标之间的量化间隙以及测度之不确定性范围，要远远超出经典统计推断分析考察的范围。

实际上，上述量化间隙中很大一部分的不确定性并不在计量学研究的视野内。这主要是由两方面原因综合所致：数据来源与变量用途。依统计数据来源分类，绝大多数经济数据属于间接的二手数据，由政府专职部门、金融市场及国际机构组织收集发布。这类数据不同于统计学一般要求的由研究者直接设计采集、专题专用的一手数据。二手数据的经济指标反映真实变量的精度十分有限（亦即这里的精度误差远超出由概率尺度的量化范围），这属于公认实情。即使就同一指标而言，统计方法不同得出的

[5] 有关经济计量学界关于建模性质评价的讨论综述，可参见 Qin (2013, Ch9)；至于经济学方法论界对经济计量学基础的反思，可参见 Rowley and Hamouda (1987) 及 Stanley (1998)。

[6] 有关这一结论的正规系统论著可参见 Stigum (2003, Part III)。

结果也会有明显差异。例如，对于一个国家的总收入来说，最常用的国内生产总值有两个统计口径：从最终支出口径统计的指标，及从三个产业生产口径统计的指标，这两个 GDP 指标常常有明显差异。至于如何增进统计指标反映真实变量的精度，这属于数据发表部门职责内的问题。在大多数应用计量模型研究中，只要所发布的指标之统计口径不变，这些指标相对于真实变量的测不准性是被忽略不计的。这主要是因为对于经济政策研究和实施者来说，综合判断大量二手数据中的有用信息是他们深入了解现实经济的一项任务。他们关注的经济变量，通常只能靠正式发表的相应指标来考察。他们的关注所在，也只是模型对特定指标在一定的规模和有限时间范围内的预测，而不是对指标所对应的真实变量的无界推断。具有明确的场景针对性及可测指标的针对性，其实是经济计量模型结果获取实用价值的必备前提。

那么在什么情形下，变量的测度误差问题是属于经济计量学需要考虑的不确定性呢？当模型研究任务是解决变量在模型外不可测的问题时，就显然属于这种情形。这时的关注变量被视为潜变量。有必要指出，经济学家对潜变量的关注反映着以可测性为标签划界变量的模糊性。仍以价格变量为例。虽然价格这一概念本身直接对应量化测度，但是在不同的加总需求下，价格指数的构造仍然成为需要由经济计量模型研究的潜变量。例如在微观经济学中，不少商品的品牌、型号及质量多种多样，如汽车、房产、各种电子产品等等，要综合测度这类商品的价格指数，就需要把商品中体现单个产品质地差异造成的价格差异部分剔除掉。于是便产生了过滤单个产品特质效用的商品价格测度模型 (hedonic price model (HPM))。

对于任何抽象定义的潜变量来说，它由某个测度模型生成的潜变量指标之间也不具有一对一的关系。这时，若用随机误差来描述由不同模型生成的指标与原定义潜变量间的差异，不同指标之间的差异就应该体现在随机误差大小及分布性质的差异之上，我们就可以把寻求随机误差最小化、纯噪声化作为筛选指标的目标。不过，由于潜变量之不可测的本质，上述筛选目标一般通过衡量指标是否基本具有我们对潜变量预期的特性来实现。显然，指标的筛选过程其实就是对生成指标的不同测度模型的筛选过程。而不同模型之间存在着无可避免、且非平凡（亦即非纯噪声可以刻画的）的不确定因素。那么，使用概率论方法是否能够有效地处理这种不确定因素呢？寻找问题的解答又把我们带回了上一节末的归论，即我们需要认真剖析采用先验模型与数据信息对接时的不确定因素之源。

从哈维尔莫所定义的变量分类视角看，上述讨论与‘理论变量’的关系密切。模型其实就是理论假说的试验载体，而测度上述潜变量的指标又是模型模拟生成的，我们就应可把这些指标归于理论变量一类。不过，这一归类难免会引起我们对理论变量与真实变量间的本质区别，及其该分类的现实作用的质疑。毕竟，真实变量也是一个纯抽象概念，与现实观测到的变量之间不具有一一对应的关系。要想解除质疑，并在学科内对变量分类方式达成某种共识并非易事。相对来说，由上述讨论较易达成的共识是，从建模方法论的角度看，有关测度模型的选择问题与非测度模型的选择问题并无本质区别。这意味着，我们可以通过分析研究者所选的模型之构成与设定的不确定性问题，来考察计量研究中有关模型所测度的变量之不确定问题。

1.3 理论模型与经济现实

让我们首先从上节例举的商品效用特征的价格测度模型 (HPM)入手。假设我们有 $T+1$ 个时段$(t = 0,1,\cdots,T)$上的横截面样本，每个样本包含某类商品中单个产品的出售价格 p_i 及主要特质信息 x_{ij}，$\left(p_i\middle|x_{ij}; i = 1,\cdots,n; j = 1,\cdots,J\right)$，而且样本量充足。为了

便于分析，我们这里避繁就简，以最常用的毗连时段虚拟变量模型形式为例[7]。将毗连两个时段的横截面样本整合为一个混合截面样本，做下述回归估计：

(1.3.1) $$\ln\left(p_i^{t,t+1}\right) = \beta_0 + f\left(z_{ik}^{t,t+1}, \beta_k\right) + \alpha^{t+1} D_i + \varepsilon_i^{t,t+1}$$

式中的z_{ik} $(k = 1, \cdots, K)$是基于特质x_{ij}上的单个产品质量变量，例如可以选择定义：$z_{ik} = x_{i1}$、或者 $z_{ik} = ln(x_{i1})$、或者 $z_{ik} = x_{i1}x_{i2}$、或者 $z_{ik} = x_{i1}^2$等等。D_i是时段虚拟变量，在 t 时，$D_i = 0$，在$t+1$时，$D_i = 1$。参数α^{t+1}表示在剔除单个产品特征效应后，该类商品从 t时到$t+1$时的纯价格变动，因此是本模型的关注参数。该类商品的时序价格指数可由 T 个混合截面样本估计来的该参数之幂指数时序 $\left\{e^{\hat{\alpha}^{t+j}}\right\}$算出。显然，基于产品特征效用理论所构造的不同模型，即$f\left(z_{ik}^{t,t+1}, \beta_k\right)$的不同设定，会生成不同的$\alpha^{t+1}$估值。实践中，从上述模型反映出的常见问题有二：残差项$\varepsilon_i^{t,t+1}$不满足统计学期待的白噪声性质的问题和模型参数估值$\hat{\beta}_k$缺乏时不变性的问题。二者均是先验选定的$f\left(z_{ik}^{t,t+1}, \beta_k\right)$存在经验解释力不足的问题之表象。应用建模者对上述两个问题的诊断和处理大致可归为两方面。一是消费者对产品某个特质的支付需求超出简单线性关系的描述范围，因此需要根据样本信息对单个$z_{ik}^{t,t+1}$做出仔细认真的设计筛选；二是数据样本涵盖的消费者需求差异超出单个连续函数的构述范围，需要考虑剔除样本中的异常值离群点，或者考虑采用广义加性模型形式。两种处理途径都需要数据试验，而处理的结果都会影响α^{t+1}的估值。值得注意的是，在处理上述建模不确定性的试验过程中，基于概率论的统计推断工具的作用是有限的。而用统计推断工具对α^{t+1}的估值分析之可信性却需要以上述后验建模试验得到满意结果为前提条件。

诚然，HPM 模型不是经济计量学的典型案例，潜变量的模型测度问题也不为大多数应用建模者所关注。验证某种先验假设的行为因果关系，这才是经济计量模型研究的基本主线。不过，在这类模型的研究中，上例所述的通过数据试验来修正选择模型的情形处处可见。是什么原因造成这种局面的呢？让我们来分别考察分析横截面样本和时序样本的两类行为模型之基本特征[8]，以探讨问题的基本答案。

分析横截面数据的模型以微观模型为主。微观模型基于的先验理论大都为局部的行为因果假说关系，如 $x_1 \to y$。现将相应的容量为n的横截面样本记为$\left(y_i \middle| x_{ij}; i = 1, \cdots, n; j = 1, \cdots, J\right)$，其中的变量集 $(x_{i2}, \cdots, x_{iJ})$ 包含其他可能作用于被解释变量的自变量，通称为控制变量。这时就有类似于(1.3.1)的回归模型：

(1.3.2) $$y_i = \beta_0 + f(x_{i1}, z_{ik}; \beta_k) + \varepsilon_i$$

式中的 z_{ik} $(k = 1, \cdots, K)$ 表示基于x_{ij} 的特征变量，其设计原理类似于 HPM 一例。这里唯一不同的是，由先验理论决定的关注参数是定义在函数$f(x_{i1}, z_{ik}; \beta_k)$之内的。实践中，横截面模型之拟合度通常较低，意味着数据信息大都被滤出模型。残差项ε_i不满足统计学期待的白噪声性质是普遍存在的现象，这一现象也基本被视为常态。一个通用的解释是个体行为者的奇异性，施加虚拟变量是控制该奇异性的主要模型途径，如固定效应和随机效应方法。不过，这些方法对于模型拟合度的改善效应并不十分明

[7] Griliches (1961) 一文是公认 HPM 模型研究的开拓篇；有关 HPM 研究的综述，可参见 Triplett (2004) 及 Hill (2013)。

[8] 横截面样本和时序样本的延申是版面数据样本。现有分析版面样本的模型技术都是从前两者延申而来的，亦即它们不是以时序分析技术为主线的延申便是以横截面分析技术为主线的延申。因此，这里就不考虑分析版面数据样本类的模型了。

显。确保关注参数估值的可信度，通常是微观模型研究的重心。为此，式(1.3.2)函数中的变量z_{ik}选取问题则是必须考虑的，且只有后验数据试验才能加以解决。考察和试验的一个目标，是从先验关注的因果关系出发，排除遗漏变量有偏性的风险。不过，该风险的规避并不能绕开模型拟合度普遍偏低的障碍。先验理论一般被当作担保关注参数估值之广义性的坚实依据，而参数估值是否的确具有不变性的问题却几乎无人问津。

在分析时序样本的宏观模型研究中，出于对模型预测性能的考虑，参数估值的不变性是一项普受关注的问题。传统的宏观计量模型的基石是静态一般均衡理论。由于模型预测的需求驱导，当今的宏观模型都是建立在动态理论基础之上的，如基于合理预期假说的模型。向量自回归模型 (vector autoregression (VAR)) 则是宏观计量模型研究中最常见的形式。以如下的开放式 VAR 为例：

(1.3.3) $$Y_t = A_0 + \sum_{i=1}^{l} A_i Y_{t-i} + \sum_{j=0}^{l} B_j X_{t-j} + \boldsymbol{\varepsilon}_t$$

式中的 Y_t 为被解释变量向量集， X_t 为其他有关变量向量集， A_i和 B_j 为参数矩阵[9]。与分析横截面样本的微观计量模型相比，VAR 模型的拟合度往往要高很多，残差向量 $\boldsymbol{\varepsilon}_t$也往往基本满足统计学期待的白噪声性质。导致这种现象的主要原因是相对小的样本容量与模型被解释变量表现出的显著时滞惯性。宏观计量模型从静态走向动态化，就是需要充分刻画这种惯性的结果。然而，通过数据试验来修正模型的情形仍是无处不见。这主要体现在两个方面：对于 VAR 的滞后阶数 l 的选择与对于X_t集的选择。由于 l 引致的过度参数化是 VAR 模型的一个薄弱环节，它不仅受到样本容量的局限，而且影响模型内含的静态均衡关系的参数估值。动态计量模型中，与静态均衡关系的参数相对应的参数为长期参数，它们是确保模型经济理论解释力的关键参数，而它们的估值与 l 密切相关。至于X_t集的选择问题，时序样本大都包含具体经济体制上的某些特质信息。这些样本所含的特征是被理论抽象掉的，却是经济预测所不能忽略的因素。宏观经济经历的“体制冲击”或者“体制转型”会导致模型预测的系统失误。这种失误就意味着模型中某些参数丧失了不变性。而且在体制冲击面前，长期参数之不变性的承受力往往要低于短期参数之不变性。

上述分析表明，在数据面前，先验理论模型普遍暴露出解释力之充分性上的不确定性。因此，参照数据信息修正模型这一步骤不仅无法规避，而且是实现理论假说统计推断检验的必要前提。主要需要修正的是两个互相关联的方面：(a) 输入变量筛选。无论先验理论模型的构造多复杂精细，都难免忽略了某些与被解释变量有关的导因，特别是变动观测的数据样本所反映的属于特定经济体的导因。只有通过对数据的后验分析，我们才能判别和解决这些被遗漏的导因是否会影响建模者所关注的因果假说之验证的问题。(b) 输入变量的设定形式。经济理论中的许多因果假说都与所涉变量的规模特征有关。如收入和消费这类流量型变量、以及储蓄和库存这类存量型变量，它们在多元回归模型中缺乏统计学所期待的解释变量间的互不相关性。这也是建模者面对理论假说检验任务时担忧遗漏变量有偏性风险的一个主要原因。因此，仔细设计输入变量形式，参照数据信息尽量避免变量间的相关性，以提高相应参数的个体解释力，是处理上述变量规模特征的唯一途径。动态模型研究中对变量长、短期的区分设计，便是为实现此目标的一个范例。 另外，在采用由开放世界被动观测得来的样本拟合的模型做统计推断时，我们还需意识到推断“总体”的不确定性。只有澄清模型在什么

[9] Sargent and Sims (1977) 一文是公认的 VAR 模型最初的系统倡导性研究；至于 VAR 模型的发展史，可参见 Qin (2013, Ch 3)。

