Chapter 12
Machine Learning and Model Selection

Agoston Reguly, Esfandiar Maasoumi and László Mátyás

Abstract The chapter provides an extensive historical overview of the methods developed in machine learning, with special emphasis on their relation to econometric practice and model discovery with observational data. The historical line given is not meant to be exhaustive, but rather serve to illuminate how ideas in econometrics and other statistical fields are finding their way to provide a better understanding and appreciation of modern machine learning in a context that emphasize policy evaluation and decision making with uncertain models.

12.1 Introduction

Evidence-based discovery of mechanisms involves at least three components, observation, theorizing, and evaluation. In observational (non experimental data) studies, there are at least two types of uncertainties, sampling (inductive) and model uncertainty. These are interactive and complex. Samples primarily inform about model objects, model objects are abstractions for latent, meaningful real life objects and policy goals.

Statistical/probability approach to inductive inference has, traditionally, emphasized sampling uncertainty, often through test statistics and confidence intervals. Model uncertainty has been generally less well attended to. Machine learning (ML) is viewed here as the latest set of tools that may address this relative neglect, but it requires greater care in sampling inference costs and uncertainty.

Ágoston Reguly ✉
Corvinus University of Budapest, Budapest, Hungary and Georgia Institute of Technology, Atlanta, Georgia, USA, e-mail: agoston.reguly@uni-corvinus.hu

Esfandiar Maasoumi
Emory University, Atlanta, Georgia, USA, e-mail: emaasou@emory.edu

László Mátyás
Central European University, Budapest, Hungary, and Vienna, Austria, e-mail: matyas@ceu.edu

We offer an overview of histories of ML and sampling inference, including model selection and averaging. This provides a constructive setting for investigation of a common paradigm in prediction and policy analysis based on empirical models. There is a basic conception of joint relations between target sets of variables, represented by a statistical model. There is also, however, uncertainty about 'nuisance' covariate effects in a less than perfect experimental setting, in terms of both which covariates and what functional effects they exert on identification of desired objects, such as treatment effects and elasticities etc. Statistical properties of samples and models provide an additional and related layer of difficulty and approximation that require careful examination.

Modern ML in social sciences, especially in Econometrics, is best seen as a set of developing tools and approaches that aspire to find inferences and predictions that are 'robust' to inevitable approximations and specification errors. All models should be considered as misspecifications of a 'true' Data Generation Process (DGP). Does ML help to alleviate the challenge of robust inference under different degrees of misspecification (local and global and in between)? We find a qualified affirmative answer, without ignoring the ability of ML techniques to expand the feasible model set and the associated increased cost of risk management. Formal risk management is not optional since model uncertainty persists under ML with many different flexible approximations may do well as assessed by 'fit', with widely different structural and policy explanations and mechanisms.

There are at least two fundamental objectives to learning from the data: prediction and inference about mechanisms and relations. The distinction may be represented by different loss and penalty functions, but is somewhat artificial since actionable prediction requires reliable mechanism discovery. Modern potential outcome (treatment effect) literature makes this historically recognized inter-relation quite explicit.

We offer in this chapter a history of ML and model selection in econometrics which has traditionally emphasized limits of observational samples and models. From an early and ongoing emphasis on predictive performance in ML, we proceed to modern 'causal inference' emphasizing robustness and model uncertainty, as exemplified with Automated Debiased ML. We try to deal with the many unifying interpretations of 'penalization', 'regularization', 'constrained optimization', 'shrinkage', Bayesian learning and formalisms, pretesting, cross validation, counterfactual and source decompositions.

12.2 Brief History of Machine Learning up to 2000s

A common starting point for machine learning's intellectual lineage is McCulloch and Pitts (1943), who proposed an explicitly mathematical abstraction of a biological neuron as a binary threshold device: given inputs x and weights w (in econometrics we call them coefficients), the unit outputs $y = \mathbf{1}(w'x \geq c)$, which is essentially a

deterministic latent-index model.[1] By showing that networks of such threshold units can implement logical operators and complex logical expressions, McCulloch and Pitts (1943) offered an early precursor to what we would now call a flexible functional form: composing simple nonlinearities to approximate complicated mappings. Although their model provided an early bridge between neurophysiology and computation, the general idea later resurfaced as the conceptual basis for artificial neural networks and learnable function approximators. A complementary milestone is Turing's (1950) essay, which reframed 'machine intelligence' into an operational, performance-based criterion (also known as the 'Turing test' or 'imitation game'). Turing's proposal helped legitimize evaluation by observable task performance rather than internal mechanism. The research agenda of artificial intelligence has been judged by the evaluation of intelligent behavior via observable performance since. This perspective is close to the econometric tradition of model evaluation, where models are judged by how well they perform under a specified loss function. The first explicit learning machines emerged in the late 1950s. Rosenblatt's perceptron marked a decisive shift from representational neuron models toward *trainable* parametric systems. Developed in 1957 at the Cornell Aeronautical Laboratory, the perceptron was introduced as a perceiving and recognizing automaton and then formalized as a probabilistic model for information storage and organization in the brain (Rosenblatt, 1958). In modern terms, the perceptron is a linear classifier $\text{sign}(w'x)$ whose parameters are updated using labeled data (x_i, y_i). This is best viewed as empirical risk minimization with a classification loss (a cousin of least squares, but with a loss tailored to 0–1 outcomes), implemented via a simple stochastic approximation rule: weights are adjusted in the direction that reduces in-sample misclassification, using one observation (or a small batch) at a time.[2] Later theoretical results clarified that a single linear separator cannot fit nonlinearly separable patterns (see, e.g., Minsky & Papert, 1969 or Cover, 1965), but the perceptron established two enduring themes that reappear throughout machine learning: (i) data-driven parameter updating framed as the optimization of a sample criterion, and (ii) the idea that 'learning' can be implemented by local computations, which later became central for multilayer networks and gradient-based training.

During the 1960s, 'learning from data' increasingly meant operational statistical procedures for prediction and classification: linear decision rules (estimated from labeled samples) alongside nonparametric methods that avoided strong parametric or distributional commitments when explicit generative models were hard to justify. This period helped standardize a workflow that remains recognizable today: (i) specify a model class or hypothesis space (e.g., linear separators, nearest-neighbor rules), (ii) fit or tune its free components on observed data (often by minimizing an empirical loss), and (iii) compare alternatives by empirical performance. This is done in particular using error rates or expected loss on new observations, which were not used. Early syntheses in statistical pattern recognition made this explicit

[1] One can see this as an early limiting case of a binary response model where the disturbance is suppressed, unlike in the probit or logit model, where $y = \mathbf{1}(w'x + \varepsilon \geq 0)$.

[2] This procedure is conceptually very similar to the recursive least squares or Robbins–Monro-type updating, see Robbins and Monro (1951).

and placed linear discriminant methods next to distribution-free classification rules as complementary tools for supervised learning (Duda & Hart, 1973). During the late 1960s and early 1970s the ideas of nearest neighbors and early backpropagation emerged. A canonical milestone for nearest neighbors is Cover and Hart (1967), who analyzed the 1-nearest-neighbor (1-NN) rule, which classifies a new point by the label of the closest training observation. Cover and Hart (1967) showed that, in large samples, its classification error is is bounded above by twice the Bayes error under mild regularity conditions. Put differently: if an 'oracle' classifier that knows the true data-generating distribution achieves the minimal attainable error rate (the Bayes benchmark), then the extremely simple 1-NN rule achieves an error rate within a factor of two of that oracle in the asymptotic regime. Beyond nearest neighbors, the 1960s also produced core nonparametric and unsupervised primitives that later reappear throughout machine learning. Kernel regression (the Nadaraya–Watson estimator Nadaraya, 1964; Watson, 1964) smooths the conditional mean by locally weighted averages, providing a continuous analogue to kNN.[3] Chapter 11 discusses more in detail the nonparametric literature. Furthermore, during this decade in unsupervised learning MacQueen (1967) employs k-means clustering, which offered an efficient prototype-based partitioning rule, while the EM algorithm soon provided a general-purpose likelihood maximization scheme for latent-variable and mixture models (Dempster, Laird & Rubin, 1977) used in statistics and (partially) in economics as we discuss it in Section 12.9. Another statistical thread in the early 1970s introduced explicit regularization for prediction under multicollinearity. Ridge regression from Hoerl and Kennard (1970) added an L2 penalty[4] to least squares, shrinking coefficients toward zero to trade bias for variance and improve out-of-sample performance. Aligned with this line of research Mallows (1973) proposed C_p for subset regression, an early complexity-penalized criterion in the Gaussian linear model. Stone (1974) proposed an alternative to in-sample fit penalization the method of cross-validation as a general criterion for choosing and assessing predictors, that we will discuss in Section 12.5 in detail. Let us mention here that another solution to model selection during this era was proposed by Akaike (1973) and Schwarz (1978), who derived information criteria (AIC and BIC) that penalize likelihood by model dimension, anticipating modern complexity control. This strand of literature is more familiar in economics as a model selection technique, therefore we discuss origins more in Section 12.4. In parallel, during the early 1970s, there were key steps toward what later became known as 'backpropagation', the central workhorse for training deep networks. Backpropagation is an efficient way to compute the gradient of a model's loss with respect to all weights (w) in a multilayer network by repeatedly applying the chain rule from the output layer back to earlier layers (see details in the Appendix). Lastly, it is worth mentioning that the decade of 1970s also exposed limits of the era's hardware and algorithmic scalability: most prominently the Lighthill, 1973 report argued that many ambitious machine learning goals were not meeting

[3] This method is the starting point of modern regression discontinuity design analysis, with local polynomial estimators. RDD combined with modern ML tools, can provide heterogeneous treatment effect estimation, see Reguly (2025).

[4] An L2 penalty adds a term like $\lambda \sum_j \beta_j^2$ to the loss function.

promised progress due to inability to scale up to 'real' problems, contributing to major funding pullbacks in the UK and helping precipitate the first 'AI winter' (Agar, 2020).

After the funding squeeze of the 1970s, the 1980s witnessed a renewed and increasingly self-conscious focus on *learning* as a topic in its own right, with a distinct research community, venues, and vocabulary that gradually differentiated *machine learning* from the broader (and often more knowledge-engineering–oriented) 'AI agenda'. In particular, many researchers interested in computational issues initiated dedicated workshops and started the journal *Machine Learning* in 1986, which helped crystallize the field's identity and standards for empirical evaluation and reproducibility (Langley, 2011; Flach, 2011). Decision-trees became the decade's most influential practical learners, a flagship approach due to its interpretability and its ability to learn structured rules from tabular data. Quinlan (1986) introduced the 'ID3 system', formalized top-down induction using information gain and demonstrated how decision-trees could be adapted to noisy and incomplete information. This property made tree-based methods attractive for early expert-system and knowledge-acquisition pipelines. Another variant of tree-based methods is, the classification and regression tree (CART) framework pioneered by Breiman, Friedman, Olshen and Stone (1984). CART can be systematized as a statistical procedure built around impurity-based splits, cost-complexity pruning, and honest assessment of predictive risk, establishing the template that underlies much of modern tree-based learning such as random forest (Breiman, 2001) and its causal versions in economics (Wager & Athey, 2018; Athey, Tibshirani & Wager, 2019). A second, highly visible axis of the 1980s revival was the rediscovery and dissemination of efficient gradient-based training for multilayer networks. The paper by Rumelhart, Hinton and Williams (1986), played a decisive role in popularizing backpropagation as a practical procedure for adjusting weights to minimize output error and for learning useful internal ('hidden') representations. This methodological consolidation quickly fed into application-driven demonstrations. Most notably, LeCun et al. (1989) showed how architectural constraints aligned with the vision task (precursors to convolutional ideas such as locality and weight sharing) could be combined with backpropagation to yield strong performance on real handwritten digit data from the U.S. Postal Service. Together, these developments established a durable template for modern neural-network practice: scalable gradient-based optimization, representation learning in intermediate layers, and task-informed inductive biases in network architecture, alongside the parallel rise of interpretable tree-based induction in mainstream machine learning.