情况下具有满足泛化的经验规律性，以明确模型适用的总体界限，我们才能确立计量模型的实用价值所在。由于经典统计学是设计在推断总体明确已知的前提基础上的，这一定界任务是无法单靠统计推断工具来完成的。综上所述，为了测度和验证经济假说理论，我们首先需要参考数据信息、尽量排除上述各不确定因素，以设计和选择出既嵌套关注理论假说又被数据接纳的模型。这一建模过程可以由理论驱导，也可以由数据驱导，但必须是一个理论信息和数据信息结合使用的过程。建模过程中各不确定因素的处理涉及一系列的选择决策，其决策范围远远超出经典统计学的框架[10]。反思起来，这足以诠释哈维尔莫为什么将该过程概述为“一门艺术”了。

哈维尔莫虽然把建模任务分离出经济计量学的研究范围，但他对该任务的重要性一清二楚。他用一个整章的篇幅描述了构建的模型须具有的基本性质，即模型中假设的行为因果关系必须具有“不变性”、“简洁性”和“自律性”（见原著第二章）。在单一依靠先验理论建模的模式下，严格的数学演绎推导被普遍视为确保模型满足这些基本性质的手段和途径。但是，大量应用结果业已表明，这一信念犹如痴人说梦，经济计量学研究者必须面对先验模型欠缺上述性质给本学科带来的麻烦[11]。近年来崛起的机器学习理论，首次将上述性质作为建模的原则标准归入统计学习的体系，将如何参考数据信息来建模的“艺术”过程转换成一个有序的科学归纳学习过程。鉴此，我们在第二章概述和讨论有关统计建模的机器学习基本理念和原理，强调了概率近似正确学习（probably approximately correct (PAC) learning）、结构风险最小化准则（structural risk minimisation）等基础概念。第二章表明，在以探索构建具有泛化性的模型目标下，要想从对样本信息的有效系统归纳来学习得到吻合数据特征的模型，就必须摆脱经典统计学框架的束缚，根据建模中决策的具体需求来选择恰当的数学工具[12]。

一旦基于机器学习理论的建模步骤被正式归入经济计量学的研究范畴，学科的疆界就大为扩展，研究问题的主次结构就面临重组。为此，我们需要重新审度哈维尔莫的概率论方法构架。后续章节就是朝这个方向的探索尝试。在第三章，通过反思和考察概率测度在计量学研究中的基本功能，我们不难发现，成功研发的大多数应用计量模型其实并没有、也不需要以学界公认的所有变量之联立概率分布为基础。从理论角度看，绝大多数经济因果假说的先验模型构述属于确定性数学逻辑推理过程，并不涉及分布函数。相应地，来自开放世界场景、并确有实用价值的计量模型大都属于辨别式，而不属于生成式。因此在建模学习过程中，概率论的主要基本用途就是辅助先验知识和后验数据信息的综合归纳辨别决策。继后的三章继续沿着哈维尔莫一书的结构，进一步分析探讨概率论的这种辅助用途在假设检验、参数估计和预测三个层面的体现特征。

第四章的讨论表明，统计假设检验的框架是远不足以忠实构述大多数经济理论所关注的假说检验问题的，对这些问题的模型构述有赖于依据机器学习原理。模型学习的需求又使统计假设检验的诊断性功能上升到先于其推断性功能的位置。另外，本章还简述了基于概率测度的诊断性检验工具的局限性。第五章转而分析主流计量学以参

[10] 值得一提的是，在 Kardaun *et al* (2003) 一文中，建模过程中不确定因素的决策问题被统计学家们描述为含义隐晦的问题。

[11] 有关经济计量学史上对于建模问题的方法论不休的争论概述，可参见 Qin (2013, Ch 9)；从科学哲学方法论角度探讨理论模型与经济现实关系的文献，可参考 Mäki (2002) 和 Rodrik (2015)。

[12] 有关机器学习中的统计学习方法与经典统计学方法的认识分歧的讨论，可参见 Breiman (2001) 和 Efron (2011)。

数估计为中心的研究策略的认知缺陷，指出估计法在模型学习过程中的首要功能也是辅助模型选择决策。只有在学习到了既满足机器学习倡导的结构风险最小化准则、又具理论解释力的模型之后，参数估计法的推断性质才进入议事日程。本章简述了有关模型选择之后的估计推断问题。第六章着重讨论了在模型学习过程中模型预测功能的关键性，该功能亦即模型的泛化力大小及其疆界。鉴此，概率测度在经济计量学中的最基本用途必须被重新定位，从经典统计推断的主线转移到为 PAC 可学性服务的统计决策工具之上。

2. 经济关系式的可学性

在讨论经济计量学的技术问题之前，哈维尔莫将第二章的整个篇幅用于讨论经济学家所关注的基本问题：“我们是否有任何希望建立起合理的模型，来增进我们对现实经济生活的认识”(见原著第 11 页)。回瞻易见，在这章中，哈维尔莫的不少思想从未被公理化融入经济计量学，但却与机器学习的原理不谋而合。他对建模的最基本要求是：模型必须体现‘经济规律的持久程度’(原著第 12 页)。这一要求与机器学习对模型泛化性的追求密切相关。哈维尔莫所描述的经济学家之被动观测者的处境，以及他所强调的‘实际试验设计’(原著第 13 页)建模理念，与机器学习中的统计数据分析基本场景如出一辙。

本章旨在引入机器学习理论来更新哈维尔莫的建模思想。第 2.1 节指出，从哈维尔莫的第二章中列举的建模所需处理的多种难题看，他所描述的建模实质上是一个机器学习问题。他对模型需有‘经济关系的逆转性’(原著第 17 页)的要求，正是机器学习算法所力求实现的。这些算法的设计主旨，是通过对现实数据生成机制进行逆转性设计，来学习选择出最吻合数据结构和规范的应用模型。第 2.2 节集中介绍机器学习中可学性理论的要旨，并且讨论可学性理论对于经济计量学的适用性。最基础的可学性理论是以一致性学习问题为出发点的，可学性定义在概率近似正确学习（probably approximately correct (PAC)）的概念上，并以经验风险最小化(empirical risk minimisation (ERM)) 作为基本优化准则，以权衡选择模型泛化的偏差-复杂度为基本理念。第 2.3 节简述有关非一致性学习问题的可学性理论，该理论是对 PAC 可学性理论的扩展，其基本准则为结构风险最小化 (structural risk minimisation (SRM))，其他的准则包括奥卡姆剃刀定律（Occam’s Razor）、模型的一致性和稳定性。这时，模型选择的偏差-复杂度权衡理念，也等价于拟合-稳定性的权衡理念。显而易见，哈维尔莫所强调的经济规律表达式的‘简化性’和‘自律性’[13]，终于在可学性理论体系中得到了精炼严谨表述，成为经验模型可学性的一部分。最后的 2.4 节综述机器学习问题的分类，并讨论这些分类对经济计量学建模问题的启迪。

有必要指出，哈维尔莫所关注的一个要题是联立模型的逆转性问题。不过，过去数十年的大量经济计量学研究业已表明，经济变量间复杂的动态相关关系体系，远远超出静态联立模型所能表述的范围。在任何多方程动态模型体系中，如何降低其中每个单一经济关系式设定中的不确定性，实现其最优泛化能力，仍然是建模研究的最基本问题。因此，这一基本问题便构成本章的重心。有关由静态联立模型欠逆转性而引致的内生有偏性估计解法，我们在第五章再做讨论。

[13] 自律性这一概念是由弗里希首倡的，有关这一概念的历史回顾，可参见 Qin (2014)。

2.1 经济关系式的构建应属统计学习问题

为了检验经济理论假说，在将假说转换为数学关系式时，关系式必须满足一定的性质条件。哈维尔莫在他的第二章详尽讨论了三条必备性质：在开放和多变现实中的常定或持久性、由被动数据的可观测性（亦即逆转性）、以及结构的高度自律或不变性。哈维尔莫对构建具有这些性质的关系式的可能性充满信心，但他也深知实现目的的难度。他讨论的难点大致为下述几个方面： (1) 先验理论通常以理性经济行为作为假设前提。在现实中，这一假设对于个体行为者来说往往过于简单笼统。(2) 经济理论往往集中关注某几个特定变量间的因果关系，此时通常假定其它情况均保持不变，作为忽略其他因素的理由。然而，在开放场景被动观测到的数据面前，这一假定条件是站不住脚的。(3) 当理论关注的因果关系中，多个输入变量间存在显著相关时，很难从被动观测数据信息中将单个输入变量的影响分离出来。(4) 现实中的政策及体制变动属于常态，个体行为者时常需要面临和适应新 '环境' (原著第 p.20 页)。因此，样本总体的不确定性、甚至是总体的变动，都是我们分析由开放场景被动观测到的数据时不能回避的问题。

上述难点所共同反映的，正是我们在上章指出的先验模型构建不可避免的经验不确定性。另外，哈维尔莫对经济关系式的性质要求，也不是纯先验模型构建所能确保的，必须经过后验实验来甄别。这表明，哈维尔莫对实证性模型构建研究的最基本标准要求，其实正是上章指出的模型设定在经验层面上的充分性。用哈维尔莫的话来说，'实验设计 … 是任何量化理论的一个核心附件'。为了强调实验设计的重要性，他还引用了罗素的名言："实际的科学过程是观测、假设、实验和理论交替的进程" (原著第 14 页)。如今看来，他的'实验设计'其实预示着基于计算机试验的统计学习方法，通常简称为机器学习方法。机器学习是专为分析开放世界中的未知数据机制而发展兴起的学科。这里值得一提的是，在 Valiant 详细解说 PAC 理论构思的一书 (2013, Ch.1) 中，他把机器学习所针对的场景称为*理论贫乏* (*theoryless*) 场景，以区别于那些科学知识相对充足的*理论丰富* (*theoryful*) 场景，例如物理学所研究的场景。当我们面临理论贫乏类问题时，通用的处理理论丰富类问题的严格数学形式演绎推理手段往往效力虚乏，现有的常识性知识仍旧是我们选择决策、处理问题的主要依赖手段。必须看到，在常识性知识的形成过程中，起着决定性作用的一环便是人类对案例的归纳学习。机器学习的主旨便是模拟人脑的这种学习认知功能。

针对在理论贫乏场景条件下的模型构建是机器学习的一项核心任务，该任务有时也被称为"函数估计"，如见 Goodfellow *et al* (2016, Ch.5)。在模型学习过程中，研究者对模型关系式的性质要求被转换为选择模型的准则，以便计算机协助研究者搜寻既具有数据相合性质、又能满足他们研究目的的模型。在当今的计算机时代，哈维尔莫所要求的'实验设计'已经成为机器学习的日常工作部分。下面我们就来简述一下机器学习对模型构建学习任务的基本构架。

将经济学的先验假说关注的因果关系记为$X \to y$，其中的 X 表示因果关系中的自变量集， y 表示因变量。对$X \to y$施用经济学通用的优化准则，一般就得出某函数式：

(2.1.1) $$y = h_p(x_1, x_2, \cdots, x_k; \kappa)$$

使得因果关系由参数集κ测度表出。由于优化准则通常被转化为数学函数的凸性，相应得出的h_p就是κ的线性函数。不过，h_p一般不具数据相合性。现将具有数据相合性的正确关系式记为：

(2.1.2) $$y = f(x_1, x_2, \cdots, x_k, z_1, \cdots, z_m; \beta) + \varepsilon$$

式中的 $(z_1, \cdots, z_m)$ 是被理论式(2.1.1)抽象省略掉的变量集，$\kappa \subset \beta$，ε则为可被忽略不计的噪声或误差项。显然，函数f 是先验未知的。而由于ε的存在，该函数的后验可知性也不确凿。这时，研究者的学习目标就是f、是寻求该目标函数的最优近似函数 $h_d \approx f$。完成任务的手段是设计学习算法。算法应包括所有有关可行关系式的准则、以及数据相合的条件等等。例如，哈维尔莫提出的关系式参数的常数性、式中围绕理论关注参数κ的结构不变性。不难看到，学习任务$h_d \approx f$的设定本身已经体现了哈维尔莫要求的‘经济关系式的逆转性’了。而他例举的难点则意味着一般有：$h_p \not\approx h_d$。

机器学习对统计学习任务的构架，实质上是以下述假设为前提的：由数学解析无法达到数据相合关系式的绝境，可以通过经验解的途径来克服。那么，这一可行性假设前提是否实际可信？采用归纳学习法得来的关系式这条路径到底有多么可靠？答案便在由计算数学而生的可学性理论之中。

2.2 体现逆转性的 PAC 可学性

让我们先来简述一下机器学习中可学性理论的基本要旨。该理论已经历了半个世纪的发展走向成熟，在不少机器学习教科书中都有系统的介绍[14]。

沿用上节的学习任务构架，即设学习目的是利用数据集$\mathcal{D}$寻求未知目标函数 $f: \mathcal{X} \to \mathcal{Y}$的最优近似函数，如见 Friedman (1994; 1997)。同时，根据先验知识，我们选定一个包含所有可能的函数之假说类$\mathcal{H}$。为了通过$\mathcal{D}$来从$\mathcal{H}$中选出尽可能最优的$h_{\mathcal{D}} \approx f$，我们需要在数据算法中设定最优判别准则。这里，最为直接明显的准则是经验风险最小化 (ERM)。选择与 ERM 相应的损失函数ℓ，求得函数：$h_{\mathcal{D}} = argmin_{h \in \mathcal{H}}\{E_{in}\}$，其中的$E_{in}$ 表示样本内误差，即 $E_{in} = \mathrm{E}[\ell(h)]$，$argmin\{\}$ 表示使目标函数取最小值的变量值。

由于$h_{\mathcal{D}}$的泛化性才是学习任务的最终目标，在完成了上述计算之后，我们还需要考察样本外的误差项E_{out}，将它与E_{in}作比较。这里，为了明确定义函数模型的可学性，机器学习理论便引入了 PAC、即概率近似正确学习的概念。将$h_{\mathcal{D}}$的样本外预测精度用参数ϵ表示。当以下概率水平δ相对小在可接受的范围内时，若有：

(2.2.1) $$P[|E_{out} - E_{in}| > \epsilon] \leq \delta$$

我们则称$h_{\mathcal{D}}$为 PAC 可学的。阈参数δ一般被称为置信参数。更精确地说，(2.2.1)表示，$h_{\mathcal{D}}$ 是在$1 - \delta$ 的置信水平下 PAC 可学的。值得注意的是，上述 PAC 可学性定义在两个近似参数$\epsilon, \delta \in (0,1)$之上，定义中并不含对数据集$\mathcal{D}$的任何概率分布假设[15]。

---

[14] PAC 可学性理论是 Viliant (1984)首创的；有关统计机器学习理论的框架及权威性综述，可参见 Vapnik (1999, 2003)。有关可学性理论的严格推导和详细讲解，可参见 Shalev-Shwartz and Ben-David (2014, Part I)；在 Abu-Mostafa et al (2012, Chapters 1 & 2) 和 Russell and Norvig (2016, Part V)两本教科书中，则有关于可学性理论更为浅显的讲解。

[15] 在 Shalev-Shwartz and Ben-David (2014, Ch. 4) 书中，不含任何规律分布假设条件的 PAC 学习理论，被更确切地定义为不可知(*agonistic*) PAC 可学性。因此，从认知角度看，PAC 学习理论是‘尽量不附加实质性假定条件、而不去约束实体对象’的理论，Valiant (2008)。另外值得一提的是，Doyle (1992) 从经济学理性原则视角把 PAC 学习解释为‘对概念的理性逼近’。

显然，为了识别和认识在较高置信度下能够学习得到$h_{\mathcal{D}}$的场景，必须仔细考察泛化误差项$|E_{out}-E_{in}|$的大小与特性。考察的关注点大都在由 ERM 准则下选择的$h_{\mathcal{D}}$所得的预测误差E_{out}的收敛条件。式 (2.2.1) 隐含着泛化界限：$E_{out}\leq E_{in}+\epsilon$。于是便有了以泛化界限为出发点的有关$E_{out}$收敛条件的各种定理，例如教科书中通常介绍的 Hoeffding 不等式、 VC (Vapnik-Chervonenkis) 维度、 Rademacher 复杂性等等。这些泛化界限大致由下式来概括：

$$E_{out}\leq E_{in}+\mathrm{O}\left(\sqrt{\frac{d}{N}lnN}\right), \tag{2.2.2}$$

式中的N 表示数据集$\mathcal{D}$的样本容量， d为反映$\mathcal{H}$复杂程度的维度参数， O 为通用的函数渐近符号，反映该项的收敛有界性。这里，括号内的表达式最为关键，它向我们揭示了泛化程度，亦即可学性，对$\mathcal{H}$ 之复杂度和 $\mathcal{D}$之容量的密切依赖关系。为了控制泛化中的误差界限不扩大，若$\mathcal{H}$越复杂，则所需的数据样本容量就越大。在给定了$\mathcal{H}$ 的复杂度情形下，若再设定参数 $\epsilon,\delta\in(0,1)$的值，我们就能根据泛化界限定理解出相应的N值，该解被通称为‘样本的复杂度’。

由于$h_{\mathcal{D}}$是未知f的近似函数，即有$\delta>0$，在非试验开放场景中，$h_{\mathcal{D}}$的误设程度便是机器学习理论关注的重心。通用的分析理念是，从f出发将E_{out}的期望值分解为：

$$\mathbb{E}[E_{out}]=[\mathbb{E}(h_{\mathcal{D}})-f]^2+\mathbb{E}\left[\left(h_{\mathcal{D}}-\mathbb{E}(h_{\mathcal{D}})\right)^2\right]=\epsilon_{app}+\epsilon_{est} \tag{2.2.3}$$

上式中的ϵ_{app}表示近似误差，ϵ_{est}表示估计误差。从分解式可看到，为了实现泛化目标，建模者需要在ϵ_{app}和ϵ_{est}间做出权衡决策。由于ϵ_{app}测度的是近似函数距f的偏离，因此通常被视作学习函数的偏差。显而易见，ϵ_{app}随着$\mathcal{H}$复杂度的增大而缩小。ϵ_{est}测度的则是在选定$h_{\mathcal{D}}$下的经验风险，该风险通常随$\mathcal{H}$复杂度的增加而上升，并随样本容量 N 的增大而减小。由于f 是未知的， ϵ_{app} 只是一个纯理论测度，现实中对最优泛化目标的追求主要靠监测ϵ_{est}来实现，即通过监测在不同复杂度模型下的ϵ_{est}的变动情况，来做出权衡决策。在机器学习文献中，通称这一权衡为偏差-方差权衡、或者偏差-复杂度权衡。

为了通过监测ϵ_{est}而选择模型，应用中通常把已有数据集$\mathcal{D}$分为两部分[16]：训练子集和测试子集，记为$\mathcal{D}=\mathcal{D}_{train}\cup\mathcal{D}_{validation}$，且有 $\mathcal{D}_{train}\cap\mathcal{D}_{validation}=\emptyset$。这样，我们就能通过$\mathcal{D}_{train}$获得$E_{in}$，并通过$\mathcal{D}_{validation}$获得$E_{validation}$，作为$E_{out}$的模拟。这两种误差都是通过综合$\mathcal{H}$ 和 ℓ的编程算法$\mathcal{A}$而计算的。在机器学习中，$\mathcal{A}$的设计十分关键。在人工智能研究的影响下，一些机器学习研究者把人工智能文献中给神经网络起的别名‘感知器’(*perceptron*) 引用到算法上。因此教科书中$\mathcal{A}$ 也常被称为感知器学习算法 (PLA)。

机器学习对于ϵ_{app}的关注和对偏差-方差权衡决策的强调，是对反思经济计量学方法论的疏漏的极好警示。计量学将任何先验理论模型必有泛化性视为不言自明的前提，相当于完全忽略理论模型的近似有偏性质，从而引致学科从认识论上误入歧途。这集中表现在学科内对估计量一致有偏性孜孜不倦的关注。在机器学习的偏差-方差权衡理念面前，经济计量学的基本前提显然是站不住脚的。

[16] 在机器学习教科书中，通常把大样本的数据集分为三个子集：训练、测试和检验子集。测试子集用于模型选择，而检验子集用于模型预测评价。

那么，机器学习的 PAC 可学性理论是否就真正适用于经济计量学呢？前面的讨论业已表明，无论经济理论的数学推导如何缜密，面对现实数据，f的先验不可知性是不容忽略的，采取后验模型近似逼近，是减少理论模型设定偏差的唯一可行路径。这里，我们再来考察一下 PAC 可学性理论中的三个元素：$\mathcal{H}$、ℓ 和 $\mathcal{D}$是否与经济计量学中的对应元素相互匹配的问题。几乎所有的实证性经济理论不外乎是从效用或利润最大化、或风险或成本最小化的理性行为优化原则而推导出来的。如上节式 (2.1.1) 所示，这类优化原则将经济关系式定义在凸函数假说类，而这一假说类正是 PAC 可学性理论形成的基础元函数$\mathcal{H}$。在凸函数大类中，机器学习中常用的线性函数种类，主要是针对两种学习目标而设定的：(i) 目标变量$\mathcal{y}$ 为连续变量； (ii) 目标变量 $\mathcal{y}$为离散变量，尤其是反映分类问题的二元变量。相应地，对应于 ERM 准则的通用损失函数ℓ有两种：用于回归式的二次损失、即差方损失函数，以及用于分类式的逻辑回归损失函数。最初的 PAC 可学性理论始于后者，随后才扩展到前者。经济计量学虽然对优化准则会有不同的诠释，但在应用模型研究中采用这两类关系式的情形明显据首。PAC 可学性理论针对的数据样本，是沿用经典统计学的独立同分布假设条件收集的样本集。不过，可学性理论的近期发展已经从独立同分布条件扩展到专门针对随机时序数据样本的理论，如参见 Zimin and Lampert (2017)及 Dawid and Tewari (2020)。总的看来，可学性理论中的三个元素与它们在计量学中对应元素有着相当高的匹配程度。

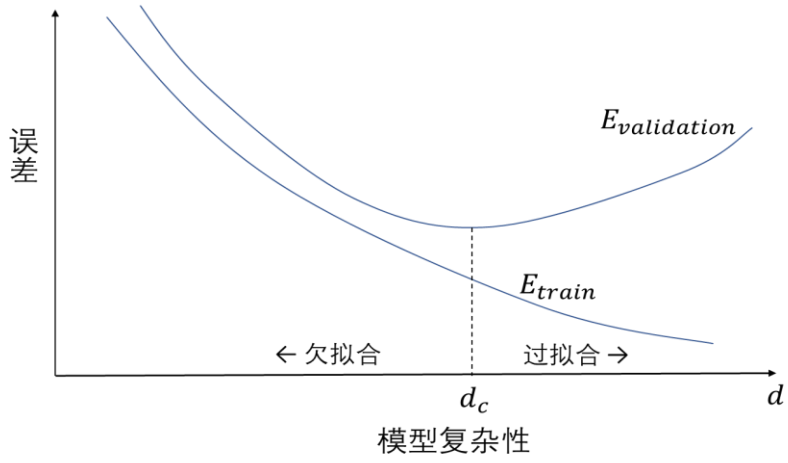
显然，可用的经济数据样本量会对 PAC 学习理论在经济计量学的引入构成一个潜在约束限制。但从现有的计量学案例所用的数据样本看，样本量不足的制约还是相对较弱的。如上所述，计量学的绝大多数案例中所涉的假说函数类和损失函数都属于机器学习中常用的类型。机器学习理论方法一旦被计量学采用，应能扩展计量模型目前所能分析的应用问题范围，使计量学释放出更大的研究潜力。不过我们必须认识到，PAC 可学性所能处理的理论贫乏类问题仍是有限的。其学习范围的一个严峻局限是算法计算能力的限制条件，亦即学习的计算复杂性问题，如参见 Valiant (2013, Chs.3 & 5), and also Shalev-Shwartz and Ben-David (2014, Ch. 8)。决定计算复杂性的一个关键点是，针对现实中的复杂问题，我们到底能够采用多么简洁的表述形式来刻化？在机器学习中，追求模型表述的简洁性其实是一条基本标准。反映这一标准的重要性的最好例子是以结构风险最小化(SRM)为准则的可学性理论的形成。PAC 可学性理论所采用的 ERM 准则，需要以经验风险一致收敛作为前提条件。但就现实中各种需要学习的问题来说，能满足这种一致收敛性的问题是有限的。以 SRM 为准则的可学性理论是针对非一致收敛型问题而构设的。这类理论所采用的基本衡量准则，恰恰吻合于哈维尔莫在他的第二章讨论的两个理想建模标准，即‘简化性’和‘自律性’。

2.3 简化性和自律性：奥卡姆剃刀定律与稳定性

当数据源于开放式不可控场景、样本又有限时，$N \to \infty$时的一致收敛条件就不能作为模型学习可依赖的状态了。这意味着，ERM 准则也就不再胜任以泛化为目标的归纳学习任务了，参见 Mukherjee *et al* (2006)。这时，我们仍可沿用式 (2.2.3) 所示的偏差-复杂度权衡理念，利用样本的$\mathcal{D}_{train}$与 $\mathcal{D}_{validation}$子集分割，试验选择出泛化度最好的关系式。图 2.1 便是基于这种试验的学习曲线示意图。图中代表训练模型拟合误差的E_{train}，随着模型复杂度 d的增加而下降。然而，代表模型测试预测误差的$E_{validation}$则不具相同属性。$E_{validation}$的变化取决于d 的特定临界阈值d_c。只有当$d < d_c$时，$E_{validation}$才随着d的增加而下降。而一旦有$d > d_c$，$E_{validation}$则随着d的增加而上升。因此，通称所选 $d < d_c$ 的模型为欠拟合模型， 而称所选 $d > d_c$ 的模型为过拟合模型。如果不做上述试验，仅据 ERM 来选取模型，其结果并不一定是泛化性最

高的模型。这时，我们可以利用$E_{validation}$的属性来取代由 ERM 准则引致的对一致收敛条件的依赖。具体地，我们在选模时，同时考虑模型拟合与预测精度，亦即把模型拟合最大化与预测误差最小化同时编入学习算法中。图 2.1 展示的$E_{validtaion}$与d_c的关系告诉我们，预测误差的最小化与模型的复杂度息息相关。该标准实际上要求选择可实现预测任务中最为简洁的模型。在机器学习文献中，这一模型简洁性原则被通称为欧卡姆剃刀定律。该定律源于 14 世纪逻辑学家欧卡姆“如无必要，勿增实体”的哲学思想。显而易见，哈维尔莫对于模型简化性的探求和思路，现已被机器学习中的欧卡姆剃刀定律有效并精准地反映出来了。