In the 1990s, machine learning increasingly embraced a statistical and explicitly *probabilistic* framing: models were evaluated by predictive performance on data, and uncertainty was represented and manipulated directly through probability distributions. This shift is evident in the rise of probabilistic graphical models (especially Bayesian networks) as a unified language for encoding conditional independencies, handling missing data, and combining prior knowledge with evidence (see for a great overview, Heckerman, 1995). At the same time, sequential probabilistic models such as hidden Markov models (HMMs) became widely adopted in speech and sequence recognition, supported by clear mathematical foundations and effective algorithms

for inference and learning (see, e.g., Rabiner, 1989). During this era ML started to face high-dimensional problems, where the number of potential variables is high. A milestone for handling such setup was the least absolute shrinkage and selection operator (LASSO) from Tibshirani (1996), which modified the ridge method of Hoerl and Kennard (1970) and used an L1 penalty, by using absolute values instead of squared.[5] This change allowed to shrink coefficients exactly to zero, turning variable selection into convex optimization and anticipating sparse learning, when there are (much) more variables than observations. Flexible regression learners also expanded beyond trees and kernels. Multivariate adaptive regression splines (MARS) by Friedman (1991) automated the construction of spline basis expansions (including interactions) for high-dimensional regression. Whereas Friedman and Stuetzle (1981) proposed the 'projection pursuit regressions', which models $E[Y \mid X]$ as sums of smooth functions of low-dimensional projections, hence avoiding the curse of dimensionality occurring in fully nonparametric setups. In parallel, the broader theoretical consolidation of statistical learning theory helped connect generalization and empirical risk minimization, providing a principled vocabulary for understanding why certain algorithms (and not just representations) generalize well from finite samples (Vapnik, 1995). A defining algorithmic development of the decade was the introduction of *support vector machines* (SVMs) and the max-margin principle: the COLT'92 work by Boser, Guyon and Vapnik (1992) presented an 'optimal margin' training algorithm whose solution depends on a subset of supporting patterns, anticipating the kernelized SVM framework that would soon dominate many high-dimensional classification tasks. Ensemble learning also matured into mainstream practice: Breiman's bagging (bootstrap aggregation) was formulated in Breiman (1996), demonstrating that averaging predictors trained on bootstrap resamples can substantially improve accuracy for unstable base learners. Soon after, boosting methods (particularly AdaBoost, see Freund & Schapire, 1996) showed how iteratively reweighting training examples and combining weak learners could yield strong predictors, with influential experimental and theoretical treatments appearing in the mid-1990s, e.g., by Freund and Schapire (1997). Finally, neural networks reach a notable milestone in the form of 'LeNet-5' and related convolutional architectures. LeCun, Bottou, Bengio and Haffner (1998) synthesized gradient-based learning with architectural inductive biases (local receptive fields, weight sharing, pooling) to achieve document and digit recognition, foreshadowing the later deep learning resurgence.

In the 2000s, machine learning became a core engine of data mining and large-scale web applications, driven by the rapid growth of digital traces (click logs, crawled documents, user-generated content) and the need to turn them into search, ranking, recommendation, and anomaly-detection systems. A defining infrastructure shift was the move toward distributed storage and computation on commodity clusters: Google's File System emphasized fault tolerance and throughput for data-intensive workloads, while MapReduce provided a simple programming abstraction that automatically parallelized data-processing jobs over thousands of machines, enabling routine

[5] Modifying the penalty term to $\lambda \sum_j |\beta_j|$ instead of $\lambda \sum_j \beta_j^2$.

processing of terabytes at web scale (see, Ghemawat, Gobioff & Leung, 2003; Dean & Ghemawat, 2004). This 'systems + statistics' co-evolution also pushed algorithmic research toward scalability: learning procedures whose time and memory footprints grow roughly linearly with data volume, and toward an explicit recognition that optimization accuracy can be traded off against statistical accuracy in the large-data regime (e.g., stochastic optimization can be preferable even when it is a poor optimizer Bottou & Bousquet, 2008). Tree ensembles took a decisive step with random forests proposed by Breiman (2001), which combine bagging with random feature selection at each split to decorrelate trees, yielding strong off-the-shelf performance along with internal validation via out-of-bag error and variable importance measures. On the regularization side, efficient path algorithms made sparse and shrinkage estimators computationally routine. Efron, Hastie, Johnstone and Tibshirani (2004) introduced the least angle regression (LARS), which provided a stagewise model-selection path closely connected to the LASSO. Another advancement was the elastic net from Zou and Hastie (2005). This method combined L1 and L2 penalties to stabilize selection under strong predictor correlation, while allowing more variables than observations, hence effectively combined the strengths of ridge and LASSO methods. Overall, the decade was marked by the broad adoption of kernel methods and support vector machines as versatile, theoretically grounded tools for classification, regression, and dimensionality reduction, supported by mature expositions and a growing catalogue of kernel constructions for structured data (see, synthesis from Schölkopf & Smola, 2002 or Shawe-Taylor & Cristianini, 2004). At the same time, *unsupervised learning* grew in importance as practitioners confronted high-dimensional, weakly labeled, or unlabeled data: manifold learning and nonlinear dimensionality reduction (e.g., Isomap by Tenenbaum, de Silva & Langford, 2000) provided geometric approaches to structure discovery, while probabilistic topic models such as latent Dirichlet allocation (LDA) by Blei, Ng & Jordan, 2003 offered scalable unsupervised representations for massive document collections. Finally, reinforcement learning advanced toward large decision problems via Monte Carlo Tree Search, with UCT algorithm[6] (Kocsis & Szepesvári, 2006), providing a principled exploration vs exploitation rule that helped drive strong performance increase in game-playing systems e.g., in the well-known game of Go (Coulom, 2007). Together, these trends define the 2000s as a period in which ML methods were increasingly judged by their end-to-end utility in real systems by their robustness, throughput, and deployability, alongside with strong generalization performance.

From an econometric perspective, the history above can be read as machine learning steadily turning prediction into a well-posed optimization problem: minimizing explicit loss under finite-sample generalization constraints. In contrast most of econometrics kept a complementary emphasis on identification, interpretability, and inference about economically meaningful parameters. The overlap is where modern empirical practice now lives: regularization and resampling stabilize high-dimensional prediction and make model complexity an object of choice rather than

[6] Upper Confidence bounds applied to Trees algorithm, which treats each node's action-choice like a multi-armed bandit and uses an upper confidence bound rule to balance exploitation (choose what looks best) vs exploration (try what is uncertain).

an afterthought. This sets up the natural bridge to the next section: once models are judged by out-of-sample performance and complexity control, the central question becomes how to select among competing specifications – a problem econometrics has long addressed through hypothesis testing, information criteria, and related selection principles, towards which we turn next.

12.3 Brief History of Model Selection in Econometrics

During the 1950s and early 1960s, model selection in econometrics largely meant specification from theory. Researchers began with behavioral or accounting equations and imposed some prior restrictions (e.g., exclusion restrictions, exogeneity/endogeneity assignments, functional forms, and other structural constraints) before turning to estimation and inference. This structural approach was closely tied to the probabilistic foundations advocated by Haavelmo (1944). The influential Cowles Commission program organized empirical work into three steps: specification, identification, and estimation (Koopmans, 1950). In applied practice, diagnostic checking was often comparatively informal by modern standards. Haavelmo (1944) explicitly notes that empirical work frequently relied on simple summaries such as standard errors and multiple-correlation measures as rough indicators of fit. Within this framework, model selection was based on classical hypothesis testing which provided an operational mechanism for refining specifications. In linear regression applications, t-tests and F-tests were routinely used to assess individual or joint restrictions, guiding inclusion or exclusion decisions for candidate regressors and checking economics motivated constraints (Johnston, 1963; Goldberger, 1964). During the mid-1960s treatments also discuss practical specification concerns (e.g., multicollinearity, specification errors, and stepwise-type procedures), reflecting the emergence of test-based strategies as part of empirical modeling practice (Goldberger, 1964). Alongside tests, a common (though mechanically problematic) heuristic was to pursue higher in-sample goodness-of-fit (especially R^2) as a proxy for the 'best' model (Theil, 1961). The prevalence of such fit-based reasoning was already implicit in early methodological discussions that emphasized multiple correlation and related fit statistics. This later became the target of formal critique in contexts where standard R^2 arguments do not fit purpose (see, e.g., Maddala, 1988; Pesaran & Smith, 1994). In parallel, applied forecasting literature emphasized that models should also be judged by their *predictive* performance and by the accuracy of their forecasts for policy purposes, an agenda synthesized early on by Theil (1961). A notable shift toward more explicitly data-driven specification appeared by the end of the 1960s in some time-series studies. The Box-Jenkins approach (popularized in Box & Jenkins, 1970, see also Chapter 1) promoted iterative model selection, estimation, and diagnostic checking for ARIMA processes, with selection guided by the time-series properties of the data and the behavior of forecast errors rather than structural interpretation.

Shrinkage methods and combined estimators and predictors began to be noted in the 1960s. Through Stein's work and Bayesian learning, optimal model averaging and

selection, and the choice between predictive and estimation criteria came into sharper focus. This relatively thin literature offered some insights on model uncertainty, and the role of widespread pretesting and data snooping, vs use of test statistics as defining potential model sets and partial identification. A priori structural uncertainties were examined in Stein-Like shrinkage estimation in simulatenous equations systems for reduced form estimation (prediction), in Maasoumi (1978). System Ridge penalization and estimation were developed in, for example, Maasoumi (1980), with an early indication of why and how these regularization methods may overcome inadequate sample information. Bayesian MELO estimators and shrinkage methods began to shed light on why mixed forecasts and estimation seemed to do well! Caner, Maasoumi and Riquelme (2016); Hansen (2016); Shi (2016); Drukker and Liu (2022); and Liao (2013) among others, shed further light on the choice of 'tuning parameters' as more than a merely computational device. This work on ML in econometrics extended earlier ideas on the many moments selection literature, especially Andrews and Lu (2001). The optimizer in ML was considered with penalization, possibly for failed moment conditions, with the use of AIC, BIC and shrinkage methods to determine the optimal weights (tuning parameters). This work also clarified the importance of 'asymptotic rates' and orders of magnitude for these tuning parameters, an area of ongoing research with consequence for econometrics risk management. Gospodinov and Maasoumi (2021) provided an explanation and demonstration of higher order model averaging as possible panacea for model uncertainty when all models are misspecified. Gospodinov, Kan and Robotti (2014) provided insightful analyses of degrees of misspecification and showed when and how robust inference was possible. Importantly for ML methods in econometrics Dovonon and Gospodinov (2024) show that including possibly irrelevant covariates in models can invalidate specification and diagnostic tests, indicating spurious model fit and causality! Chudik, Kapetanios and Pesaran (2018) have shown how old fashioned stepwise regression (one covariate at a time) methods can be modified to provide reliable variable selection and significance testing, avoiding the instability and bias of methods such as simple LASSO.

A major conceptual shift in the 1970s was the move from in-sample fit and sequential testing toward *explicit* trade-offs between goodness-of-fit and model complexity, formalized through penalized likelihood criteria. Akaike's information criterion (AIC) proposed selecting the model that maximizes the log-likelihood penalized by the number of free parameters, providing an asymptotically motivated approximation to expected (out-of-sample) information loss and thereby turning model choice into a well-defined optimization problem rather than a sequence of ad hoc tests (Akaike, 1973). In parallel, Mallows' model selection statistic (called C_p) emerged in linear regression as a closely related predictive-statistics device. By correcting the residual sum of squares with a degrees-of-freedom penalty, C_p targets (scaled) mean squared prediction error and supplies a calibrated way to compare subset regressions (Mallows, 1973). These developments made explicit what had previously been implicit in specification work: additional regressors and parameters must improve predictive performance enough to offset increased estimation variance and overfitting risk. Later, Schwarz's Bayesian information criterion (BIC) provided a complementary, large-sample approximation motivated by Bayesian model

comparison as it selects the model that maximizes the log-likelihood penalized by a function of the number of regressors and the number of observations. This yields a criterion that can be interpreted as an asymptotic Bayes factor approximation under mild conditions (Schwarz, 1978). While AIC and C_p are typically oriented toward minimizing expected predictive loss, BIC's stronger penalty emphasizes parsimony and tends to favor smaller models as number of observations grows. These criteria also sharpened an emerging methodological concern: *specification searches* (trying many models and reporting the most favorable) in fact undermine the nominal meaning of classical t- and F-tests by implicitly conditioning on the search process. The Leamer (1978, 1983a) and Lovell (1983) critiques framed these issues as a central problem of empirical practice and argued for making the researcher's objectives explicit via loss functions and for incorporating prior information and sensitivity analysis, rather than relying on 'fishing' through specifications until statistical significance appears. Finally, from this decade, let us mention Stone (1974), who formalized cross-validation as a model-choice tool and evaluation principle by assessing the predictive performance on held-out data, which is discussed in the next section.