图 2.1 学习曲线示意图



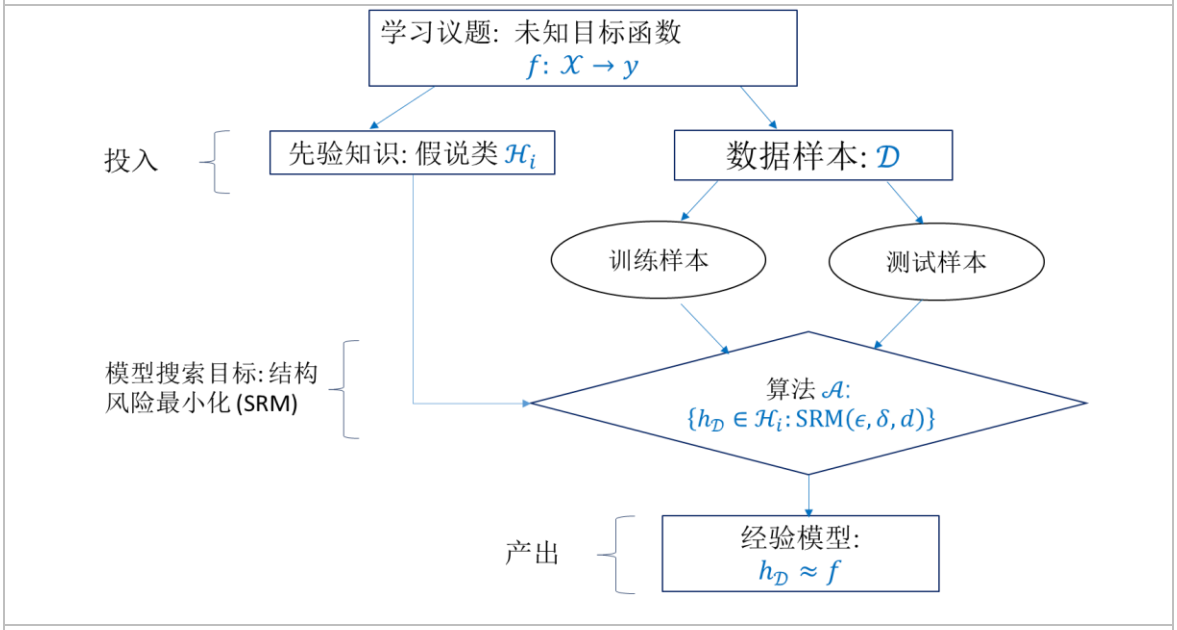
正是由于 ERM 无法顾及非一致收敛型问题，以结构风险最小化(SRM)为准则的可学性理论才应运而生。基于 SRM 的可学性理论的思路大致如下（详见 Abu-Mostafa *et al* (2012, p178)）。首先，将假说类内的所有元素按‘结构’等级排列为一嵌套序列：$\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \cdots \subseteq \mathcal{H}_d \subseteq \cdots$。再对每个$\mathcal{H}_i$依 ERM 准则选择：$h_{i,\mathcal{D}} = argmin_{h\in\mathcal{H}_i} E_{i,in}$。最后，在$E_{in}$最小化准则之上增加一条针对模型复杂性的惩罚法则$\Lambda$，从而在上述所选出的一系列$h_{i,\mathcal{D}}$中做出最后选择：$h^*_{i,\mathcal{D}} = argmin_{1,2,\cdots}\left[E_{i,in}\left(h_{i,\mathcal{D}}\right) + \Lambda(\mathcal{H}_i)\right]$。$h^*_{i,\mathcal{D}}$ 即为依 SRM 准则学习得出来的模型。由于它规避了过拟合风险、复杂度吻合图 2.1 中d_c，因此是满足欧卡姆剃刀定律的最优模型。以上的最后一个步骤也相当于从所有互具竞争力的假说模型中选取泛化界限最小的一个模型[17]。

以 SRM 为准则可学性理论把预测误差最小化明确设为$h^*_{i,\mathcal{D}}$的选择约束条件。由于测试模型预测功能在学习中的关键作用，就产生了利用重采样技术对模型性质在不同样本容量情形下的关注分析考察。从变化$\mathcal{D}_{train}$ 与 $\mathcal{D}_{validation}$分割的样本容量角度看，基于 SRM 所选模型的预测误差项的期望值必然具有一致收敛性。在可学性理论中，这种一致性被公理化为模型的稳定性，并由测度$E_{validation}$的稳定程度表出。业已证明，模型的稳定性是泛化性或可学性的一般必备条件，如见 Poggio *et al* (2004)、及 Shalve-Shwartz *et al* (2010)。检验模型稳定性现已成为学习算法设计中的一个重要环节。实践中，谋求偏差-复杂度权衡的最优选择也等价于谋求拟合-稳定性权衡的最优选择，详

[17] 值得注意的是，依 SRM 准则的可学性理论与经济计量学建模研究中 Hendry and Richard (1982) 倡导的‘累进式研究策略’的基本思路不谋而合。特别地，Hendry 推举的各种简洁包容原则（如参见韩德瑞和秦朵 (1998, 第 14 章)），虽然并未集中强调模型泛化性目标，却显然与奥卡姆剃刀定律密切相关。

见 Shalev-Shwartz and Ben-David (2014, Ch13)。不难看到，机器学习理论对于“稳定规则必无过拟合”理念的公理化，不仅是对哈维尔莫诉求模型构造的自律性和简化性的简洁数学形式化，而且为这些建模标准的应用实施铺平道路。

图 2.2 函数学习问题的可学性理论示意图



根据 Abu-Mostafa *et al* (2012)书中的图 1.2 改编。

图 2.2 是机器学习模式的简要图示，以更便于我们对经济计量学方法论的薄弱之处进行反思。值得一提的是，以上简述的可学性理论及图 2.2 的图示与经济计量学界对机器学习的流行看法有着显著差异。计量学界的流行看法是，机器学习是黑箱操作，单纯靠数据驱动追求预测拟合最优；而计量学的重心是解释经济因果关系，因此与机器学习存在本质分歧。其实，机器学习要比经济计量学更为重视理论模型的构述和选择。只不过由于假说类$\mathcal{H}_i$的最初构建对专注题材知识的依赖性，机器学习教科书的主体只能放在后续的数据分析方法技术方面。在可学性理论中，$\mathcal{H}_i$的选择对于样本复杂性和泛化界限紧密度起着先决作用，而这两个参数又是学习成败的关键测度。机器学习理论特别要求研究者在选择假说空间时综合考虑两方面因素：一方面假说空间相应于所关注的现实问题必须足够大，以满足研究问题的基本需求；另一方面假说空间相应于现有数据样本又必须足够小，以满足选模对于泛化性检验的需求。这意味着，研究者在选择假说类时就需要有既防止欠拟合又防止过拟合的意识。不幸的是，这种建模意识在经济计量应用研究中普遍匮乏。更不幸的是，作为经济计量学中的两个基础概念——偏差或有偏性概念与一致性概念，与机器学习中的对应概念之内涵大相径庭。它们在经济计量学中是狭窄定义在先验给定参数估计场景内的。机器学习的建模视角，促使我们要重新审视和评价这种窄义的认识论基础。

显然，SRM 准则的实施需要在算法编程中引入多重决策准则与条件，其所用数学工具远远超出经典统计学工具箱的范围。更具体地说，$\mathcal{A}$的设计需要兼顾到计算数学工具与统计学工具的相合作用性。更为重要的是，这一建模途径要求研究者注重对先验知识与后验数据信息的融合过程做出更好的改革设计。为了借鉴引入机器学习的成功经验，我们必须要接受机器学习理论中的基础理念，这就必然在经济计量学界引起一场学科基本规则对策的变革。

2.4 由机器学习范式到经济计量学习任务的构建

以 PAC 概念为基础的各种函数可学性定理，都是针对所谓的有监督学习一类问题构造的。这类学习问题的划界要求是，所涉数据样本中的变量已被划分为输入和输出两类。输出变量亦称响应变量，是模型学习任务的目标，通称为*标签*目标。正是因为这些响应变量的标签，我们就可把它们作为机器学习的范例，引导我们探求和选择输入变量得以有规律性地生成这些标签目标的最优模型。处理有监督学习问题最常见的有两种模型，一种是处理分类问题的逻辑回归模型，其目的是预测分类标记；另一种是一般的回归模型，其目的是预测数值标记。

与有监督学习问题相对应的是无监督学习问题。无监督学习问题所涉的数据样本中没有带标记的输出变量，所有变量都是输入变量，因此学习任务缺乏明确的范例引导。统计学中传统使用的主成分分析法就被视为一种处理无监督学习问题的模型。不过，目前无监督学习问题的主要模型莫属于聚类分析模型。与有监督学习问题的建模任务相比，无监督学习问题的建模任务要更为艰巨。由于缺乏标记范例的引导，研究者需要自行选取设定建模所需的准则和标准条件。这些准则和标准条件不仅要针对数据样本充分必要，而且要针对关注问题富有意义。显然，这些准则和标准条件的选取必含有偏性，该有偏性也必然在模型输出结果中体现出来。

半监督学习问题是由有监督与无监督学习两类问题扩展而来的，详见 Chapelle *et al* (2006)、van Engelen and Hoos (2020)。常见的半监督学习问题是，所涉数据样本中只有有限数量的有标记响应变量。也就是说，引导学习任务的标记范例有限。另一种半监督学习问题是，所涉数据样本虽然没有任何有标记的响应变量，但从关注问题的先验知识中可以得到某种约束条件，它们对学习任务起一定的引导作用。与处理半监督学习问题的视角密切相关的是转导 (transduction) 推理的概念。转导推理与一般的归纳推理不同，它针对的不是从特例到一般（亦即从样本到总体）、而是从特例到特例的归纳学习任务[18]。转导学习的一种常用应用工具就是基于聚类准则的近邻分类算法，如见 Gammerman *et al* (1998)。

不难看到，经济计量学中的大多数应用研究都可归于有监督学习问题的范畴。这些研究不仅有先验因果理论关系的定位，而且还有含充分标记响应变量的数据样本。一般的回归模型与逻辑回归模型在计量学中的广泛应用是有目共睹的。诚然，在预测目标变量之外，经济理论往往会含有更多的条件要求。例如，某个因果关系属于负数关系、某个因果关系应具变量增量间的正向关系等等。从机器学习的视角看，这些附加条件其实为学习任务提供了更多的引导信息，因此不改变有监督学习问题的范畴分类。而基于有监督学习问题的泛化可学性理论，为系统解决我们在第一章中例举的各种建模不确定性的难题指出一条可行之路。可学性理论启迪我们，对于假说和数据做出什么样的条件要求，我们就能将建模的艺术部分转化为有章可循的科学部分。它还启迪我们，归纳学习模型的研究规范是与经典统计学的研究规范有实质性区别的。从机器学习的角度出发，经典统计学专注的参数推断学习属于演绎问题，不属于归纳问题。

机器学习中根据输出标记信息对不同学习问题的分类，也警示我们有必要反思经济计量学可能存在的相关疏漏。这里提及两例。第一例是模型构造用于宏观经济预测

[18] 转导归纳学习概念是 Vapnik 在 1970 年代中期提出的。有关转导学习的细节可参见 Gammerman *et al* (1998)、以及 Chapelle *et al* (2006)一书中的第 24 和 25 章。这里值得一提的是 Vapnik 原则：在为某题寻解时，必须避免将一个新的一般性问题设为解决问题方案的中间步骤。

的前导指数。这一课题源于约一个世纪前的经验式商业周期研究，是一项经济学家一直未能攻克的难题。从建模工具看，虽然现有技术手段已从早期通用的主成分分析法大有扩展[19]，但原则上仍然是把前导指数建模任务视作无监督学习问题来处理的。机器学习分类法警示我们，这一任务其实是含有一定的监督学习的目标信息的，该信息即为构造前导指数所期待预测的宏观经济变量！由忽略目标信息的模型生成的前导指数，系统缺欠对目标的预测力也就不足为怪了。第二例是微观计量学通称的限值因变量(limited dependent variable) 模型。这类模型针对的数据样本，正属于有标签响应变量的数量有限的一类。在泛用的一般回归模型的视野局限下，计量学把这类样本称为截尾 (truncated) 数据或者删失 (censored) 数据样本，把一般回归模型的欠适用性诊断为选择有偏性问题，并采用修改估计法的策略来处理。机器学习分类原理警示我们，这类模型应该属于半监督学习的研究范畴。的确，从相关的微观经济学课题考虑，这些课题最关注的并不是一般的归纳推理问题，而是转导推理问题，即与有限的有标签响应变量相关的群体的推理效应，如该群体的消费倾向或者劳动力供给倾向等等。我们在第四章再详细讨论有关上述两例的建模设计问题。

模型构造失误的危害及其后果之严重性无需赘述。这里仅引用英国统计学家 D. Hand (1994) 发表的“解构统计问题”一文中的一段陈述：模型构造与‘用统计学工具识别数据中的结构和模式’相比，属于‘更高层次的问题’；这是因为模型构造‘从起点上决定了被研究问题是什么、应用什么工具来处理的问题’ (p317)。鉴此，统计学界流行将模型误设称为第三类错误，用来描述为研究问题误构的模型研制正确统计检验和估计方案的做法。

其实，上述‘更高层次的问题’在人工智能学界被归为‘知识表示’或‘机器推理’问题，并且受到广泛深入的探讨和研究，如见 Russell and Norvig (2016, Part III)。这方面研究的主题是，为了便利于机器学习任务，如何将人脑知识忠实而有效地转换表述为计算机可运算的工具手段。从认知角度看，除了要满足逆转性和简洁性之外，知识表示都具备两个最基本特性：主体论承诺和认识论承诺 (ontological and epistemological commitments)（见上述引文中的第 8 章）。知识表示的本质是做出一组主体论承诺，亦即对现实的本性假定。这组承诺的选择本身并不包含对数据结构的任何承诺，但对研究的框架与视角起着奠基作用。从认识论角度看，知识表示必然包含一组推断，这组推断的选择与后续的计算机工程处理知识信息的方式和路径息息相关。任何推断都是不完全的，这一不完全性必然寓意着不确定性。Davis *et al* (1993) 把这一特性称为‘智能推理的不完全理论’。显而易见，数学处理上述不确定性的一种自然选择是采用概率论的因果推理方法。但这并不是唯一的推理方法，泛化可学性理论所选择的*无分布*学习的框架便是最好的佐证。这里，推断的不确定性只是明示在 PAC 概念中，但在泛化可学定理中，并未对所涉数据的生成机制、或假说模型类中的变量做出任何概率分布的假定。这种处理方法其实是对人工智能知识表示的主体论承诺的具体体现。将机器学习的可学性理论引入经济计量学，就意味着我们必须重新审视概率论在经济计量学内的作用，质疑其目前享有的绝对基础地位。那么，依据机器学习理论，我们该如何重新认识概率论在经济计量学中的作用呢？下一章的主要任务便是探索这一问题。