By the early 1980s, the econometrics literature had begun to consolidate model selection and specification analysis into a more explicit toolkit rather than a collection of case-by-case practices. A key marker of this consolidation was the *Handbook of Econometrics* edited by Griliches and Intriligator (1983), which assembled survey chapters on identification, estimation, testing, and importantly on model selection. This Handbook treatment made model selection an explicit topic of econometric methodology (not merely a by-product of estimation), and it framed specification as a disciplined process with attention to the inferential consequences of searching over many candidate models. During this time Leamer's influential work (Leamer, 1983a, 1983b) categorized the main theoretical and practical considerations. The core message was that much of applied econometrics rests on nonexperimental data where researchers have substantial discretion over specification, so conventional 'significance' can be misleading because results may hinge on modeling choices rather than on robust information in the data. Leamer (1983a) contrasted the credibility of randomized experimentation with observational work and argued that, without transparency about identification assumptions and specification sensitivity, empirical claims can look more definitive than they truly are. Furthermore as multiple selection procedures become popular in applied work (sometimes contradicting to each other), Leamer (1983b) explicitly catalogued different techniques such as *stepwise regression*, *cross-validation*, and *goodness-of-fit tests*. The growing use of such procedures was facilitated by improved computing and packaged software, which made iterative fitting and comparison of many models feasible in routine empirical work. At the same time, Bayesian approaches gained traction as a coherent way to incorporate prior beliefs and to regularize over-parameterized specifications. Zellner (1985) surveyed Bayesian econometrics explicitly in terms of using prior information for estimation, testing, and prediction and argued that formal Bayesian methods can improve empirical results when prior information is used wisely.

During the 1990s, information criteria such as AIC and BIC became standard, 'default' comparators in many likelihood-based time-series (and increasingly panel)

applications, and they were treated as routine tools for selecting lag length and model dimension in leading multivariate time-series texts used throughout the 1990s (see e.g., Lütkepohl, 1991). Beyond information criteria, forecast-based model comparison became increasingly formalized through tests of equal predictive accuracy. Notably the Diebold and Mariano (1995) framework for comparing competing forecasts under general loss functions and serial dependence, led to a model selection criterion via testing equal predictive accuracy, still used today. Along with the information criteria, the London School of Economics (LSE) tradition associated with David F. Hendry, popularized *general-to-specific* (Gets) modelling as a systematic approach to empirical specification (Hendry, 1995). In this tradition one begins from a relatively general dynamic model intended to capture the salient features of the data, and then simplifies it via statistically tested reductions while monitoring diagnostic adequacy (called 'congruence') and economic interpretability. A flagship statement of this programme is Hendry (1995), which explicitly addresses model discovery, evaluation, and the role of 'data mining' in nonexperimental time-series modelling and illustrates the practical workflow using the `PcGive` software package. Hendry (1995) also helped normalize the view that specification is not a one-shot choice but an iterative reduction-and-checking exercise grounded in a theory of reduction and encompassing, thereby linking model selection to explicit criteria for adequacy and empirical reliability. In the econometric community PcGive evolved into a mature platform for dynamic modelling, with widely used releases in the mid-1990s (e.g., PcGive Professional 8.0) and accompanying documentation, reflecting how routine empirical practice increasingly involved iterating across candidate specifications (Judge & Harris, 1995). By the late 1990s, the software ecosystem expanded toward the integrated `OxMetrics/PcGive` family, which supported structured modelling, diagnostics, and model selection across time-series and panel settings, facilitating increasingly automated reductions that later culminated in dedicated Gets algorithms such as `PcGets` (Doornik & Hendry, 1998; Krolzig & Hendry, 2000).

A further selection challenge during this period is related to macro time-series applications with parameter instability. Bai and Perron (1998) develop estimation and tests for multiple structural breaks and procedures to determine the number of changes, treating break selection as part of specification search. Apart from these advancements, in cross-sectional modeling a new approach emerged with *shrinkage-based* selection procedures designed to curb overfitting when many regressors are available. A leading example is the LASSO (Tibshirani, 1996), which performs variable selection by adding an L1 penalty[7] that shrinks many coefficients exactly to zero while stabilizing estimation in high-dimensional or collinear designs (see also Chapter 4).

In parallel, Maasoumi's information-theoretic synthesis highlighted how model comparison and prediction can be framed in terms of entropy/Kullback-Leibler notions of information loss, thereby encouraging approaches that explicitly balance fit against complexity under pervasive model uncertainty (Maasoumi, 1993). Finally, let us mention that during this decade, the $(EC)^2$ conference series (launched in

[7] An L1 penalty adds a term like $\lambda \sum_j |\beta_j|$ to the loss function, so large coefficients are discouraged and the model is pushed toward simpler solutions.

1990) provided a recurring forum for econometric model selection and evaluation, including dedicated themes reflecting the centrality of these issues in econometrics (EC^2, 1990).

By the early 2000s, a clear synthesis had emerged in which econometric model selection was no longer framed as a choice between theory driven selection versus data-driven selection, but as a disciplined attempt to construct models that are simultaneously useful approximations and fit for the economic purpose. In particular, Hansen (2005) argues that standard selection practices too often (i) assume a true finite-dimensional DGP, (ii) rely excessively on in-sample fit, and (iii) neglect model uncertainty, advocating instead that models be treated as approximations and evaluated relative to their intended purpose (e.g., prediction or causal/policy analysis). In the same spirit, work on the focused information criterion emphasized that the relevant bias–variance trade-off (and hence the 'right' degree of parsimony) can depend on the particular parameter or estimand of interest, not just global fit, especially under moderate misspecification (Claeskens & Hjort, 2003). In moment-based settings, Andrews and Lu (2001) propose consistent model and moment selection criteria for GMM and show how they can guide lag-length choices and specification decisions in dynamic panel models. Together, these contributions helped consolidate a pragmatic stance: combine theory-based structure with data-based diagnostics, while being explicit about the inferential target and the cost of complexity. At the same time, the early 2000s sharpened awareness of persistent difficulties that remain even with a mature toolbox (AIC, BIC, C_p, cross-validation, diagnostic/encompassing checks, Bayesian methods): overfitting through specification search and model uncertainty in inference. A central theoretical message is that ignoring the selection step can invalidate subsequent standard errors, tests, and confidence intervals. Leeb and Pötscher (2005) provides a systematic account of why post-model-selection inference is intrinsically difficult and why naive inference that conditions on the selected model can be misleading. In response, the literature increasingly promoted multimodel perspectives (either explicitly Bayesian, e.g., Bayesian model averaging in settings with many plausible regressors or frequentist model averaging) as a way to reflect model uncertainty rather than pretending a single selected model is known with certainty (Fernández, Ley & Steel, 2001; Hansen, 2007; Burnham & Anderson, 2002). This shift reframed the classic tension 'complexity vs. parsimony' as a problem of risk management under model uncertainty. More complex models may reduce approximation bias but increase variance and search-driven overfitting, so robust empirical practice requires either principled penalties or averaging methods that acknowledge uncertainty about specification (Hansen, 2005, 2007).

12.4 Cross-validation

Cross-validation (CV) is a resampling-based approach to model assessment in which the available sample is repeatedly split into an estimation (or training) part and an evaluation (or validation/test) part, so that model performance is measured on

observations not used to fit the model. This is specially useful since ”over fitting” is a common problem with simple ML methods. In its simplest, but nevertheless useful form, CV amounts to dividing the sample into two subsamples, fitting a predictor on one, and evaluating the discrepancy between predictions and realizations on the other, thereby providing an (often approximately) unbiased estimate[8] of predictive efficacy when compared with the in-sample fit. First, Stone (1974) gave an influential formalization by treating CV as a general criterion for the choice and assessment of statistical prediction and illustrating it in linear regression and ANOVA settings. Closely related ‘sample reuse’ ideas were developed around the same time. Geisser (1975) explicitly frames predictive sample reuse as a synthesis of cross-validatory assessment and function fitting, with the key emphasis placed on prediction of observables rather than estimation of parameters. Historically, CV was also tied to variable selection via prediction-error criteria. Allen (1971), an early work from engineering proposed mean squared error of prediction as a variable-selection target precisely because residual sum of squares alone always favors including more regressors, motivating criteria that approximate out-of-sample error. Modern surveys (e.g., Arlot & Celisse, 2010) emphasize that the proliferation of CV variants reflects trade-offs among bias, variance, and computational cost in estimating predictive risk , to be discussed in detail later.

As seen earlier, in econometric practice, model selection has traditionally relied heavily on penalized-likelihood criteria such as AIC and BIC (and related devices like C_p), which compare fitted models using in-sample likelihood while penalizing complexity to approximate an out-of-sample trade-off without explicitly holding out data. Cross-validation introduces a complementary, more directly *data-driven* principle: models are evaluated by their *out-of-sample* predictive accuracy under an explicit loss function, rather than by fit statistics computed on the same observations used for estimation (Konishi & Kitagawa, 2008). This perspective sits naturally alongside long-standing concerns in econometrics about specification search and overfitting. By forcing evaluation on held-out data, CV mitigates the tendency to improve in sample fit mechanically with added parameters, which encourages overly complex models. Leamer (1983b) explicitly lists cross-validation among the practical tools used in specification searches, reflecting its role as part of the model-selection toolkit rather than merely a forecasting afterthought. In applied work, however, CV typically requires *more data* than purely in-sample criteria because each split reduces the effective estimation sample, and predictive-error estimates can become noisy when the validation sets are small. This data requirement and computational burden helps explain why CV was historically less common in applied econometrics than information criteria. In modern applied econometrics, CV is used in at least three recurring ways. First, it is a workhorse for *regularization and tuning*, where the penalty parameter (e.g., λ in lasso/ridge) is chosen to minimize estimated out-of-sample loss rather than in-sample fit. This is standard in statistical learning toolkits for penalized regression and generalized linear models (see more in Hastie, Tibshirani & Friedman, 2009; Chan & Mátyás, 2022b) Second, CV ideas underpin *forecast*

[8] Conditional on the training fit, otherwise it may exhibit systematic bias, which often occurs with k-fold cross-validation with few folds.

evaluation in macroeconomic and financial applications via rolling or expanding origins, where prediction errors are averaged over multiple forecast origins to reduce sensitivity to a single split. This connects directly to the forecasting literature's emphasis on rolling-origin out-of-sample testing (Tashman, 2000). Third, CV is useful for model comparison when theoretical guidance is weak, because it evaluates candidate specifications by their predictive performance under an explicit loss function (rather than by relying on a particular likelihood-based penalty), a role emphasized in foundational and survey treatments of CV (Stone, 1974; Arlot & Celisse, 2010).

In econometrics, CV entered the general discussion of model selection as one of several practical tools (alongside stepwise procedures and goodness-of-fit testing), emphasizing that model adequacy should be judged by predictive performance on data not used for estimation. Mostly, there are three canonical forms of CV commonly discussed in econometrics:

- *k-Fold CV*: Data are divided into k subsets (folds); each fold is used once for validation while the model is trained (estimated) on the remaining $k-1$ folds.
- *Leave-One-Out CV (LOOCV)*: Each observation acts as a validation set ($k = n$).
- *Time-Series CV*: Uses rolling or expanding windows to respect temporal ordering and dependence.

In k-fold CV, the sample is partitioned into k approximately equal folds and the model is fit k times, each time leaving out one fold as a validation set and averaging the resulting loss across folds to approximate out-of-sample prediction error. Historically, the core logic of repeatedly splitting data to assess predictive performance predates modern machine learning. Stone (1974) describes cross-validation as the controlled division of the data sample into subsamples, fitting a predictor on one part and assessing it on the other, and extends the idea to regression problems including the *choice of variables*. In machine learning, k-fold CV became a standard during the 1990s as a practical accuracy-estimation and model-selection protocol. Kohavi (1995) provides a canonical large-scale empirical comparison and argues that 10-fold CV is often a strong choice for model selection in general settings. In practice, however k-fold CV comes in several design variants that differ from each other in how folds are constructed. Random (or classical) k-fold CV assigns observations to folds by random partitioning and is appropriate when observations are approximately i.i.d.. Its main advantage is simplicity and low implementation cost. Stratified k-fold CV is most common in classification problems. The fold assignment is randomized subject to preserving (approximately) the same class proportions in each fold. Kohavi (1995) shows that this procedure can reduce the variance of performance estimates under class imbalance and is often recommended in empirical comparisons. Thirdly, when observations are dependent within clusters (e.g., repeated observations, panels, or spatial clusters), clustered or grouped k-fold CV assigns whole clusters to folds (so that all observations from a cluster appear either in training or in test), because random or stratified folds can leak within-cluster information and severely understate the true out-of-cluster predictive error (see e.g., Roberts et al., 2017). This grouped-fold logic is directly analogous to using cluster-robust inference, as the goal is not to 'fix' dependence but to ensure the evaluation protocol account for it.