[19] 有关综述可参见 Marcellino (2006)。

3. 经济计量学中概率论的基本功能

所有可测经济变量可由一个相应的联合分布的随机变量空间来刻化，这是哈维尔莫构思和倡导概率论方法的最基本理念。经济计量学发展至今，特别是从理论上对经典统计推断技术的全盘系统接纳和拓展，都构建在哈维尔莫上述基本理念之上。然而必须看到，上述理念需有如下前提假设的支撑：变量所在系统是个闭系统，该系统的结构需相对简单，从而能近似于经典统计学所基于的随机试验场景。虽然几乎没有文献明确阐明这一假设，但它显然是经济计量学的一条“关键性”假设。“与现实不符的假设尚可采纳，但与现实不符的关键性假设不可纳”， 这是 Rodrik “对经济学家之十诫”中的一诫 (2015, p116)。上述关键性假设显然与现实不符，但就经济计量学的研究对象而言，其不现实程度是否仍可被忽略不计呢？

第一章业已阐明，概率的概念适用于表示经济现实中不确定性的场景是十分有限的。概率概念并不适用于分析任何无模型任务限定的经济变量。第二章进而阐明，对于开放式场景中的建模任务而言，其不确定性大大超出单靠数学先验演绎推导路径便可有效驾驭的范畴。将建模任务视为机器学习方法中的学习任务，才是目前最为有效的路径。这主要是因为，机器学习方法摆脱了上述关键性假设，并且警示我们，在学习到确凿可信赖的应用模型之前，将研究课题归于统计推断的范畴为时过早。为了进一步审慎评价机器学习方法的可学性和优越性，本章将讨论视角集中在概率概念在应用模型研究中所起到的基本作用。

在 3.1 节里，我们从经济变量的联合分布理念出发，考察该理念在变量作为模型目标变量（亦即响应变量）时的作用。概率论的一条基本法则是连锁法，即随机变量的联合分布可被因子分解为条件分布和边缘分布。然而，在计量模型实践中，连锁法却难以兑现。若按连锁法把应用模型划分为基于边缘分布、条件分布和联合发布的三种模型类型，我们便可发现，只有基于条件分布的模型类能够得到数据支持，而基于边缘分布或者联合分布的模型类通常缺乏足够的经验数据支持。我们在 3.2 节追根寻源，集中考察条件模型形成过程的特点。不难发现，作为条件模型中核心部分的因果关系，其理论构造过程一般不需要引入概率概念。理论构造完成之后，为了应用分析，通常在因果关系上附加一随机误差项或随机扰动项，于是概率论才被引入。按机器学习的术语描述，如此形成的条件模型属于辨别式 (discriminative) 模型，而不属于生成式 (generative) 模型。在 3.3 节里，我们借助人工智能对于知识表示的分析视角，将经济学理论推导过程定位在逻辑推理过程上。因此，我们得出结论：经济计量学的应用条件模型研究符合机器学习中的无分布学习范式。正是因为由逻辑推理而来的因果关系，与现实之间有着不可忽略的不确定间隙，我们才必须采用归纳性学习手段来寻求有效地综合先验知识与数据实例信息的最佳模型。概率论则是辅助这一学习的必备工具。在学习过程中，概率论的基本作用是辨别功能，即为模型选择决策做量化裁断。只有把概率测度集中用于模型学习中的诊断性用途上，我们才有希望系统改善目前计量应用模型研究中普遍存在的对数据信息提取效率低下的问题。

3.1 作为目标变量时的可测经济变量之随机模型特征

在计量学的应用研究中，将可测经济变量表述为由随机变量组成的联合分布系统的理念究竟起了什么作用？这一理念是否大致符合现实需要？这便是本节的议题。第一章业已阐明，经济变量的不确定特征所涉范围广泛，采用概率的概念来笼统表述过于简单化。这里，我们将考察范围限定在经济变量作为模型研究的目标变量的情形，

并且从上述理念对于应用计量学的研究策略与方向之影响的视角，来评判用联合分布刻化可测经济变量的基础理念到底有什么实用可行的成效。

概率论的一条基本法则是连锁法，即任何随机变量的联合分布都可被因子分解为条件分布和边缘分布。如：

(3.1.1) $$P(x,y) = P(y|x)P(x) = P(x|y)P(y)$$

就模型研究中的目标变量而言，上述连锁法意味着，我们既可以从联合分布角度对所涉变量联立建模，也可以分别按照条件分布和边缘分布的理念建模；从第二种路径所依的分解原理看，它应比前一路径更具科学深度和基础性。应用计量模型研究的发展轨迹也大致可由这两条路径来归纳。标志学科初起的宏观计量联立模型研究[20]，正是第一种路径的直接产物。这里最能集中体现联立模型研究的学科基石作用莫属‘内生有偏性’概念。这一概念源于哈维尔莫在二元变量静态联立模型基础上对普通最小二乘法有偏性的推证。虽然静态联立模型早已在宏观应用研究中销声匿迹，‘内生有偏性’仍属学科之大忌甚至最忌，特别是在应用模型归于条件模型类的情形下。其实，自 1973 年的石油危机以来，宏观计量模型研究便转向基于分解原理的第二条路径了。其最为明显的标志便是突出强调模型所涉单个变量之数据生成过程（data generation processes (DGP)）。如今，单个变量的随机时序特征被广泛认为是建模初始所需考察的基本特征，特别是模型动态设定的基本特征。

与宏观经济计量学可观的理论发展相比，应用模型的研究成果明显相形见绌。不难发现，凡是基于联合分布或者边缘分布的应用模型结果都不尽人意，缺乏泛化性，颇经推敲的成果大都来自条件模型。应用建模研究似乎一直陷于僵局，即停留在概率分布刻化‘不完全’的条件模型之上。为了仔细寻求僵局的症结所在，让我们来依次考察基于上述三种分布理念的时序变量模型。

根据连锁法的分解原理，基于边缘分布的时序模型应该属于最基础的一类模型。我们就先来考察一下这类模型。在宏观计量研究中，通用的做法是依可观测到的单个时序对于时间的依赖性来进行变量分类。具体地，依变量的一、二阶矩与样本量间的关系特征将它们区分为（弱）平稳过程与非平稳过程。应用建模中通常采用自回归模型（autoregressive (AR)）来刻画这一特征。以最简单的一阶自回归模型为例：

(3.1.2) $$y_t = \rho y_{t-1} + u_t,$$

当式中的参数估值$|\rho| < 1$时，便把 y_t 归为平稳过程，而当$\rho = 1$时，我们则称之为非平稳或者单位根过程。文献中有关宏观变量的单位根结果比比皆是，一个重要原因是单位根现象被视为获得协整关系的基础前提，而协整关系从概念上非常接近于多数经济学家所钟爱的均衡理论。遗憾的是，经济变量的单位根结果一般都经不起严格推敲。首先，单变量时序模型的参数估值一般缺乏样本外的不变性，因此泛化功能弱。即使选择更复杂的边缘分布函数、采用参数非线性模型，也难以改变上述缺陷。也就是说，描述单个变量分布的基础矩参数通常不具有时不变性。其次，在经典统计学的随机抽样试验场景下，通常将威胁样本参数估值不变性的异常观测值视为意外野点，把它们从样本中剔除。但在开放式的经济现实面前，这些违反常态观测值往往含有特殊信息，不容忽略。从大多数应用研究课题来看，单变量自回归模型的实用功能在于概括变量的动态表象，模型所展示的是变量经济动态活动的结果，而不是变量内在的生成机制。把模型测度的

[20] 有关经济计量学形成史，可参见 Qin (1993)。

表象归为变量的动态本征，显然是逻辑有误的推理。从经济学家的视角看，解释很简单：变量间的相互影响作用是经济活动的最基础本征。单变量模型忽略这一本征，必受遗漏变量之扰，因此基于边缘分布的单变量模型显然不足为信。

下面我们来看看实践中相对成功的条件模型这一类。让我们拿颇受欢迎的自回归延迟分布（Auto-regressive Distributed-lag (ARDL)）模型为例。现设如下的 ARDL(1, 1) 模型已被证实是与数据吻合的模型：

(3.1.3) $$y_t = \alpha y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + \varepsilon_t$$

必须强调，(3.1.3) 被证实为是与数据吻合的模型的一个必备条件是：$|\alpha| \ll 1$。显然，既然有 (3.1.3) 与数据吻合，就不可能同时也有 (3.1.2) 与数据吻合。对照 (3.1.3)，(3.1.2) 因不含x_t和x_{t-1}导致模型欠拟合，而且参数ρ必受遗漏变量有偏性之扰。这时，如果由 (3.1.2) 估计得出$\hat{\rho} \approx 1$，那么ρ的遗漏变量有偏性则属于上偏。这意味着，认为y_t的基础数据生成过程为单位根过程的认知是有逻辑漏洞的。当然，据连锁法的分解原理，相对于 (3.1.3) 而言，需考察的单变量模型应以x_t为目标变量。但是在应用条件模型的场景下，寻建x_t的边缘分布模型既没有必要也不可行。就以y_t为目标变量的建模学习任务而言，是没有必要去学习x_t的随机分布机制的。而且，前面对自回归模型的分析业已表明，单变量时序模型并不等同于刻化经济变量生成机制的模型。以 (3.1.1) 的连锁法分解原理作为建模研究的策略是与经济现实大相径庭的策略。

最后我们来反思一下基于联合分布原理的应用建模研究结果。计量学初始注重联立方程模型的研究，是对以描述经济变量间相互作用为重心的经济学假说的响应。当今，上述传统则反映在许多宏观应用经济学家规避单一方程模型、偏爱向量自回归（Vector Auto-regression (VAR)）模型的倾向。不过，实际可操作的多方程模型其实也属于条件模型一类[21]；VAR 模型就是基于滞后变量的条件模型。现将上述 ARDL(1, 1) 模型扩展为一个简单的闭式 VAR 模型：

(3.1.4) $$\begin{pmatrix} y_t \\ x_t \end{pmatrix} = \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix} \begin{pmatrix} y_{t-1} \\ x_{t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix}, \text{ 假设: } \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix} \sim IIN \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} \right]$$

选用模型(3.1.4)而不用(3.1.3)的通用理由是，(3.1.4)与变量联合分布的本征相呼应，即不对式(3.1.1)的因子分解方向、亦即是$P(y|x)$ 还是 $P(x|y)$、做出选择。然而，这种维持联立性的信念过于虚幻肤浅。首先，应用 VAR 模型的变量集选择都是具有明显因果方向偏倚性的。以包含国内生产总值、通货膨胀、失业率和利率四个目标变量的 VAR 模型为例。这四个变量的选择显然是以描述实体经济、特别是国内生产总值为重心的，变量集的选择对解释利率变量的考虑要明显少于对解释国内生产总值的考虑。因此，VAR 模型中各单方程的数据解释力或拟合度通常具有明显差异。其次，在使用 VAR 模型做脉冲响应分析时，我们必须对模型中变量做出明确的冲击先后排序，这种排序就相当于动态因果设定，必然打破初始模型设定的联立对称性。另外还需指出的是，不少经济学家都把 VAR 视为‘简约式’模型，而不是‘结构式’模型。为了将结构模型特征引入(3.1.4)，通常的做法是把变量间的联立性作为一条结构参数约束条件，如：

(3.1.5) $$\begin{pmatrix} 1 & \beta_{12} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} y_t \\ x_t \end{pmatrix} = \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix} \begin{pmatrix} y_{t-1} \\ x_{t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix}, \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix} \sim IIN \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \right]$$

[21] 传统静态联立模型所需的模型识别条件，就反映着联立模型必须满足条件模型设定的基本可操作前提要求；有关静态联立模型与条件模型之争有着渊源历史，可参见 Qin (1993, Chs 4 & 6)。

这样一来，原模型表面的静态因果对称性也就消失了。另外，实践中 VAR 模型的对称性缺失还可从两个方面观测到。其一是模型估计后一般都有某个方程内的某些参数统计不显著（即为零）的结果，导致使用中的 VAR 模型不具对称性。其二是出于提高模型拟合度的需要，建模者时常放弃闭式 VAR 模型、采用开式 VAR 模型，即试在模型中加入其他控制变量。这种做法其实是对变量联合分布基础理念的摒弃。

简言之，对变量间做某种因果不对称的假说是任何统计上可操作模型的必要前提，如参见 Cox (1992)。反思大量经济计量模型研究的经验教训，无论是单方程模型还是多方程模型，凡是颇经推敲的应用成果都来自条件模型一类。是什么原因造成经济变量联合概率分布的理念与应用经济建模研究的实践如此事与愿违的呢？

回顾哈维尔莫书中将概率测度在模型中正式引入之处：他的式 (11.2)，该式也为条件模型类。该式相当于在我们上章中的 (2.1.1) 内附加一个随机分布误差项 s。按哈维尔莫原话：s 是 “具有某种概率分布的随机变量”（哈维尔莫原著第 51 页）[22]：

(3.1.6) $$y = h_p(x_1, x_2, \cdots, x_n; \kappa) + s$$

遍览应用计量模型研究文献就不难发现，关系式(3.1.6) 占据中心位置。随机变量s的核心功能是为了使参数κ的区间估计和统计推断合理化。将所有变量定义为随机变量只不过是一种习惯形式。值得注意的是，哈维尔莫在引入s之后，也没有继续讨论由误差项的概率设定可能引致的有关y和$x_1, x_2, \cdots, x_n$的随机变量设定问题。他继而关注的问题是：如何将y足够合理地分解为可系统描述的部分h_p和随机残差部分s。他用了一个章节的篇幅详细探讨这一问题（见原著第 12 节）。他的做法充分反映了该问题在计量模型研究中的核心位置。