LOOCV is the limiting case of k-fold CV when the number of folds equals the number of observations ($k = n$). The model is trained on $n - 1$ observations and evaluated on the single omitted observation, repeated for each observation and averaged to estimate predictive loss. Its early appearances are closely tied to error-rate estimation and 'leave-one-out' (or jackknife-like) logic in classification and discriminant analysis; for example, Lachenbruch and Mickey (1968) evaluate sampling-based approaches to estimating misclassification error rates in discriminant analysis and became a widely cited early reference for leave-one-out style assessment in that literature. Stone (1974) explicitly treats single-observation omissions as a cross-validatory device and discusses variants for prediction assessment and variable choice, while Geisser (1975) reframes such sample-reuse ideas as *predictive* methodology emphasizing prediction of observables rather than inference on parameters. In modern machine learning practice, LOOCV is often viewed as nearly unbiased in some settings but potentially *high-variance* for model selection and computationally expensive. Kohavi (1995) highlights that, for selecting among models, 10-fold CV can outperform LOOCV despite LOOCV's greater computational effort. In econometrics, LOOCV's attraction historically stemmed from small-sample environments (common in macro and time-series), but dependence structures limit its validity: leaving out a single time point may still allow biases through lagged regressors or serial correlation, motivating dependence-respecting alternatives (e.g., rolling-origin evaluation).

Time-series CV modifies the basic CV principle to respect temporal ordering: training sets contain only observations prior to the forecast origin, and evaluation uses future observations, repeated over multiple origins to obtain a distribution (or average) of forecast errors. Tashman (2000) analyzes and reviews out-of-sample forecast-accuracy tests and emphasizes design choices such as *fixed-origin* versus *rolling-origin* evaluation, updating versus recalibration, and rolling windows, all of which directly correspond to time-series analogues of cross-validation. In more recent forecasting texts bridging statistics, econometrics, and machine learning, 'rolling forecasting origin' time-series CV is presented explicitly as the analogue of cross-validation for sequential data. Each test set is a future observation and the training set expands (or rolls) forward through time, avoiding the data leakage that random folds would induce (Arlot & Celisse, 2010; Tashman, 2000).

Beyond these canonical forms, several cross-validation variants are widely used in modern applied work as they address (i) the variance of the CV risk estimate, (ii) hyperparameter tuning bias, or (iii) dependence and grouping in the data. Repeated k-fold CV, repeats the k-fold split multiple times with different random partitions and averages the risk estimates, trading higher computation for lower Monte Carlo variance in the estimated test error. Nested CV uses an inner CV loop to tune hyperparameters and an outer CV loop to estimate the resulting model's generalization error, preventing the optimistic bias that arises when the same CV is used both for tuning and for performance reporting. (Nadeau & Bengio, 2003; Varma & Simon, 2006) Bootstrap-based CV (e.g., the '.632' and '.632+' rules) uses bootstrap resamples for training and out-of-bag observations for testing. (Efron & Tibshirani, 1997) Finally, blocked or grouped CV (cluster, panel, spatial blocks) assigns entire groups to folds so that

dependence within groups does not leak from training to test, which otherwise can severely underestimate predictive error. (Roberts et al., 2017)

CV is central in machine learning workflows because it provides a generic, algorithm-agnostic way to tune complexity and compare models by out-of-sample performance. In modern econometrics, this role is especially visible in *causal machine learning*, where sample splitting and cross-fitting are used to estimate high-dimensional nuisance components without overfitting, while still enabling valid inference for low-dimensional causal or structural parameters (Chernozhukov et al., 2018). CV-style evaluation is also routinely used for ensemble methods such as random forests and boosting, either through explicit resampling or closely related internal error estimation devices (e.g., out-of-bag error in random forests) (Breiman, 2001; Friedman, 2001). Finally, CV is deeply embedded in nonparametric econometrics through data-driven bandwidth (smoothing parameter) selection in kernel and local polynomial methods, where predictive criteria guide the bias-variance trade-off (Racine & Li, 2004; Li & Racine, 2004).

A key advantage of cross-validation over information criteria such as AIC and BIC is conceptual: CV evaluates models by observable out-of-sample predictive performance and therefore does not rely on the existence of a single correctly specified finite-dimensional 'true' model within the candidate set. This addresses a central critique in econometrics that many selection devices implicitly assume: a true DGP or reward in-sample fit in ways that can be misleading when all models are approximations.[9] Hansen (2005) argues explicitly that model selection should treat models as approximations and incorporate model uncertainty rather than presuming a true parametric specification. In contrast, CV directly targets the predictive risk under a user-chosen loss function, which makes it naturally aligned with the purpose of prediction and robust to moderate misspecification, as emphasized in broader surveys of CV procedures (see, e.g., Arlot & Celisse, 2010). A second advantage is practical relevance for modern, high-dimensional and machine-learning settings: when number of parameters (or variables) k is large relative to n, the main problem is controlling overfitting via regularization and tuning hyperparameters (e.g., penalty strength), for which CV provides a generic, model-agnostic selection rule. This is why CV is routinely used to tune shrinkage/regularization methods such as ridge regression (Hoerl & Kennard, 1970) and the lasso (Tibshirani, 1996), selecting the penalty parameter to optimize out-of-sample error rather than relying on likelihood penalties that may be poorly calibrated in data regimes. In the statistical learning literature, model assessment and selection is framed explicitly in terms of minimizing test error, with CV presented as a core tool for choosing complexity/regularization (Hastie et al., 2009). Likewise, practical ML implementations for penalized models (lasso/ridge/elastic net) are built around cross-validated tuning along regularization paths, underscoring CV's role as a unifying selection principle across algorithms beyond classical likelihood models (Chan & Mátyás, 2022b).

[9] That said, for completeness AIC can also be viewed as a predictive criterion, in which Stone (1977) shows asymptotic equivalence between LOOCV (under log score) and AIC. However, Shao (1997) shows, that BIC, is more naturally aligned with consistent identification when a finite-dimensional true model is among the candidates.

Although cross-validation is appealing for its direct focus on predictive risk, it has its limitation with small sample sizes as each split leaves fewer observations for estimation and only a small validation set to measure prediction error. This can inflate the variability of the estimated risk and make model ranking unstable. Arlot and Celisse (2010) explicitly surveys how CV performance depends on the problem setting and highlights that CV procedures involve bias-variance trade-off aligned with sample-sizes, so their practical behavior can deteriorate in small-n environment. In the machine-learning literature, this instability is also reflected in empirical comparisons showing that more data reuse (e.g., LOOCV) is not automatically better for model selection, since high-variance error estimates can lead to unreliable choices. In contrast, information criteria such as AIC and BIC are computed from a single fit using the full sample and thus can be more stable when data is limited. Choosing between cross-validation (CV) and information-criterion methods depends on the goal and the assumed nature of the data-generating process. If the objective is prediction and models are viewed as approximations, CV is natural because it estimates out-of-sample risk directly, whereas if the objective is identification or consistent selection of a parsimonious 'true' model, BIC-type penalties are better aligned with that aim (see more in Shao, 1997). CV can be computationally costly in large datasets, since the procedure requires repeated refitting across folds or rolling windows (and sometimes repeated tuning over hyperparameter grids). Arlot and Celisse (2010) discusses how the practical performance of CV depends on bias-variance and computational trade-offs, and why these considerations matter for choosing among the setup for CV. Another challenge is the aforementioned dependence structure of the data. With autocorrelation, clustering, or panel dependence, naive random folds can leak information and distort performance estimates, which is why time-series and grouped/blocked CV designs are typically required in econometric settings (Tashman, 2000; Bergmeir, Hyndman & Koo, 2018).

Although cross-validation (CV) is primarily used to estimate hyperparameters in economics inference has a key role. In CV setting this leads to inference for either (i) the generalization or prediction error itself or (ii) parameters reported after a CV-based model choice.

For the first goal, uncertainty quantification for generalization error is nontrivial as CV reuses observations and induces dependence across splits, so naive standard errors can be severely misleading (Nadeau & Bengio, 2003). Recent theory provides a sharper asymptotic picture. Austern and Zhou (2025) establish a central limit theorem (CLT) for the cross-validated risk over a broad model class, which yields asymptotically accurate confidence intervals for predictive risk and clarifies when and how CV achieves valid statistical inference. In particular, a key message is that parametric M-estimators can attain a 'full' variance-reduction benefit when CV evaluates the training loss, whereas with surrogate training losses or regularization the variance reduction can be smaller or larger than this benchmark depending on the model and distribution.[10] These asymptotic results complement earlier impossibility statements (Bengio & Grandvalet, 2004) in full generality there is no distribution-free

[10] Note that Austern and Zhou (2025) analysis also allows the number of folds k to grow with n at an arbitrary rate.

unbiased estimator of the variance of k-fold CV based only on observed fold errors, so finite-sample uncertainty quantification remains procedure-dependent even though asymptotic approximations can be valid under additional structure. This result is quite new, and opens up multiple possibilites in the future, that we discuss in Section 12.9.

For the second goal – inference after CV-based model selection or tuning – reporting the same cross-validated error used to tune hyperparameters (e.g., λ in LASSO) as the final test error is typically optimistically biased. This result hold for other cross-validation techniques such as grouped CV shown by Varma and Simon (2006). As we discuss in Section 12.7, there happened a great breakthrough in this aspect when the inferential target (parameter) is a low-dimensional causal or structural parameter with high-dimensional ML nuisance parameters. Econometric practice pioneered by Belloni, Chernozhukov and Hansen (2014); Chernozhukov et al. (2018); Wager and Athey (2018) started to use further sample splitting methods to control overfitting bias and recover asymptotically valid $\sqrt{n}$ inference for the parameter of interest.

12.5 Lessons from Model Selection and Why Machine Learning Seems to be Winning

A cornerstone of classical econometric modeling is the *strict exogeneity* condition, meaning that the regressors are uncorrelated with the contemporaneous error terms and are not affected by past, current, or future shocks once conditioning on unit effects. This assumption underpins the usual causal interpretation of the regression coefficients. When it holds, coefficient estimates can be interpreted as partial effects (under the maintained functional form), and model selection criteria such as AIC/BIC are applied in a setting where in-sample fit is indeed informative about the underlying conditional mean structure (Hayashi, 2000; Wooldridge, 2002). In practice, especially with observational data, strict exogeneity is frequently implausible as several regressors may in fact be endogenous. Regressors move with unobserved determinants of the outcome (omitted variables), respond to shocks (reverse causality/feedback), or are measured with error, all of which induce dependence between the observable regressors and the unobserved error terms (Angrist & Pischke, 2009). Panel settings make this tension more explicit: strict exogeneity fails mechanically when lagged dependent variables appear among regressors and can fail more generally when shocks today influence future choices of covariates (Wooldridge, 1997). This is one reason why 'best' model selection based purely on fit can be misleading for causal work. Adding covariates that are correlated with the error terms may improve in-sample fit and even out-of-sample prediction, while at the same time worsening bias in the estimated causal effect (Angrist & Pischke, 2009).

Machine learning (ML) methods are typically optimized for *prediction*. These methods are designed to minimize predictive loss and can freely exploit correlations among features, including highly collinear or proxy variables, without requiring strict exogeneity. This flexibility helps ML deliver strong predictive accuracy even when

the regressor set is 'contaminated' from a causal standpoint. However, it also means that the fitted relationships generally lack a causal interpretation unless combined with an additional identification structure; e.g., unconfoundedness, instruments, or design-based variation (see a great overview in Shmueli, 2010).