我们在上章业已指出，哈维尔莫的分解问题实为 PAC 可学性问题，超出了经典统计学研究的范围。从 PAC 可学性的理论视角，我们应能将基于 (3.1.6) 的应用建模任务纳入机器学习中的无分布模型学习方法轨道。这一可能性将使经济计量学从前述的困扰中解脱出来。不过，在接受这一可能性之前，我们需要弄清下述问题：概率论推理在经济计量学建模中究竟起到多么关键的作用？或者说，概率论推理在经济计量学建模中的关键作用具体何在？

3.2 概率论在经济学/经济计量学建模中的作用

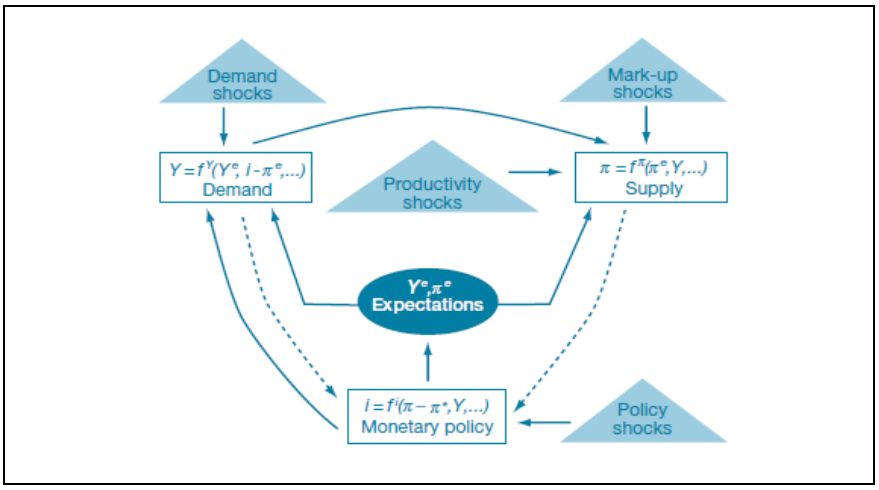
让我们先从经济学建模的环节入手，考察经济学家一般是如何处理建模任务中的不确定性、得出 (3.1.6) 中的关系式h_p。建模初始，经济学家通常是把理论视角集中在所谓‘有代表性主体的合理预期行为’之上，该视角得以滤掉不少现实中的不确定性。继而，经济学家对于因果关系假说的演绎都是以某种功效最优法则为基础，如利润或效用的最大化、成本或风险的最小化等等。这些法则被公认为驱使有代表性主体做出经济行为决策的最基本动机。由于这些法则都不是随机法则，推导h_p的数学工具基本归于微积分及有约束优化解的范畴。另外，‘假设其它因素均保持不变’也是一条理论模型通用的前提假定条件，以覆盖理论模型的偏倚角度，维护理论仅关注数量极小的输入变量集的合理性。在h_p的推导完成之后，随机不确定性因素才被纳入关注日程。通常的做法是在h_p上附加一个随机误差项s，从而实现模型与统计学所涉模型之对等形式。哈维尔莫把h_p中的变量描述为随机变量，其实完全是为了给这种对等形式正名。

[22] 为了保持本书中数学公式的连贯性，这里对哈维尔莫书中的式 (11.2)所用的符号有所改动。但是改动不影响原式本意。

在附加了随机误差项之后，不确定因素在 (3.1.6) 中便有了两个出口：参数κ的估计值和s，后者主要是为完成参数估计任务而立的。可见，在经济学理论模型推导过程中，概率论推理基本没有起什么作用。

如何缩小理论模型与现实间的差距，是经济学发展的主动力。缩小差距的关注点一般落在如何放宽‘有代表性主体的合理预期行为’这一理想假定条件上。具体地，经济学家力图识别出某种特定场景，在这种场景下，功效最优法则的表现更为显著清晰，因此更容易得到数据验证[23]。从数学模型的先验推导看，对这种特定场景的表述相当于给按效用优化法则的因果关系附加更多的约束条件。因此，模型的复杂度相应增加。这些附加约束大都落在两方面：动态约束和需求双方的相互制约。后者一般由对不同方程式中参数的约束条件表出。至于动态约束，虽然大多数结构复杂的动态模型都采纳了随机变量的概念，对模型内理论关注的动态结构关系的推导，仍限于确定性数学范畴。就以协整分析为例，其理论关注的是长期静态均衡关系的解的特征。相应的数学推导完全可由确定性系统论来承担。只要对目标变量的轨迹做出不同类别假设，就能得出相应的微分或差分方程的动态阶数特征。可见，这时动态模型结构的先验设定并不需要用概率论推理。

图 3.1 DSGE 模型的基本构造示图



来源: Sbordone *et al* (2010)

在通用的模型类别中，随机因素明确影响模型先验动态设定的莫属动态随机一般均衡 (DSGE) 模型。DSGE 模型属于多层结构模型，也是公认宏观应用模型中理论含量最为丰富的一类模型[24]。虽然微观理论基础是 DSGE 模型的特征标志，刻化微观行为也是模型结构层次的伸展功能设计之初衷，但模型最终关注的还是一组宏观目标变量。就模型内单个方程的函数形式而言，通用的方式是采用效用优化法则的欧拉方程近似解得出的线性或对数线性方程。方程目标变量的随机性则来自一个附加的外生随机扰动变量。这类变量是根据所在方程式性质赋名的，如图 3.1 所示的‘需求扰动’、货币‘政策扰动’。如此的赋名做法就给这些变量增添了与其它经济变量为伍的‘结

[23] Gilboa *et al* (2014) 将这类研究归类为以案例为基础的推理分析，以区别于以效用法则为基础的推理分析。可惜他们的分类法并不准确，没有考虑到案例背后的分析仍然是依据效用法则推理的事实。

[24] 有关 DSGE 模型的介绍和综述可参见 Canova (2009)、Sbordone *et al* (2010)、 Fernández-Villaverde *et al* (2016)以及 Christiano *et al* (2018)。

构’性质。同时，由于这类扰动变量是潜变量，建模者得以通过模拟试验扰动变量不同的动态结构设定，来尽量缩小所设理论因果函数式与可测目标变量之间的差距。对扰动变量通用的动态设定是 AR 自回归过程。就以图 3.1 中的供给方程的加成（mark-up）扰动变量π为例。可设该变量服从一阶 AR(1)：

(3.2.1) $$s_{\pi t} = \rho s_{\pi t-1} + \varepsilon_{\pi t},\ 0 < \rho < 1$$

式中的$\varepsilon_{\pi t}$是由计算机模拟生成的白噪声误差项。值得强调的是，由于外生随机扰动潜变量的引入，目标变量的随机标签便有了明确的内涵，这进一步维系了方程与统计学模型的同类匹配观念，加强了建模者对先验设定的模型可维持给定不变的信念。

从多方程模型类这一角度看，DSGE 模型可以被划为有约束的 VAR 模型类，其应用路径手法也和 VAR 模型一样，即通过脉冲分析试验来做政策实证分析。不过，DSGE 模型所含的参数约束一般比 VAR 模型复杂得多，而且模型规模也会大得多。因此，传统的统计估计方法一般难以完成模型参数的估计任务，需要通过校准法 (calibration) 来获取。模型校准的一项基本标准是对宏观目标变量之矩参数的预测准确度。这一标准被简称为对宏观矩的识别 (identification of “macro moments”)，参见 Nakamura and Steinsson (2018)。宏观目标变量的一、二阶矩是最主要的两个参数，其估值一般来源于对所关注变量的自相关回归模型。例如，采用第 3.1 节中的式(3.1.2) 来描述图 3.1 中的变量 *Y*。

不过，概率论推理在 DSGE 模型建模中的运用方式是经不住推敲的，至少存有两方面的明显认知失误：宏观变量矩参数的估值泛化性弱，和单方程内过度约束的动态设定。前一个失误在上节已经充分说明了。显然，当矩参数估值受单变量模型遗漏变量之扰而泛化有偏时，依据这种估值校准出来的模型就不可能是 ERM 意义上的最优模型。第二个失误出自建模者对扰动变量为 AR 过程的假定，无视这一假定隐含对模型动态结构强约束的影响。对扰动变量为 AR 过程的假定可追溯到 Cochrane-Orcutt 估计法，该估计法其实是通过认可静态模型残差项的 AR 形态对模型进行隐性动态扩展的做法（见第五章 5.1 节中的例子）。正是由于这种隐性动态扩展，模型才能有估计效益的改善。因此，Cochrane-Orcutt 估计法是从表象上对付模型动态设定不足的权宜之计。但是，如此被隐性动态扩展的模型附带有严格地参数共享动态特征约束，该约束已被大量数据案例所拒绝[25]。在 DSGE 模型中，构建因果动态方程通用的法则是适应性预期假设。由该假设推导出的模型类在计量学中通称为偏调模型。由于该类模型忽略了自变量滞后项的动态效应，在现实数据面前，通常被证实为动态设定不足。这时若引入计算机按 AR 自回归过程生成的扰动变量，不但掩盖了模型先验动态设定不足的问题，而且给模型的长期均衡解强加了欠数据支持的共因子约束。认识到这两方面的建模失误，面对 DSGE 模型的规模预测精度效益一直难与 VAR 模型相抗衡的现象，如见 Gürkaynak *et al* (2013)，我们也就不足为怪了。

总之，就 DSGE 模型内的单方程构述过程而言，所有的微观理论关系都属于条件型关系，有关附加扰动潜变量的做法，并不能改变概率论推理在理论条件关系推理中无足轻重的作用。另外，可测变量的边缘模型只起到对宏观变量的描述作用，并未被视为变量的微观生成机制。因此，DSGE 模型的基本建模思路仍然忠实维持在模型

[25]从建模视角对 Cochrane-Orcutt 估计方法问题的讲述，教科书可参见韩德瑞和秦朵 (1998, 第 7 章)；相关的历史信息可参见 Qin (2003, Chs 4 & 7)。这里有必要指出，参数共享动态特征的约束（亦简称为共因子约束）不仅将有关长、短期参数约束为等值，还将模型的动态结构约束为先验给定的结构。

(3.1.6) 的认知路径上。哈维尔莫对于所有经济变量服从某种联合随机生成机制的理念仍然不过是个粗幻的比喻故事罢了。

那么，如果经济学家可以实现和享有统计学所要求的随机抽样试验场景的话，哈维尔莫的理念是否就能得到数据的支持呢？这一问题将我们引向微观计量学中的项目评价模型 (Program Evaluation Model (PEM))[26]。PEM 的任务是，通过估计政策项目的平均处理效应 (average treatment effect)，对项目实施的有效性做出量化评价。PEM 的运作有一个关键前提条件：数据样本必须来自随机对照试验 (randomised control trials)。严格控制的随机抽样是经典统计假说推断的前提条件，以确保样本数据分布属性的数学易导性。然而，对于用作政策项目评价的随机对照试验数据样本来说，无论数据随机采集的设计和控制多么到位，建模中的不确定性依然是一个不可忽略的因素。

现令T代表政策项目变量，我们可用下式来概述项目评价模型：

(3.2.2) $$y = \alpha T + f_z(z_1, z_2, \cdots, z_m; \beta) + \varepsilon$$

式中的α为平均处理效应，$\{z_j\}$为其它有关控制变量，这些变量存在与T统计相关的可能性。随机对照试验的抽样要求是为了通过控制数据样本的选择来实现条件概率分布$P(y|T)$的数学易导性。然而，随机抽样并不能完全控制$\{z_j\}$内变量的属性。$f_z(\cdot)$仍然具有先验不确定性，其最优近似形式必须由经验学习而来。现有大量的 PEM 应用结果已足以佐证这一点。从机器学习的视角看，(3.2.2)只能属于一个假说模型类。而且，$f_z(\cdot)$的学习任务不需要涉及任何有关概率分布$P_j(z)$的先验知识，可以通过机器学习中的无分布建模路径来完成。正是由于该学习任务对参数α的估计会有不可忽略的影响，这使得针对αT一项就先验地把 (3.2.2) 等同于经典统计随机模型的认知过于天真。这也意味着，随机对照试验这一前提也不能确保哈维尔莫的基本理念成真。

在机器学习教科书中，通过数据学习来做出从式 (2.1.2) 中的y 分割出ε 的这一途径，被归述为 ‘判别式’ 的建模路径，以区别于 ‘生成式’ 的建模路径，参见 Jebara (2004)、Shalev-Shwartz and Ben-David (2014, Ch24)。不难看到，生成式路径是经典统计学范式的特征代表。在该路径下，先验设定的概率模型是统计推断的出发点，被视为已知确定的维持假设。这一视角为假设残差项服从噪声概率分布提供了正当理由，如通用的零均值正态分布。而当构述理论假说的模型是先验不确定的时候，实证分析的首要任务显然是搜索最可维持的经验模型。一旦搜索成功，即若建模者按第二章中图 2.2 的路径实现了 $h_{\mathcal{D}} \approx f$ 的 PAC 学习，相应得出的ε就应该与一个遵循正态分布的白噪声随机变量十分近似。于是，虽然$h_{\mathcal{D}}$是通过无分布路径得出的[27]，但是它的背景式 (2.1.2)与表述条件分布预期的随机模型形式相同。也就是说，统计学习得出的模型往往具有和条件分布预期的随机模型非同质但同形的特点。

3.3 逻辑推理与 PAC 学习合成过程中的判别式概率用途

既然经济学理论推理中对应不确定性的主要手段不是概率论，概率推理的理论作用实属无足轻重，那么经济计量学该如何认识和对待概率论方法呢？人工智能领域的研究其实已为我们指出系统解决路径。如何将不确定性由计算机可识别和操作的语言

[26]英文教科书的有关讲解，可参见 Camero and Trivedi (2005, Chs 25-26)；有关随机对照抽样试验假定对于政策评价的局限性问题，可参见 Deaton and Cartwright (2018)。