Econometrics has traditionally prioritized causal inference. The central question is not merely whether X predicts Y, but whether changing X would change Y in a counterfactual sense. Angrist and Pischke (2009) emphasize that credible causal claims require explicit identification strategies, such as randomized or quasi-experimental variation, instrumental variables, difference-in-differences, or regression discontinuity designs, and careful attention to threats like omitted variables and 'bad controls'. Athey and Imbens (2017) similarly frame modern applied econometrics as a toolkit for policy evaluation, where the validity of causal interpretation hinges on research design, robustness, and (when used) instrument validity rather than goodness-of-fit alone. In contrast, most machine learning methods are optimized for predictive accuracy. These methods are aiming to minimize out-of-sample prediction error, often using flexible function classes and regularization to capture complex associations in high-dimensional settings. Shmueli (2010) argues that explanatory (causal) and predictive modeling are distinct scientific goals, and that conflating them is common, especially when high in-sample explanatory power is mistakenly taken as evidence of strong predictive performance (or vice versa). From this viewpoint, ML methods can appear to 'win' in forecasting and classification tasks because they are designed to exploit correlations among features to reduce predictive loss, even when the underlying causal structure is unknown or ignored (Shmueli, 2010).

In the 2000s an important strand of literature in econometrics addresses the problem of 'too many moments', see Andrews and Lu (2001), and the special issue of Econometric Reviews, Carner and Medeiros (2016). Moment restrictions multiply as the number of 'features'/variables increases in the high dimensional treatments. Orthogonality conditions of this type abound. Here variable selection in ML gives way to moment selection, and criteria such as GMM or Empirical Likelihood (EL) are augmented with ML like penalization. This work is directly focused on inference on parameters and partial effects, rather than 'fit'. These techniques can also accommodate weak IV and incorrect moment restrictions, with appropriately robustified test statistics and estimation routines. Asymptotic inference is carefully developed that highlights the asymptotic rates which control degrees of misspecification and relative weights to sample and implicit priori information. An example of this latter issue is Maasoumi (1980) who proves Ridge 3SLS system estimator (which may likely have infinite moments and poor asymptotic properties) would have all its sample moments be finite! This is a demonstration of how strong the a priori information implied by penalization constraints can be relative to sample information. Shi (2016) provides another example of high dimensional treatment of 2SLS and its properties.

The same features that help ML predict well can undermine policy analysis: if the goal is to estimate the effect of an intervention, naive ML predictions may reflect spurious correlations, selection bias, or endogenous regressors, and therefore fail to support counterfactual policy conclusions without additional structure. Angrist

and Pischke (2009) stress that policy evaluation requires identification assumptions that justify causal interpretation; prediction alone does not deliver this. Athey and Imbens (2017) note that recent progress therefore comes from *integrating* ML with econometric ideas (e.g., design-based identification, orthogonality, robustness checks), rather than replacing them. A leading example is 'Debiased' double machine learning, which employs Neyman orthogonalization and Reiss representations for robust estimation of sparse effects, such as treatment effects, and is available for GMM and other settings. Many different MLearners may be used and optimized with sample splitting (cross-fitting). Valid inference for low-dimensional causal/structural parameters are obtained with a wide range of nuisance functionals, affording great flexibility in model choice in a setting that is reminiscent to partially linear models. (Chernozhukov et al., 2018).

A large part of econometric theory is built around *asymptotic* concepts: consistency, asymptotic normality, and efficiency. Typically under 'large n, fixed k' regimes in which sample size grows while the parameter dimension remains stable. These asymptotic benchmarks provide sharp guidance for estimation and model selection when their regularity conditions are plausible. For example, Shao's asymptotic framework for linear model selection makes explicit how criteria such as AIC, BIC and cross-validation can be classified by their limiting behavior depending on whether a finite-dimensional 'correct' model exists or models are inherently approximations (Shao, 1997). In this classical view, asymptotics are not merely technical, they justify using large-sample approximations to assess estimator properties, derive selection consistency, and motivate penalty forms intended to balance fit and complexity. Machine learning methods, by contrast, are usually judged by finite-sample predictive performance, rather than by asymptotic optimality within a prespecified parametric family. Accordingly, ML workflows treat model choice as a data-driven calibration problem, with cross-validation (or related resampling) used to estimate out-of-sample risk and tune complexity (e.g., depth of trees, regularization strength) directly (Hastie et al., 2009; Arlot & Celisse, 2010). This emphasis is well aligned with the predictive goal: cross-validation is explicitly designed to approximate test-set performance, while asymptotic guarantees may be weakly informative about performance at the sample sizes encountered in practice. The contrast becomes most pronounced in high-dimensional environments, where k is large relative to n (and can exceed it), so that classical large-n approximations for fixed-k models no longer describe the dominant sources of error (Bühlmann & van de Geer, 2011). In such regimes, traditional textbook inference can be ill-posed without additional structure (e.g., sparsity), and methodological focus shifts toward regularization, stability, and predictive validation, precisely the terrain where ML-style tools are operationally effective. At the same time, modern econometrics has not abandoned asymptotics. Rather, it has retooled ML methods for high-dimensional nuisance components, showing that valid inference can still be recovered when combining ML prediction with orthogonality and cross-fitting, because classical complexity assumptions (e.g., Donsker conditions) may break down in high dimensions (Athey & Imbens, 2019).

The first reason, why and how the ML approach may seem to outperform classical econometric methods is flexibility in high-dimensional data. Many ML procedures

are explicitly designed to operate when the feature space is large, often even when k is comparable to or exceeds n, by imposing regularization or using adaptive function classes. For example, the Lasso achieves scalable variable selection and stabilization via an L1 penalty (Tibshirani, 1996), while modern high-dimensional statistics emphasizes that sparsity-regularized estimators (and their variants) provide a principled way to make inference and prediction feasible in 'wide' settings (Bühlmann & van de Geer, 2011). Ensemble learners further broaden this flexibility: random forests and boosting adapt to complex nonlinearities and interactions with minimal parametric structure, and are routinely deployed as high-performing predictive baselines (Breiman, 2001; Friedman, 2001).

The second reason is robustness to functional-form misspecification. Classical econometric models often obtain interpretability and asymptotic tractability by committing to a relatively tight parametric specification. When that specification is wrong, both prediction and (especially) structural interpretation can degrade. ML methods, in contrast, typically treat the conditional mean/function as unknown and approximate it with flexible learners (trees, splines, kernels, boosting, forests) or with regularized linear expansions, which can reduce sensitivity to moderate misspecification when the goal is prediction. This 'predict-first' stance is central to statistical learning treatments of model assessment and selection, where the key target is generalization error rather than correct parametric specification (Hastie et al., 2009; Athey & Imbens, 2017). More broadly, the distinction between explanatory/causal modeling and predictive modeling clarifies why predictive robustness can dominate. Models optimized for prediction can succeed even when the assumed structural story is incomplete or inaccurate, because they are judged by out-of-sample loss rather than by structural correctness (Shmueli, 2010).

A third reason is that ML workflows adopt data-driven model selection as the default. Complexity is chosen by minimizing an estimate of out-of-sample risk, most commonly via cross-validation, rather than being determined primarily by theory or by in-sample fit statistics. Surveys of cross-validation emphasize that its many variants are motivated by practical trade-offs in estimating predictive risk (bias-variance cost), and that CV is used precisely to target generalization performance (Arlot & Celisse, 2010). Combined with regularization, this yields a coherent recipe: fit a flexible model class, tune its complexity by cross-validated risk, and report performance in terms of predictive loss, an approach that is directly aligned with the predictive objective (Hastie et al., 2009).

Finally, ML appears to win because it benefits from major computational advances, both in hardware and in algorithmic design which make complex models feasible at scale. For instance, modern penalized regression toolkits compute entire regularization paths efficiently via coordinate descent and warm starts, enabling routine cross-validated tuning of lasso/ridge/elastic-net models on large problems (Friedman, Hastie & Tibshirani, 2010). Likewise, ensemble methods such as random forests rely on computationally efficient resampling and aggregation, with internal error estimates (e.g., out-of-bag error) that further reduce the practical cost of evaluation (Breiman, 2001). Together, scalable optimization plus inexpensive model assessment makes

it practical to search over large hypothesis spaces and tuning grids, something that would be prohibitive under traditional, manually specified econometric workflows.

12.6 Causal Machine Learning and Breakthroughs in Asymptotic Behavior

Causal Machine Learning (Causal ML) refers to a set of methods that combines flexible machine learning models with the identification concept of econometrics and statistics in order to estimate causal quantities such as average and heterogeneous treatment effects, and policy impacts (Athey & Imbens, 2017; Chernozhukov et al., 2018). The motivating observation is that modern datasets often contain many potential confounders (or many transformations/interactions of a smaller set of confounders), making rigid parametric specifications risky and ad hoc, while purely predictive ML can exploit correlations that do not support counterfactual reasoning. Athey and Imbens (2019); Chernozhukov, Hansen, Kallus, Spindler and Syrgkanis (2024); Wager (2025) discusses in detail how ML retains causal parameters flexibility in high-dimensional settings, while enforcing the conditions needed for valid identification (e.g., unconfoundedness given controls, valid instruments, design-based variation; see more in).In the following, we briefly overview the main strands of this fast evolving literature.

Double/Debiased Machine Learning (DML) provides a general recipe for estimating a low-dimensional causal/structural parameter in the presence of high-dimensional *nuisance* functions, such as outcome regressions, propensity scores, or instrument regressions. There are two central failure modes of naive 'plug-in' ML methods, when estimating causal parameters. The first problem is regularization bias from using shrinkage or other estimators (ML methods will neglect variables that does not add to predictive power, but can confound the parameter of the insterest). The second is the overfitting bias, which emerges from reusing the same data for training and estimation. Chernozhukov et al. (2018) influential paper addresses both problem, by proposing a novel three step procedure: (i) estimate nuisance components using any suitable ML method, but (ii) construct an estimating equation (a 'score') that is Neyman-orthogonal (i.e., locally insensitive to small errors in nuisance estimation) and (iii) use sample splitting / cross-fitting so the data used to train nuisance models are separated from the data used to evaluate the score. As a result of their procedure, DML yields estimators that can achieve root-n rates and asymptotically valid confidence intervals under high-dimensional nuisance estimation, providing a bridge from ML flexibility to econometric inference. DML approach is used in many setting, e.g., Chang (2020) uses double/debiased difference-in-differences setup and identifies causal effects from before/after changes in treated versus control units under (conditional) parallel trends. Lieli, Hsu and Reguly (2022); Chernozhukov et al. (2024) provide an extensive overview on these and other applications of DML.

A complementary line of work focuses on heterogeneous treatment effects. Instead of estimating only an average effect, the goal is to learn how effects vary with

covariates (e.g., which subpopulations benefit from a program). First, 'causal forest' by Wager and Athey (2018) established asymptotic guarantees for treatment effect heterogeneity estimation under unconfoundedness and enable confidence intervals for individualized or local average treatment effects. Folloiwng causal forest, Athey et al. (2019) proposed Generalized Random Forests (GRF) which worked out the idea into a framework for estimating parameters defined by local moment conditions using forest-derived adaptive weights. GRF provides consistent and asymptotically normal estimates (with feasible variance estimation) for targets including heterogeneous treatment effects and other instrumental-variable versions of heterogeneous effects, thereby offering both flexibility and a formal inferential apparatus. A closely related development parallel to GRF is the Orthogonal Random Forest (ORF) by Oprescu, Syrgkanis and Wu (2019) which explicitly fuses Neyman-orthogonal scores (as in DML) with forest-based local weighting (as in GRF) to obtain heterogeneity estimators that are robust to nuisance estimation error while retaining nonparametric flexibility. Parallel to frequentist forest-based inference, Bayesian approaches have also been adapted to the causal setting, most notably the Bayesian Causal Forest (BCF, by Hahn, Murray & Carvalho, 2020), which incorporates propensity information to reduce confounding-induced bias and regularizes treatment-effect heterogeneity separately from prognostic variation. While causal forests are often presented under unconfoundedness (selection on observables), the same 'honest' forest logic can be adapted to settings where identification comes from other quasi-experimental designs by changing the within-leaf estimand and score. For example, Reguly (2025) extend causal forest in sharp and fuzzy regression discontinuity designs (RDD), with the use of local polynomial estimator, providing a specialized tool for RDD. This illustrates a broader point emphasized in modern causal learning: forest-based heterogeneity estimation is best viewed as a generic local identification tool, where the identification strategy (e.g., RD, IV, randomized assignment) determines the estimating equation, and the forest supplies adaptive localization and honest sample splitting for valid inference. Wager (2025) provides a great overview on the theoretical and application side as well.