[27] 值得一提的是，无分布建模其实与近半个世纪前 H. Wold (1975, 1980) 所极力倡导的“推测性软建模”的理念是基本一致的。

表出，这是在人工智能学科倍受重视的研究课题。在这一研究方向里，积累了大量对各种信息的特征提取和分类编码、以及对各种学习任务的特征表述和分类的实证成果，可参见 Dubois and Prade (2009)、 Costa *et al* (2018)、以及 Russell and Norvig (2016, Part III)。通过缜密仔细地反省人脑在现实中做出决策行为的过程，人工智能的研究发现，在理论贫乏的日常生活场景下，人脑执行认知任务和行为决策的过程大都是模态命题逻辑思维 (modal propositional logic) 的过程，该过程一般不属于概率推理模型表述的范畴。由于人的逻辑推理认知在开放世界场景下具有不完全性的特征，我们才容易将这种逻辑推理和概率推理混淆起来，因此导致混淆模型表述功能的认知错误。从采用条件模型来表述模态命题逻辑推理的角度看，模型的目标变量实为*认知不确定*的变量，而不是统计学意义上的随机变量。为此，Dubois and Prade (2009，第 6 节) 将这种模型所担负的学习任务称为“纯真”条件决策任务，以别于基于概率推理的条件上的决策任务，如贝叶斯逆概条件推理法。

从人工智能的分析角度看，经济学对应现实中不确定性的主要手段是效用论或经济合理性的信条，如见 Doyle (1992)。经济学所关注的决策问题和学习任务一般都属于模态命题逻辑思维范畴。因此，经济理论关系式的数学优化推导过程以微积分为主，如式 (3.1.6) 中先验构造h_p的过程就不需要使用概率分布数学。换言之，h_p的主体论承诺决定了该条件式不属于概率生成式一类，(3.1.6) 的学习任务也不是简单的统计推断任务。值得一提的是，上述分析其实印证和肯定了第一章开篇就提到的经济学思想史及方法论学界对概率论理念之局限性的不信任态度。

前面两章业已表明，在理论贫乏的开放世界场景中，先验构建的h_p与经典统计学中的目标模型之间的差距甚远，寻找与数据吻合的最优模型是一个学习任务。为了有效地完成这一任务，经济学和经济计量学都需要改变研究策略。经济学的传统策略是，依赖经济学家的逻辑归纳推理能力来先验创建出参数具有经济解释意义的模型。人工智能领域的成功经验，激励我们对上述策略认真反思。具体地，经济学应该详细学习借鉴人工智能对于知识表示的处理研究途径，如基于知识的归纳学习手段、归纳逻辑编程系统等等，可参见 Russell and Norvig (2016, Ch19)。虽然知识表示的各途径间存有差异，但它们的研究策略目标是一致的，即如何最有效地将人脑中的常识性知识正规表述为可为机器学习服务的人工知识[28]，并且为实现从人机知识转换上升到扩展知识总和的目标服务。这一目标也被描述为“从机器学习转向机器推理”，见 Bottou (2013)。

简言之，人工智能的知识表示是一个逻辑语言构建过程。逻辑可量化关系法则的表达力强，因此解释力就强，而且在计算系统中的毗连功能和转换功能都很强。但是，面对理论贫乏的日常生活场景下的决策任务，由于逻辑法则内涵的全局普遍性忽略了具体场合的相应特性，这类法则难免反映出明显的脆弱性特征，如见 Valiant (2013, Ch7) 的描述。按 Valiant 的话来说，“由于数学逻辑有着清晰的语义学和严谨的推证过程，因此是一种颇具迷惑力的描述语言。但是，以数学逻辑作为大型编程系统的基础语言，必然导致系统的脆弱性。这是因为在实践中，通常不可能保证实现计算系统中的各种谓语赋名之间的一致性应用” (Valiant 2000, p231)。为了克服脆弱性，就必须针对含有噪声的数据案例进行归纳学习。只有通过对数据案例的学习过程来评价、选择和约束关系法则，才能达到依据特定现实场景的特征来强化逻辑推理的目的，如

[28] 这里值得引用的是著名经济学家 T. Sargent 在美国加州伯克利大学 2007 年毕业典礼上的一句话：“经济学不过就是有组织的常识而已”。

见 Valiant (2000, 2008)。细思忖量，这种先验逻辑推理与后验归纳学习的‘知识融汇’理念，其实是对上一章 2.3 节中所介绍的结构风险最小化准则要旨的强调和补充。

还需看到，正是在后验归纳学习的过程中，概率测度才起到不可忽略的基础作用。这时，概率测度被用作一种辨别手段、实现对归纳学习中失误风险的控制。这种作用在第二章介绍的 PAC 可学性理论中就显示在置信参数 δ 和 精确度参数 ϵ这两个概率参数上。它们协助建模者尽量规避先验逻辑法则未经证实的泛化程度，同时尽量减小后验学习归纳过程中的失误程度。具体地，它们协助建模者通过试验找到偏差-复杂度间的最佳权衡决策点。为此，建模者需要对训练样本和测试样本的估计结果、特别是残差项做详尽的诊断性分析。Seeger (2006) 将上述学习过程简称为“诊断性范式”。这一范式应该是经济计量学的研究策略转移的主方向。从计量学科的研究现状看，浏览文献中发表的应用建模结果，对于数据信息低效和误导的分析比比皆是。因此，学科的主策略转向问题并不是一个纯方法论的问题，而是系统提高应用研究水准的唯一出路。

下面我们就针对如何借鉴机器学习方法克服应用计量研究中的通病、改善研究水平的问题做进一步分析。让我们先从微观计量应用建模入手。微观计量应用研究中一个普遍存在的问题是模型欠拟合，且欠拟合程度往往随着数据量的增加而加剧。这一现象恰恰是前述有关基于逻辑推理的微观经济关系式之脆弱性特征的集中反映。其实，应用经济学家对建模中一些常见的并发杂症早有意识。例如，微观主体行为间的显著奇质性、理论关系式中忽略的因素或自变量在数据上的不可分离性、方程式内输入变量间相互作用形式的不确定性、理论关注自变量可能存在的非线性效应及其不确定性等等。由于这些并发杂症都属于模型的函数估计学习范畴，从以给定模型内的参数估计推断为重心的微观经济计量学教科书中，难以找到对它们系统处理的方略和手段。因此，当应用者转向机器学习的教科书、试用机器学习中常用的函数估计工具来对付上述问题时，轻而易举得出显著超越按主流经济计量学思路建模的结果也就顺理成章了，如见 Bajari *et al* (2015)。从本质上看，出现模型系统地欠拟合意味着模型设定思路过于简单。鉴此，机器学习倡导在模型函数选择时采用尽量灵活的近似函数形式。由于经济学的基本法则一般可表述为凸函数学习任务，一种处于首选的假设模型类是广义加性模型 (generalised additive model (GAM)) [29]。以下便是一个含两个解释变量的 GAM 例子：

(3.3.1) $y = \alpha + f_1\left[\sum_{j=1}^{k} b_{1j}(x_1, x_2)\right] + f_2\left[\sum_{j=1}^{k} b_{2j}(x_1, x_2)\right] + \cdots + f_p\left[\sum_{j=1}^{k} b_{pj}(x_1, x_2)\right] + \varepsilon$

其中的$f_i\left[\sum_{j=1}^{k} b_{ij}(x_1, x_2)\right]$ 表示线性可分离的输入函数，它们的函数形式可以互不相同；每个输入函数都是由某种基表示 (basis representations) 因素 $b_{ij}(x_1, x_2)$ 组成的。这些基表示的形式多种多样，它们可以是解释变量本身，也可以是两个变量的比值或者它们的差或积，也可以是单个变量的多项式、或者是基于某变量的核函数。这些输入函数的设计也被通称为‘特征设计’。无庸赘述，特征设计对于理论假说的验证至关紧要。为了实现与各个输入基表示相关的参数的可解释性，并确保它们的泛化性，特征设计必须参照先验知识、细致地将所有有关的因果法则公式化表出。在公式化过程中，特别需要考虑的是，如何使基表示的设计尽量满足各个基表示之间具有数据变动之独立可分离性的要求。显然，对于f_i的分离设计，是为了应对微观主体行为差异显著所

[29] 教科书有关 GAM 的讲解，可参见 Hastie *et al* (2009, Chs 5 & 9)。在经济计量学中，GAM 其实并不是前所未闻的模型类。例如，在 Cameron and Trivedi (2005, Ch9) 中就有对该模型的简述。遗憾的是， 计量学教科书忽略了这类模型的统计学习方法的基础，把它简单归类于非参数估计方法的范畴了。

反映出的数据分层。f_i的分离一般是通过各种随机树或决策树分类算法来实现的。在经济计量学教科书中，这类算法被归为半参数法。不过必须明确的是，这类算法的首要目的是函数估计，而不是计量教科书所强调的参数估计。还须看到，由后验学习得出的f_i的个数，p，是与基表示的设计密切相关的。就样本量大、模型函数学习情况复杂的课题而言，基表示的设计越简单，p值就很可能会越大。可见，基表示的设计与树分类这两个任务是需要通过交替迭代过程来完成的。这一学习过程集中体现着先验专业知识与后验数据信息的密切互动。正因如此，特征设计也被归类为‘特征学习’，如见 Shalev-Shwartz and Ben-David (2014, Ch25)，并且在‘特征表示学习’（及其缩写‘表示学习’）题目下受到专门研究关注，如见 Goodfellow *et al* (2016)。总之，GAM 建模思路及其背后的 PAC 可学性理论为系统改善传统的微观计量应用研究中普遍存在的模型欠拟合现象提供了一条有望可行路径。

下面我们再来看一下宏观应用计量建模研究方面的问题。这里仅以测度长期均衡关系的研究问题为例。从某种角度说，该研究问题在宏观计量模型研究中占据核心位置。通用建模路径有两条。一条是协整分析路径，另一条则是伦敦经济学院派所推出的由一般到具体的动态建模路径，参见韩德瑞和秦朵（1998）。前一条路径遵循以估计法为重心的计量学理念，属于经济计量学研究的主流路径。后一条路径则与机器学习方法有着不少共性。从目前文献中发表的大量应用案例看，沿后一条路径得出的模型结果通常要比沿前一条路径得出的结果要更具泛化性和精确性。协整分析的数学推证虽然精致严谨，但应用模型结果往往不是欠拟合就是过拟合。从机器学习的视角，我们就不难看清这两条路径的实践差距背后的原因。由于协整分析起源于误差修正模型[30]，我们就来从误差修正模型入手。

现设 ARDL 式 (3.1.3)为经数据验证是动态设定适当的模型。由于该式中的输入变量之间通常具有显著的共线性，相应参数的解释力很差。通常的做法是通过参数转换将 ARDL 模型转换为误差修正模型， 如下式：

$$\Delta y_t = \beta_0 \Delta x_t + (\alpha - 1)\left[y - \frac{\beta_0+\beta_1}{1-\alpha}x\right]_{t-1} + \varepsilon_t,\ (\alpha - 1) < 0 \tag{3.3.2}$$

$$= \beta_0 \Delta x_t + \gamma[y - \kappa x]_{t-1} + \varepsilon_t, \qquad \kappa = \frac{\beta_0+\beta_1}{1-\alpha}$$

式中的β_0可被解释为测度x_t的短期冲击效应，κ可被解释为测度该变量的长期均衡效应，亦即理论假说：$y = \kappa x$之效应，γ ($\gamma = (\alpha - 1) < 0$) 则可被解释为测度该理论关系偏移：$[y - \kappa x]_{t-1}$的负反馈动态效应。与 (3.1.3) 比较，(3.3.2)显著更强的参数解释力其实来源于误差修正模型的变形标准，即尽量使输入因子、亦即基表示的设计满足数据变动之独立可分离性的要求。就模型对于长期均衡关系的测度而言，有必要强调两点：一是参数κ是 ARDL 式 (3.1.3) 中所有参数的函数。这意味着，一般来说，它取决于经特定场合下数据学习得出的动态模型的特定阶数，ARDL(p_1, p_2)。若按未经学习而先验假定的阶数来估计κ 必然会使估值精度受到损失。二是当长期均衡关系中解释变量个数较多时，由于难以避免解释变量间的共线性，不可能单靠统计估计法来精确得出各变量的长期均衡效应。处理解决这一难题的可行路径是参数校准试验法。在试验过程中，所需考虑的不仅有负反馈参数γ，还包括所有相关的短期效应参数及其表述的经济特征。

[30] 有关协整分析如何从伦敦经济学院动态建模应用研究中的误差修正模型成果衍生而来的历史，可参见 Qin (2013, Ch4)；有关伦敦经济学院动态建模计量学派的介绍，还可参见 Gilbert (1986) 。

相对而言，协整分析的关注面要窄得多，它只考虑如何在先验给定动态模型结构条件下得出κ的统计最优估计量。为此，协整分析需要对长期关系所涉变量的动态特征做出明确的先验假设，即设它们拥有单位根、属于非平稳时序过程。协整估计的主要途径有二。一是 Engle-Granger 两步法。该方法因应用简便而广受青睐。在 Engle-Granger 两步法中，以下静态式中：

(3.3.3) $$y_t = kx_t + e_t, \Rightarrow e_{t-1} = [y - kx]_{t-1}$$

k的普通最小二乘法 OLS 估计量被证明是 κ 的一致估计量。协整估计的另一条途径是基于 VAR 模型的 Johansen 步骤。必须看到，这两种估计法都要求协整关系所涉的单个非平稳变量是滞后阶数齐次的变量。对于现实样本数据而言，这一要求显然过强。我们在 3.1 节业已阐明，基于边缘分布和联合分布的应用模型一般欠缺数据相合力。另外，静态模型由于忽略时序中的动态信息而严重欠拟合。在有限时序样本面前，一定规模的 VAR 模型又往往过拟合。因此，以这两种模型为前提的κ 的估计很难是满足与数据相合条件的最优估计。相比之下，由伦敦经济学院派的动态建模路径得出的 (3.3.2) 不仅是构建在条件模型基础上，而且其由一般到具体的动态建模约化思路也利于模型的学习，参见韩德瑞和秦朵 (1998, Ch9)。具体地，先验初始模型类 ARDL$(p_1, p_2, \cdots, p_n)$ 及其后验转型的误差修正模型都不含对任何单个变量的单位根假定，也不含对它们滞后结构的齐次性假定。另外，由于经建模约化路径后验选择出来的与数据吻合、且最为简洁的误差修正模型既避免了欠拟合又避免了过拟合，从这种模型导出的长期参数 κ 的估值，显然要比由协整分析得来的估值更为精确。

在经济计量学界，伦敦经济学院学派以其推崇模型选择步骤和注重各种诊断性检验方法见长。虽然该学派也倡导对数据生成过程的关注，但是在关注的背后，其动态建模思路和体系都与机器学习的辨别式建模路径极其相似。鉴于沿用该学派方法的应用模型研究所取得的成效，也鉴于经济模型应用研究中至今罕见纯属生成式的概率模型，我们实无理由再继续维系所有经济变量的基础生成机制是一联合随机分布机制这一幼稚幻想了。机器学习中的成功经验启迪我们，实践可行的经济建模研究应该系统转向以无分布的辨别式方法论为重心的研究。相应地，概率论应被主要用于协助模型选择的分辨工具，而不是被用作经济计量学直接套用经典统计学工具、对先验模型进行简单估计推断的理论说辩修辞凭据。

References

Abu-Mostafa, Y. S., M. Magdon-Ismail, and H.-T. Lin (2012) *Learning From Data*, AMLBook.

Bajari, P., D. Nekipelov, S. P. Ryan, and M.-Y. Yang (2015) Machine Learning Methods for Demand Estimation, *American Economic Review*, 105(5): 481-85.

Bateman, B. W. (1990) Keynes, Induction, and Econometrics, *History of Political Economy*, 22(2): 359-79.

Bottou, L. (2014) From machine learning to machine reasoning, *Machine Learning*, 94: 133-49.

Bradley, R. and M. Drechsler (2014) Types of Uncertainty, *Erkenntnis*, 79(6):1225-48.

Breiman, L. (2001) Statistical modeling: The two cultures (with comments and a rejoinder by the author), S*tatistical Science*, *16*(3), 199–231.

Brodeur, A., M. Lé, M. Sangnier and Y. Zylberberg (2016) Star Wars: The Empirics Strike Back, *American Economic Journal: Applied Economics*, 8(1): 1-32.

Cameron, A. C. and P. K. Trivedi (2005) *Microeconometrics: Methods and Applications*. Cambridge: Cambridge University Press.

Canova, F. (2009) How much structure in empirical models? in T. Mills and K. Patterson eds. *Palgrave Handbook of Econometrics, Volume II: Applied Econometrics*, pp.68-97.

Chapelle, O., B. Scholkopf and A. Zien eds. (2006) *Simi-Supervised Learning*, MIT Press.

Charpentier, A., E. Flachaire and A. Ly (2018) Econometrics and Machine Learning, *Economie et Statistique*, Institut National de la Statistique et des Études Économiques (INSEE), issue 505-506, 147-169.

Christiano, L. J., M. S. Eichenbaum and M. Trabandt (2018) On DSGE models, *Journal of Economic Perspectives*, 32(3), 113-40.

Costa, P., A.-L. Jousselme, K. Laskey, E. Blasch, V. Dragos, et al. (2018) URREF: Uncertainty representation and reasoning evaluation framework for information fusion. *Journal of Advances in Information Fusion*, ISIF, 13(2): 137-57.

Cox, D. (1992) Causality: Some statistical aspects. *Journal of Royal Statistical Society A*, 155: 291-301.

Cox, D. (2006) *Principles of Statistical Inference*, Cambridge University Press.

Davidson, P. (1991) Is probability theory relevant for uncertainty? A post Keynesian perspective, *Journal of Economic Perspectives*, 5(1): 129-43.

Davis, R., H. Shrobe and P. Szolovits (1993) What is a Knowledge Representation? *AI Magazine*, 14(1):17-33.

Dawid, A. P. and A. Tewari (2020) On Learnability under General Stochastic Processes, *arXiv preprint arXiv:2005.07605*.

Deaton, A. and N. Cartwright (2018) Understanding and Misunderstanding Randomized Controlled Trials, *Social Science and Medicine*, 210: 2-21.

Doyle, J (1992) Rationality and its Roles in Reasoning, *Computational Intelligence*, 8(2): 376-409.

Dubois, D. and H. Prade (2009) Formal representation of uncertainty, in D. Dubois, M. Pirlot and H. Prade eds., *Decision-Making Process-Concepts and Methods*, ISTE and Wiley, pp. 85-156.

Efron, B. (2011) The future of Statistics, in M. Lovric ed., *International Encyclopedia of Statistical Science*. New York: Springer, pp. VII-X.

Efron, B. and T. Hastie (2017) *Computer Age Statistical Inference: Algorithms, Evidence and Data Science*, Cambridge University Press.

Eichenbaum, M. (1995) Some Comments on the Role of Econometrics in Economic Theory, *The Economic Journal*, 105(433): 1609-21.

Fernández-Villaverde, J.F., J. Rubio-Ramírez and F. Schorfheide (2016) Solution and Estimation Methods for DSGE Models, in J. B. Taylor and H. Uhlig eds., *Handbook of Macroeconomics* Vol. 2, Elsevier, pp. 527-724.

Friedman, J. H. (1994) An overview of computational learning and function approximation, in V. Cherkassy, J. H. Friedman, H. Wechsler, eds. *From Statistics to Neural Networks, Theory and Pattern Recognition Applications*, NATO/ASI Workshop, Springer-Verlag, pp. 1-55.

Friedman, J. H. (1997) Data mining and statistics: What's the connection? Keynote Address in *Proceedings of the 29th Symposium on the Interface Between Computer Science and Statistics*.

Gammerman, A., V. Vovk and V. Vapnik (1998) Learning by transduction, *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pp. 148-55.

Gilbert, C.L. (1986) Professor Hendry's econometric methodology, *Oxford Bulletin of Economics and Statistics*, 48: 283-307.

Gilboa, I., A. Postlewaite, L. Samuelson and D. Schmeidler (2014) Economic Models as Analogies, *The Economic Journal*, 124(578): F513-33.

Goodfellow, I., Y. Bengio and A. Courville (2016) *Deep Learning*, MIT Press.

Griliches, Z. (1961) Hedonic price indexes for automobiles: an econometric analysis of quality change, in Price Statistics Review Committee ed., *The Price Statistics of the Federal Government*, National Bureau of Economic Research, Cambridge, pp173–196; reprinted in: Qin, D. ed. (2013) *The Rise of Econometrics*, vol. 3, London, Routledge, pp 96-124.

Gürkaynak, R.S., B. Kısacıkoğlu and B. Rossi (2013) Do DSGE models forecast more accurately out-of-sample than VAR models? *Advances in Econometrics, VAR Models in Macroeconomics – New Developments and Applications: Essays in Honor of Christopher A. Sims*, 32: pp. 27-79.

Haavelmo, T. (1944) The Probability Approach in Econometrics, *Econometrica*, 12, supplement. 中文版：哈维尔莫，经济计量学的概率论方法，商务印书馆 (1994)。

Hand, D. J. (1994) Deconstructing statistical questions, *Journal of Royal Statistical Society* A, 157(3): 317-56.

Hastie, T., R. Tibshirani and J. Friedman (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd Edition, Springer.

Hempel, C.G. (1965) *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science Aspects of Scientific Explanation*, The Free Press.

Hendry, D. F. (1995) *Dynamic econometrics*, Oxford: Oxford University Press.

韩德瑞和秦朵（1998）动态经济计量学，上海人民出版社。

Hendry, D. F. and J.-F. Richard (1982) On the formulation of empirical models in dynamic econometrics, *Journal of Econometrics*, 20: 3-33.

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

Hill, R. J. (2013) Hedonic price indexes for residential housing: A survey, evaluation and taxonomy, *Journal of Economic Survey*, 27(5): 879-914.

Iskhakov, F., J. Rust and B. Schjerning (2020) Machine learning and structural econometrics: contrasts and synergies, *Econometrics Journal*, 23(3): 81-124.

Janssen, M. C. W. (1993) *Microfoundations: A Critical Inquiry*, Routledge.

Jebara, T. (2004) *Machine Learning: Discriminative and Generative*, Springer.

Judge, G. G., W. E. Griffiths, R. C. Hill and T.-C. Lee (1980) *The Theory and Practice of Econometrics*, John Wiley and Sons.

Kardaun, O. J. W. F., D. Salomè, W. Schaafsma, A. G. M. Steerneman, J. C. Willems and D. Cox (2003) Reflections on Fourteen Cryptic Issues Concerning the Nature of Statistical Inference, *International Statistical Review*, 71: 277-303.

Koopmans, T. C. (1950) When is an equation system complete for statistical purposes? in Koopmans ed., *Statistical Inference in Dynamic Economic Models*, Cowles Commission Monograph 10, John Wiley.

Mäki, Uskali ed. (2002) *Fact and Fiction in Economics*, Cambridge University Press.

Marcellino, Massimiliano (2006) Leading Indicators, in G. Elliott, C.W.J. Granger and A. Timmermann eds., *Handbook of Economic Forecasting*, Chapter 16, Elsevier, pp.879-960.

McCann, C. R. (1994) *Probability foundations of Economic Theory*, London: Routledge.

Mukherjee, S., P. Niyogi, and T. Poggio (2006) Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization, *Advances in Computational Mathematics*, 25(1-3): 161-93.

Mullainathan, S. and J. Spiess (2017) Machine Learning: An Applied Econometric Approach, *Journal of Economic Perspectives*, 31(2): 87-106.

Nakamura, E. and J. Steinsson (2018) Identification in Macroeconomics, *Journal of Economic Perspectives*, 32(3): 59-86.

Pesaran, M. H. (1987) Global and Partial Non-Nested Hypotheses and Asymptotic Local Power, *Econometric Theory*, 3(1): 69-97.

Poggio, T., R. Rifkin, S. Mukherjee and P. Niyogi (2004) General conditions for predictivity in learning theory, *Nature*, 428: 419-22.

Qin, D. (1993) *The Formation of Econometrics: A Historical Perspective*, Oxford: Clarendon Press.

Qin, D. (2013) *A History of Econometrics: The Reformation from the 1970s*, Oxford University Press.

Qin, D. (2014) Inextricability of confluence and autonomy in econometrics, *Oeconomia*, 4(3): 321-41.

Rodrik, D. (2015) *Economics Rules: The Rights and Wrongs of The Dismal Science*, W. W. Norton & Company.

Rowley, R. and O. Hamouda (1987) Troublesome probability and economics. *Journal of Post Keynesian Economics* **10**: 44-64.

Russell, S. and P. Norvig (2016) *Artificial Intelligence: A Modern Approach*, 3rd Edition, Pearson.

Sargent, T. and C.A. Sims (1977) Business cycle modeling without pretending to have too much a priori economic theory, in *New Methods in Business Cycle Research: Proceedings from a Conference*, Federal Reserve Bank of Minneapolis, pp 45-109.

Sbordone, A., A. Tambalotti, K. Rao, and K. Walsh (2010) Policy Analysis Using DSGE Models: An Introduction. *Economic Policy Review*, 16(2): 23-43.

Seeger, M. (2006) A taxonomy of semi-supervised learning methods. in O. Chapelle, B. Schölkopf, and A. Zien, eds., *Semi-Supervised Learning*, MIT Press.

Shalev-Shwartz, S. and S. Ben-David (2014) *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press.

Shalev-Shwartz, S., O. Shamir, N. Srebro and K. Sridhara (2010) Learnability, Stability and Uniform Convergence, *Journal of Machine Learning Research*, 11: 2635-70.

Sloman, S. and P. Fernbach (2017) The Knowledge Illusion: Why We Never Think Alone, Riverhead Books.

Stanley, T. D. (1998) Empirical economics? An econometric dilemma with only a methodological solution, Journal of Economic Issues, 32(1): 191-218.

Stigum, B. P. (2003) *Econometrics and the Philosophy of Economics*, Princeton University Press.

Triplett, J. (2004) *Handbook on hedonic indexes and quality adjustments in price indexes: special application to information technology products*, Directorate for Science, Technology and Industry Working Paper 2004/9, OECD.

Valiant, L. (1984) A theory of the learnable, *Communications of the ACM*, 27(11): 1134-42.

Valiant, L. (2000) Robust logics, *Artificial Intelligence*, 117(2): 231-53.

Valiant, L. (2008) Knowledge infusion: In pursuit of robustness in artificial intelligence, in R. Hariharan, M. Mukund and V. Vinay eds., *Foundations of Software Technology and Theoretical Computer Science*, Bangalore, pp 415-22.

Valiant, L. (2013) *Probably Approximately Correct: Nature's Algorithms for Learning and Prospering in a Complex World*, Basic Books.

van Engelen, J.E. and H. H. Hoos (2020) A survey on semi-supervised learning, *Machine Learning*, 109: 373-440.

Vapnik, V. (1999) *The Nature of Statistical Learning Theory*, 2nd edition, Springer.

Vapnik, V. (2003) An overview of statistical learning theory, *NATO Science Series Sub Series III Computer and Systems Sciences*, 190: 1-28.

Varian, H. R. (2014) Big data: New tricks for econometrics, *Journal of Economic Perspectives*, 28(2): 3-28.

Wold, H. O. A. (1975) From hard to soft modelling, in H. O. A. Wold ed., *Modeling in Complex Situations with Soft Information*. Research Report, University Institute of Statistics, Uppsala, Chapter 1.

Wold, H. O. A. (1980) Model construction and evaluation when theoretical knowledge is scarce: Theory and application of partial least squares, in *Evaluation of econometric models*, Academic Press, pp. 47-74.

Zimin, A. and C. Lampert (2017) Learning Theory for Conditional Risk Minimization, in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 213-22.