Let us mention another approach, which estimates heterogeneous treatment effect uses 'meta-learners'. These methods differ from the GRF approach by first fitting (modified) outcome models and then combines the predictions, therefore they do not optimize splits for heterogeneity per se. Künzel, Sekhon, Bickel and Yu (2019) introduce a unifying framework, that treats different learners (e.g., S-, T-, and X-learners), which wrap generic supervised learners (forests, BART, nets, boosting, etc.) into procedures that target the conditional average treatment effect (CATE) directly. Among these learners, Künzel et al. (2019) shows that the X-learner performs well under treatment-group imbalance and can exploit structural properties of the response and CATE functions to achieve fast rates in favorable regimes, making it attractive in many applied problems where treated and control sample sizes differ markedly. A conceptually important refinement by Nie and Wager (2021) is the *R-learner* (or residualized learner), which first partials out outcome and treatment assignment via nuisance fits and then learns heterogeneity by minimizing a residualized objective that isolates the causal signal.

A different problem, where ML (can) plays a role in economics is the use of surrogates. Many policy-relevant causal effects concern outcomes observed only with long delay (e.g., lifetime earnings, long-run health), which motivates surrogate outcomes. Athey, Chetty, Imbens and Kang (2025) investigates short-run proxies that can be measured quickly and used to infer long-run treatment effects. They formulate many short-term outcomes into a surrogate index, which becomes a predictor of the long-term outcome. Under surrogacy conditions (roughly, the long-term outcome is independent of treatment conditional on the surrogates), the treatment effect on the surrogate index equals the long-run treatment effect. This approach explicitly blends ML prediction (constructing a high-dimensional index) with causal identification, and it allows faster and more precise evaluation of long-run impacts. Recent work by Kallus and Mao (2025) also clarifies robustness and bias when surrogacy assumptions fail, and develops empirical methods using additional outcomes or experimental variation.

Beyond estimation of policy changes, causal ML also changes experimental design through multi-armed bandits, which adaptively allocate units to treatment arms to balance exploration and exploitation of the treatments and their effect. In contrast to fixed-probability A/B tests, bandit algorithms update assignment probabilities as outcomes arrive, reducing exposure to clearly inferior treatments and improving cumulative welfare when the goal is to find a good arm rather than estimate all arm means precisely. Canonical approaches include Thompson sampling (Thompson, 1933) and upper confidence bound (UCB) methods (see e.g., Lai & Robbins, 1985). Athey and Imbens (2019) stress that bandits are especially important in online settings where outcomes are quickly observed and units arrive sequentially (e.g., testing templates in Facebook). A key econometric challange here is that adaptively collected bandit data are not i.i.d., so naive regression-based inference can fail. Zhang, Janson and Murphy (2020) develops inference tools tailored to batched bandits and other adaptive designs, showing when ordinary estimators break and proposing modified estimators with asymptotic normality under adaptive sampling. Probably, the most important extension of bandits for economics is the contextual bandit, where treatment assignment adapts based on observed covariates so that 'what works for whom' can be learned. Dimakopoulou, Zhou, Athey and Imbens (2019) explicitly imports balancing ideas from causal inference into contextual bandits, using inverse propensity weighting inside outcome model estimation to reduce bias under misspecification.

A final topic is policy learning. Rather than reporting (complicated or simplified) treatment effects, the goal of policy learning is to reveal an explicit treatment-assignment rule that maximizes a welfare criterion subject to operational constraints such as budget, simplicity, fairness or eligibility rules. (see, e.g., Kitagawa & Tetenov, 2018; Athey & Wager, 2021) In this literature, empirical welfare maximization provides minimax-optimal welfare regret rates in benchmark settings, while recent extensions show how to leverage observational identification strategies and doubly robust scores to obtain strong asymptotic regret guarantees for learned policies.

Reviewing these advancements, it is clear that causal ML has become influential over the last decade because it provides the kind of theoretical rigor that econometricians demand. Explicit conditions for consistency, asymptotic normality, and

valid inference even when nuisance estimation is present in high dimensions have become the norm. These results prompted the use of these methods, but first, an interpretational challenge emerged. Results of ML methods are not easy to interpret and summarize in a few sentences. The generic ML inference framework of Chernozhukov, Demirer, Duflo and Fernández-Val (2025) treats ML-based CATE proxies as inputs and then post-processes them to estimate policy-relevant features such as best linear predictors, sorted group average effects, and characteristics of the most/least impacted units. These summaries treat ML-based CATE estimates as proxies and then use sample splitting and aggregation to obtain valid inference on low-dimensional, policy-relevant features of heterogeneity, even when the first-stage proxy is imperfect. However, the interpretation of these summaries is not fully unified: results depend on the chosen estimand, the proxy learner, the grouping scheme, and the target population (e.g., overlap-weighted versus population-weighted). These challenges remain to be solved, and establishing a unified framework would be beneficial in the future, guiding us towards the limitations of this agenda.

Causal ML inherits a central limitation from causal inference: no amount of functional flexibility can fix an invalid covariate (control) set. In particular, 'more controls' can be harmful when ML pipelines inadvertently condition on bad controls, post-treatment variables, colliders, or their descendants (see more in Cinelli, Forney & Pearl, 2024). Causal ML can therefore induce selection bias and cause coefficients to diverge from the intended causal estimand when the variables used are not carefully treated. A second limitation related to the previous. Concerns with proxy controls (and related negative-control approaches. When unobserved confounding is present, identification via proxies requires additional structural assumptions (e.g., proxy relevance/independence or proximal conditions) that are typically harder to justify and diagnose than conditional ignorability. Weak identification can lead to unstable estimates and wide or misleading uncertainty statements (see e.g., Tchetgen Tchetgen, Ying, Cui, Shi & Miao, 2024). Furthermore, as we argue in Section 12.7, ML methods are inherently prone to unstable results; hence, the emphasis on sensitivity analysis and careful assessment of results is extremely important.

We summarize these advancements in an applied workflow, that separates design and identification from estimation and inference and treats ML as a flexible tool for estimation.

1. Define the causal question and estimand. Specify the target parameter (ATE/ATT/CATE, policy value, LATE/IV estimand, RD estimand, DiD effects, etc.) and the target population.
2. Justify identification. State the assumptions (randomization; conditional ignorability; IV/proxy/negative controls; RD continuity; DiD parallel trends, etc.) and use reasoning to determine admissible adjustment sets and avoid bad controls.
3. Engineer variables and choose model, which satisfies identification assumptions. Construct pre-treatment covariates, treatment timing/cohort indicators (DiD), running-variable windows (RD), and sample restrictions (overlap/common support), keeping post-treatment variables out of the adjustment set unless explicitly part of the estimand.

4. Estimate causal estimand with nuisance components or CATE with ML. Choose algorithm/learners for outcome, propensity, and (if needed) instrument/proxy models. Use cross-fitting or honest splitting to prevent overfitting-induced bias and to make subsequent inference valid.
5. Do inference and report uncertainty. Report standard errors/confidence intervals for low-dimensional targets and - when reporting heterogeneity - prefer policy-relevant summaries or validated subgroup contrasts rather than uninterpretable pointwise CATE curves.
6. Stress-test assumptions. Perform overlap diagnostics, placebo/pre-trend checks (DiD), bandwidth/window robustness (RD), and sensitivity analysis for unobserved confounding; when using proxies, discuss proxy validity and potential weak identification explicitly.
7. Translate results into decision-relevant outputs. For policy learning, convert estimated effects into explicit decision rules and evaluate welfare/regret where feasible.

This pipeline should clarify the division of labor: ML contributes flexible estimation of nuisance components or heterogeneity discovery, while econometrics contributes identification logic and inferential guarantees needed for policy-facing conclusions.

12.7 Why Machine Learning is Not (Yet) the Benchmark Currently in Econometrics?

Despite impressive (predictive) performance, machine learning is not (yet) the default benchmark. Economics and applied econometrics in contrast with theoretical econometrics places heavy weight on stability, interpretability, and policy-relevant inference. A useful way to summarize this gap is that theoretical econometrics is typically asked to deliver a defensible causal or structural statement (with theoretical proofs), whereas in applied work robustness checks, and a narrative that can be communicated to policymakers is a must. As we argued in the previous section, this stability and ease of interpretation is not (yet) present in (causal) ML. This difference does not imply that ML is 'wrong' rather, it clarifies why ML tools are not often adopted, but used as complements (exploration, nuisance estimation, forecasting) for current traditional econometric workflows.

A first obstacle is *instability*. Many ML algorithms require choices of hyperparameters (regularization strength, tree depth, learning rates, subsampling rates) and rely on data-splitting or randomized components (fold assignments, bootstrap samples, random feature subsampling), meaning that small changes in tuning or splits can lead to meaningfully different fitted models and predictions (Arlot & Celisse, 2010; Mullainathan & Spiess, 2017). This issue is not incidental, one of the foundational motivations for ensemble methods is precisely that many learning procedures are unstable (perturbing the training set can produce large changes in the fitted predictor) and aggregation can reduce variance and improve accuracy (see e.g., Breiman, 1996).

In other words, part of what makes ML powerful (high adaptivity) also makes it sensitive to seemingly minor perturbations, which can complicate reproducibility when the research goal is not only prediction but also credible policy interpretation. Cross-validation, the workhorse of ML model tuning, can also contribute to apparent instability because it introduces randomness through fold construction and because the estimated risk has nontrivial variance in finite samples, especially when the validation sets are small or the learner is unstable (Arlot & Celisse, 2010; Kohavi, 1995). From an econometric perspective, this means that the 'best' ML model under CV can change across runs unless researchers carefully fix seeds, use repeated CV, or adopt stability-enhancing procedures. Meinshausen and Bühlmann (2010) explicitly address this problem and select models based on stability across subsamples, which pairs subsampling with a base selection algorithm to control false discoveries and improve replicability. The existence and popularity of such methods underscores that instability is a real concern, and it resonates with econometric expectations of robustness and replicability for policy-facing results.

A second reason why ML is not (yet) the benchmark is that, in many econometric applications, ML is better suited as an exploratory tool than as a self-contained inferential framework. ML excels at screening variables, discovering interactions, and approximating unknown functional forms in high-dimensional settings, which is valuable when economic theory is silent about the appropriate specification or when the analyst faces many candidate controls (Mullainathan & Spiess, 2017). This concern apply directly to many ML-style workflows (feature selection, tuning, adaptive model choice), which are inherently data-dependent; hence, econometric practice often insists on either (i) reframing the exercise as prediction, or (ii) using specialized tools that deliver valid inference after selection or under high-dimensional nuisance estimation. This is precisely the motivation behind 'causal ML' approaches. The point is not that ML is unusable for causal work, but that causal analysis requires extra structure beyond predictive optimization, and that structure remains firmly rooted in econometric identification and inference principles.

A third obstacle is interpretation. Many high-performing (causal) ML models are 'black boxes' in the sense that their mapping from inputs to outputs is too complex to be transparently summarized by a small set of parameters with direct economic meaning (Rudin, 2019; Shmueli, 2010). This creates a communication gap that we previously addressed. Economists often need to explain why a result arises, quantify trade-offs, and connect estimates to theory (elasticities, marginal effects, welfare implications), whereas black-box methods alone does not yield an interpretable policy narrative. The growing literature on explainable/interpretable ML acknowledges this issue, but it also highlights the limits of post-hoc explanations for black-box models in high-stakes settings. Explainable AI developed many model agnostic tools, which help interpret (causal) ML outputs. Popular approaches include SHAP values[11] (Lundberg & Lee, 2017), which provide additive, local feature attributions grounded in Shapley-value principles and can be aggregated to yield global importance summaries. Partial dependence plots (PDPs) visualize the average change in model estimates

[11] SHapley Additive exPlanations.

as a feature varies while averaging over the empirical distribution of other features (Molnar, 2019). Individual conditional expectation (ICE) curves, which disaggregate PDPs into unit-level trajectories to reveal heterogeneity and interactions in the fitted response (Goldstein, Kapelner, Bleich & Pitkin, 2015). Permutation feature importance (PFI) complements these plots by quantifying predictive relevance via performance deterioration under feature shuffling (Molnar et al., 2021). In causal applications, these tools should be interpreted primarily as explanations of the learned effect model (not as causal mechanisms), because correlation among covariates and violations of adjustment assumptions can make PDP/SHAP-style summaries reflect model-induced associations rather than causal pathways (Zhao & Hastie, 2019). Align with these cautions, Rudin (2019) argues that for high-stakes decisions it is often preferable to use inherently interpretable models rather than attempting to explain black boxes. Rudin (2019) argues that explanations can be unreliable or misleading and accountability requires transparency. In economics, this concern is amplified by the need to communicate results to policymakers and stakeholders. When the model's concept and intuition cannot be clearly stated, it becomes harder to justify recommendations, assess external validity, or evaluate the plausibility of underlying mechanisms.

Taken together, ML is not (yet) the default benchmark in policy-facing econometric work because the core comparative advantages of ML come with frictions that matter precisely for causal communication and robustness. Many ML pipelines are intrinsically sensitive to tuning, sample splits, and algorithmic randomness, so seemingly minor perturbations can change fitted objects and downstream narratives, raising replicability concerns unless stability-enhancing practices are used. The econometric mandate typically demands identification and uncertainty quantification for well-defined causal estimands, therefore high analytical skills are mandatory for such use. Finally interpretation is improving but still fragmented. Recent work promotes mainly post-processing of complex CATE proxies into policy-relevant, low-dimensional summaries with explicit inference safeguards. Again, an approach which is not easy to follow and demands extraordinary understanding of the modeller. In parallel, model-agnostic AI tools can help communicate what drives the *estimated* effect function, but they should be treated as explanations of the fitted model rather than causal mechanisms.

12.8 Data Mining and Unsupervised Machine Learning in Econometrics?

A recurring theme in applied econometrics is the need to understand the statistical properties and patterns of the observed data. Ideally, this should complement the principal activities of *causal interpretation* under explicit identifying assumptions. The parallel is the neglect of properties of economic time series properties which was remedied by extensive attention to stationarity and co-integration characteristics of economic observations.

In the modern 'design-based' paradigm (Angrist & Pischke, 2009; Athey & Imbens, 2017), economists want to learn what would happen under counterfactual interventions (policies, treatments, shocks), and they therefore emphasize research design, identification strategies, and falsifiable robustness checks. Unsupervised machine learning and generic data mining, by contrast, focus on discovering regularities in the joint distribution of observables (Hastie et al., 2009; Shmueli, 2010).

Following the taxonomy in Hastie et al. (2009), unsupervised machine learning covers a family of tools that aim to summarize or organize the joint distribution of covariates without reference to an outcome variable. A first class is *cluster analysis*, which partitions observations into groups that are internally similar and externally dissimilar; canonical examples include K-means (see e.g., Hartigan & Wong, 1979) and hierarchical clustering that depend on a user-chosen dissimilarity metric and linkage rule (see e.g., Murtagh, 1983). A closely related, more explicitly statistical approach is *model-based clustering*, where the data are assumed to arise from a mixture distribution (often Gaussian mixtures) and clusters are latent components estimated via likelihood methods such as expectation-maximization algorithm (EM). A second broad class is *dimension reduction and representation learning*, which seeks low-dimensional structure (factors or embeddings) that preserves salient features of the high-dimensional covariate space. Linear methods such as principal components analysis (PCA) provide variance-maximizing directions and are often used for visualization, compression, and pre-processing (see e.g., Jolliffe, 2002), while distance- or kernel-based embeddings (e.g., multidimensional scaling and spectral methods) aim to preserve pairwise geometry or graph structure when linear variance summaries are inadequate (Ng, Jordan & Weiss, 2002). A third class emphasizes *matrix factorization* and parts-based representations, such as nonnegative matrix factorization (NMF), which can yield more interpretable latent factors in applications where additivity and nonnegativity are meaningful (Lee & Seung, 1999). Departing from the classical unsupervised toolkit, recent advances have shifted unsupervised learning toward *representation learning* with deep, self-supervised objectives and modern generative modeling. On the representation side, contrastive and bootstrap-style methods learn embeddings by making different views or votes of the same observation predict each other,[12] producing features that transfer well to downstream tasks even without labels (see more in T. Chen, SKornblith, Norouzi & Hinton, 2020; He, Fan, Wu, Xie & Girshick, 2020; Grill et al., 2020). A complementary line uses *masked modeling* (predicting withheld parts of the input) popularized in Natural Language Processing (NLP) by Bidirectional Encoder Representations from Transformers (BERT, see Devlin, Chang, Lee & Toutanova, 2019) and extended to vision with masked autoencoders (MAE, see He et al., 2021), which scale efficiently and often rival supervised pretraining methods. In generative modeling, variational autoencoders (VAEs, Kingma & Welling, 2013), Generative Adversarial Networks (GANs, Goodfellow et al., 2014), and diffusion models (Ho, Jain & Abbeel, 2020) provide flexible density estimators and used for compression, anomaly detection, and synthetic data generation. Finally, nonlinear manifold learning for visualization and

[12] E.g., SimCLR, MoCo, BYOL.

geometry-preserving embeddings has matured with widely used methods such as t-SNE (van der Maaten & Hinton, 2008) and UMAP (McInnes, Healy & Melville, 2018). These developments substantially broaden what 'unsupervised' ML can accomplish.

Although the applicability of these methods is large, they also amplify econometric concerns. Unsupervised ML objectives are typically chosen for predictive transfer or perceptual fidelity rather than parameter interpretation with careful identification. Again this is not surprising as many unsupervised/data-mining techniques are atheoretical. They optimize purely statistical criteria (e.g., within-cluster sum of squares for k-means; explained variance for PCA, etc.), without enforcing constraints implied by economic theory or causal identification. Conceptually, this issue is more severe, as unsupervised ML is intrinsically economically ill-posed. There is no universally accepted definition of what constitutes a 'true' cluster, factor or embedding, and therefore no ground truth to specify the loss function.[13] This ill-posedness creates substantial challenge: different reasonable model setup choices (e.g., distance metrics, scaling, linkage rules, etc) can lead to different partitions, echoing formal results. For example Kleinberg (2002) formalizes this problem by showing that no single clustering rule can satisfy a set of seemingly natural axioms simultaneously. Finally, the resulting 'discovered structure' can be fragile. Cluster assignments and nonlinear embeddings are often sensitive to initialization, perturbations, and preprocessing, motivating an extensive literature on resampling-based stability diagnostics and uncertainty quantification for clustering outputs (see Hennig, 2007; Yu et al., 2019), and practical warnings that popular visualization embeddings such as t-SNE can be misleading if interpreted as preserving global geometry or as uniquely identifying clusters (see Wattenberg, Viégas & Johnson, 2016; van der Maaten & Hinton, 2008). These issues matter for econometrics because policy narratives typically require interpretable, robust, and externally defensible statements. Purely descriptive partitions of observables provide informative exploratory summaries that may guide further econometric modeling.

Such data analysis and factor identification can become influential when embedded in an economic/statistical model that is coherent with high dimensional settings. Unsupervised learning typically returns objects that are not naturally labeled, hence additional work to provide suitable interpretation is needed. The canonical example for the use of unsupervised technique in economics is factor models. Principal components or unobserved factors are used to estimate latent factors that have an economic meaning (e.g., 'common business cycle' components), and the resulting factors are then plugged into forecasting or structural analysis under an approximate factor model (see e.g., Stock & Watson, 2002; Bai & Ng, 2002). Behavioral economics is another example, where latent heterogeneity in preferences and decision rules are investigated with unsupervised models. In social-dilemma experiments, researchers often recover distinct behavioral profiles (e.g., conditional cooperators vs. free riders) from strategy-method response patterns, and then study how these types interact with institutions such as peer punishment (Fischbacher, Gächter & Fehr, 2001; Fehr &

[13] For example von Luxburg and Ben-David (2005); von Luxburg (2010) emphasize this statistical perspective in the case of clustering.

Gächter, 2000). In risky choice, model-based clustering via finite mixtures is used to estimate qualitatively different preference types and their posterior class probabilities. For example, Bruhin, Fehr-Duda and Epper (2010) use a finite mixture model to show that most subjects exhibit prospect-theory-style probability distortion, while a minority behave close to expected-value maximizers. In learning and belief-updating settings, mixtures identify heterogeneity within and across individuals in the reliance on Bayesian updating versus reinforcement-type heuristics (Alós-Ferrer & Garagnani, 2023). Finally, in strategic interaction, 'level-of-reasoning' heterogeneity is modeled as a latent discrete distribution of thinking steps. The cognitive hierarchy model treats depth of reasoning as an unobserved type and fits its population distribution to experimental play in one-shot games (Camerer, Ho & Chong, 2004). All these research shows that with careful assessment unsupervised ML can be used in economics. At last, let us note that a prominent class of unsupervised tools in economics comes from text-as-data, which we discuss more in detail in the next section.

Based on the aforementioned, a central practical limitation of unsupervised ML is that many of its outputs are not designed to be stable objects. Small, economically innocuous changes in preprocessing, tuning choices, or sampling can produce materially different results. This instability is not an implementation detail but a structural feature of the problem. In the absence of ground-truth labels, there is no unique notion of the 'correct' representation, and reasonable algorithms can disagree even on the same dataset (see e.g., von Luxburg & Ben-David, 2005; von Luxburg, 2010; Kleinberg, 2002). In applied economics this matters because it becomes easy to over-interpret patterns that are, in fact, artifacts of a particular setup.[14] These generic stability issues become more severe in economics as economic datasets often feature strong dependences (serial correlation, persistent unit or hierarchical clustering). Many unsupervised routines are developed under i.i.d. intuition,[15] so naive use can turn dependence into spurious 'structure'. The econometrics literature has long highlighted how autocorrelation and common trends can create compelling-looking patterns that are unrelated to any substantive relationship. Classic 'nonsense correlations' show that two unrelated but persistent time series can exhibit high correlation simply because each is internally autocorrelated (Yule, 1926; Granger & Newbold, 1974; Phillips, 1986). The same logic carries over to unsupervised learning: clustering raw data can mechanically group units by shared trends or common confounders rather than by economically meaningful heterogeneity.

For econometric applications, this means that unsupervised outputs should typically be treated as exploratory summaries unless they pass stability checks and are disciplined by domain structure. Ultimately embedding the unsupervised step inside a model with clear estimands and inferential targets helps to assess the uncertainty. Although synthetic control, matching estimators and recent matrix completion methods in panel settings are fundamentally causal estimators rather than generic unsupervised learning tools, they rely on unsupervised-like primitives (distance-based similarity for matching, and constrained reweighting/balancing for synthetic

[14] Stability diagnostic in clustering aims to assess such sensitivity, see e.g., Hennig (2007); Yu et al. (2019).

[15] Or rely on distance notions that behave poorly when observations are correlated.

control, and low-rank imputation for matrix completion) to construct counterfactuals. Importantly, unlike many purely descriptive unsupervised methods, these approaches come with comparatively disciplined uncertainty assessments: synthetic control relies on design-based placebo and sensitivity-style inference (Abadie, Diamond & Hainmueller, 2010; Abadie, 2021), matching admits formal large-sample theory and consistent variance estimation (Abadie & Imbens, 2006), and matrix completion frames counterfactual prediction as a regularized low-rank estimation problem with explicit links to synthetic control and related identification frameworks (Athey, Bayati, Doudchenko, Imbens & Khosravi, 2021). These solutions may provide future fusions of more elaborate unsupervised ML and economics. Otherwise if these safeguards are ignored, the combination of instability and dependence can turn unsupervised pattern discovery into a generator of spurious narratives rather than reliable empirical evidence.

12.9 Open Questions about the Use of Large Language Models (LLMs) in Econometrics

Large Language Models (LLMs) are a prominent subclass of 'foundation models' trained on broad, web-scale corpora and then adapted to downstream tasks. LLMs create new opportunities for econometrics mainly by turning unstructured inputs (text, voice, video) into structured measures that can enter standard economic analysis (see Chapter 4, and Bommasani et al., 2021; Gentzkow, Kelly & Taddy, 2019). Such practices raises unresolved questions that sit squarely at the intersection of econometric identification, measurement theory, privacy law, and algorithmic accountability. The open issues below determine whether LLM-based variables can be treated as reliable measurements, whether empirical results are reproducible, and whether policy-facing outputs can be responsibly communicated and audited.

In the last decade, we can see a fast growing number of empirical papers, which use text-as-data. Such papers from leading journals construct economically interpretable measures from large corpora and then discipline them with validation and econometric analysis, which serves insights to more developed text processing methods such as LLMs. Baker, Bloom and Davis (2016) use newspaper language and coverage frequencies and transform them into indices of policy uncertainty to show that higher uncertainty is associated with greater stock-price volatility. Caldara and Iacoviello (2022) investigate geopolitical risk utilizing newspaper articles. They show that higher geopolitical risk is associated with lower investment and employment. Chahrour, Nimark and Pitschner (2021) analyse sectoral media attention and show that news coverage is disproportionately highlights unrepresentative sectoral developments. They highlight that media coverage itself can be a driver of business-cycle-like dynamics. Gentzkow and Shapiro (2010) uses political language in newspapers and then convert into a quantitative measure of ideological slant. A last example now is Hassan, Hollander, van Lent and Tahoun (2019) who use corporate disclosures to obtain firm-level exposure to political risk. Across applications, the text step serves

as a high-frequency measurement which is compressed into scalar or topic-specific signals—after which standard empirical designs relate these signals to behaviour (investment, hiring, volatility, participation) and assess credibility via internal and external validation, making textual measurement an input to causal or structural inference rather than a stand-alone descriptive exercise.

As shown in the previous examples, economics typically demands that variables and parameters correspond to interpretable economic objects (preferences, beliefs, uncertainty, sentiment, expectations, constraints) and that measurement error can be reasoned about. 'Text as data' work shows that converting text into variables is feasible but stresses the need for careful design choices (representation, dictionaries vs. supervised learning vs. topic models) and problem-specific validation (Gentzkow et al., 2019).

LLMs expand the menu of representations dramatically (e.g., embeddings, summaries, extracted themes, inferred intent), but they also make the mapping from raw text to a numeric measure less transparent. Although this is present, there are more and more (working) papers that use LLMs to create new measures. For example, Jha, Qian, Weber and Yang (2024) uses conference-call transcripts to create a firm-level ChatGPT investment score that measures managers' anticipated changes in capital expenditures. In labor markets, Y. Chen, Fang, Zhao and Zhao (2024) use LLM to simulate an HR specialist and generate a match-quality score for applicant–job pairs, by extracting information from job/worker text labels to address labor market mismatch. In international political economy, Clayton, Coppola, Maggiori and Schreger (2025) tackle the difficulty of systematically identifying coercive policy episodes in large corpora by having LLMs generate structured classifications. They demonstrate that firms affected by tariffs respond primarily with price changes, whereas firms affected by export controls respond disproportionately by investing in research and development. Asirvatham, Mokski and Shleifer (2026) provides a general-purpose pipeline that uses GPT to generate attribute ratings/labels (e.g., 'pro-innovation' stance) from qualitative text, enabling the creation of new datasets across speeches, curricula, and other corpora at scale. Central-bank communication is also investigated by LLMs and transformed into quantitative inputs (topic, stance, audience, sentiment). These methods enable high-frequency monitoring of how messaging evolves and how it aligns with expectations, markets, and the real economy.

Although these methods getting more popular, LLM model's internal features are not inherently aligned with economic theory, and the same prompt can yield different outputs depending on context, sampling settings, or model version. A core open question is whether LLM-generated measures are reproducible and *stable and comparable* across time, domains, and populations. Econometric credibility often relies on some form of measurement invariance (e.g., the same construct is measured similarly across groups) and on the ability to assess noise and bias in the measurement process. The automated text analysis literature (see e.g., Grimmer & Stewart, 2013; Gentzkow et al., 2019) emphasizes that validity and reliability cannot be assumed and require explicit validation strategies (hand labeling, robustness to alternative preprocessing, sensitivity checks, and external benchmarks). With LLMs, this validation problem becomes sharper because models may hallucinate plausible

but incorrect interpretations, may be sensitive to prompts, and may change behavior across versions or fine-tunes, threatening reproducibility of derived variables. The foundation model literature explicitly highlights such reliability and evaluation risks, including distribution shift and undocumented model changes (Bommasani et al., 2021). Even if an LLM-derived score is strongly predictive, it may behave like a noisy proxy for an economic construct, creating attenuation bias, misclassification, or endogeneity when used in causal models. Classical 'text as data' guidance already warns that automated representations are imperfect models of language and must be evaluated for construct validity. LLMs intensify the concern because their representations are learned implicitly from broad data rather than from a task-specific measurement model (Bender, Gebru, McMillan-Major & Shmitchell, 2021). An open econometric agenda is therefore to develop principled ways to (i) quantify uncertainty in LLM-generated measures, (ii) propagate that uncertainty into standard errors and inference, and (iii) design robustness checks that treat the LLM step as a potentially endogenous 'measurement technology' not a black-box oracle.

Although these are the most prominent issues of the use of LLMs in economics, let us mention two further points, which should be considered: i) data privacy and confidentiality, and ii) bias and fairness.

Let us start with data privacy and confidentiality considerations. Many economic datasets are confidential by design (administrative tax records, employer-employee matched data, banking and transaction data, health-insurance claims, proprietary firm data, see Chapter 15). Privacy regulation frameworks such as the EU's GDPR impose strict requirements on lawfulness, purpose limitation, data minimization, transparency, and security when processing personal data. These principles raise a direct question for LLM use: how can economists fine-tune or adapt LLMs on sensitive microdata while ensuring that neither the training process nor the released model leaks personal information? A well-studied technical response is *differential privacy* (DP), which bounds how much any single individual's data can influence the trained model. DP training for deep networks is a prominent approach to mitigating membership inference and memorization risks (Abadi et al., 2016). However, applying DP to LLM adaptation in econometrics raises open questions about the privacy-utility trade-off: how much predictive/semantic performance is lost under a privacy budget appropriate for sensitive administrative data, and how should researchers report privacy guarantees in ways that are meaningful for economic replication standards and legal compliance? Chapter 15 addresses such questions in more details.

The last point is bias and fairness with LLMs. LLMs inherit statistical patterns and social biases present in their training corpora and may reproduce or amplify them in generated text, classifications, or extracted measures. Critical analyses of large language models highlight risks including skewed representation, harmful stereotypes, and the downstream harms of deploying models trained on broad, poorly documented data (Bender et al., 2021; Bommasani et al., 2021). In econometrics, the concern is not only ethical but also methodological. Biased LLM outputs can distort economic measurement, affect forecasts, and contaminate policy analysis if the bias correlates with protected characteristics or with economic outcomes in systematic ways.

Overall, the open questions around interpretability, privacy, and fairness point to a unifying perspective: in econometrics LLMs should be treated as measurement technologies whose outputs must be validated, whose failure modes must be documented, and whose use must be governed. These points are similar to unsupervised ML, discussed in the previous section, however the employed models' complexity is much larger. The foundation model literature explicitly frames broad-purpose models as powerful but risky general infrastructure that can propagate defects downstream, while the text-as-data literature emphasizes the necessity of domain-specific validation for language-based measurements. A major research frontier therefore must be methodological: building econometric workflows that (i) formalize and test the stability and construct validity of LLM-derived variables, (ii) maintain legal and ethical guarantees for sensitive data, and (iii) provide transparency and auditability comparable to the standards expected for policy evaluation.

12.10 Conclusion

This chapter has provided an extensive discussion of the methods that have developed in ML, with special emphasis on their relation to econometric practice and model discovery with observational data. The selective history that is given is not meant to be exhaustive, but rather serve to illuminate how historical ideas in econometrics and other statistical fields are finding their way to provide a better understanding and appreciation of modern ML in a context that emphasize policy evaluation and decision making with uncertain models. Our view is that that twin challenges of finding and controlling covariates and functional forms, both rather absent from underlying economic theories, can be better met with tools and algorithms of modern ML and the vast new computational reach. We point out that this opportunity comes with some potential costs. The vastly greater set of possible models in high dimensional settings, can provide many equally good prediction models (universal approximants such as NN and basis polynomials and seave methods...), but model choice becomes more challenging, and inference on traditional objects of interest, such as partial effects, elasticities, and averaged nonlinear effects, require far more careful treatment for robustness, interpretation, and statistical properties. We also made a number of references to a subtle issue of relative value of sample and non sample information. There is fundamentally no mystery why ML methods work so well, at least for fitting and prediction, through greater flexibility in model specification. There is also no mystery that a cost has to be paid in terms of statistical reliability of inferences on parameters and partial effects. The main reason this may all work well for both fit and inference is the value of a priori information that is being used. This is very clear in the case of larger dimensionality than the sample size. There is a fundamentally 'empirical Bayes' interpretation of balancing the relative order of magnitude of a priori information and the sample information/size. We have provided extensive references and citations, but of course, by no means an exhaustive set.

Appendix

Brief overview of backwardpropagation

To briefly outline the method, let the loss function given by $L(w)$ and the gradient be $\nabla_w L(w)$. The objective with multilayer network is a composition of $L(w) = \ell\big(y,\ f_K(f_{K-1}(\cdots f_1(x;w_1)\cdots;w_{K-1});w_K)\big)$. The method has two stages, the forward pass and a backward pass (the backpropagation). Let us consider a multi-layer predictor defined by repeated composition. The first layer is the input sequence $h_0 = x$, followed by $k = 1,\ldots,K$ layers, with $h_k = f_k(h_{k-1};w_k)$. The resulting predictor is $\hat{y} = h_K$, which relates to the loss function $L(w) = \ell(y,\hat{y})$. First, one takes the forward pass, by computing $h_1,\ldots,h_K$ (and thus $L(w)$), while caching intermediate values needed for differentiation. The second stage is the 'backward pass', which propagate derivatives from the output back through the layers using the chain rule, producing the gradients $\nabla_{w_k} L(w)$ for all k in one sweep. Finally, one can use any gradient-based optimizer, e.g., $w^{(t+1)} = w^{(t)} - \eta\,\nabla_w L(w^{(t)})$ and repeat until convergence. Werbos (1974) contains the core computational insight – efficient gradient evaluation for nested/dynamic models via an 'ordered derivative/dynamic feedback' procedure used within gradient optimization – while the standard layered backpropagation exposition and widespread neural-network applications were popularized later (see Griewank, 2012).

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K. & Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 acm sigsac conference on computer and communications security (ccs '16)* (pp. 308–318). doi: 10.1145/2976749.2978318

Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, *59*(2), 391–425. doi: 10.1257/jel.20191450

Abadie, A., Diamond, A. & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association*, *105*(490), 493–505. doi: 10.1198/jasa.2009.ap08746

Abadie, A. & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, *74*(1), 235–267. doi: 10.1111/j.1468-0262.2006.00655.x

Agar, J. (2020). What is science for? the Lighthill report on artificial intelligence reinterpreted. *The British Journal for the History of Science*, *53*(3), 289–310. doi: 10.1017/S0007087420000230

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), *Proceedings of the second*

international symposium on information theory (pp. 267–281). Budapest: Akadémiai Kiadó.

Allen, D. M. (1971). Mean square error of prediction as a criterion for selecting variables. *Technometrics*, *13*(3), 469–475. doi: 10.1080/00401706.1971.10488811

Alós-Ferrer, C. & Garagnani, M. (2023). Part-time bayesians: Incentives and behavioral heterogeneity in belief updating. *Management Science*, *69*(9), 5523–5542. doi: 10.1287/mnsc.2022.4584

Andrews, D. W. K. & Lu, B. (2001). Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models. *Journal of Econometrics*, *101*(1), 123–164. doi: 10.1016/S0304-4076(00)00077-4

Angrist, J. D. & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Arlot, S. & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79. doi: 10.1214/09-SS054

Asirvatham, H., Mokski, E. & Shleifer, A. (2026, February). *Gpt as a measurement tool* (Working Paper No. 34834). National Bureau of Economic Research. Retrieved from https://www.nber.org/papers/w34834 doi: 10.3386/w34834

Athey, S., Bayati, M., Doudchenko, N., Imbens, G. & Khosravi, K. (2021). Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, *116*(536), 1716–1730. doi: 10.1080/01621459.2021.1891924

Athey, S., Chetty, R., Imbens, G. W. & Kang, H. (2025, 09). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. *The Review of Economic Studies*, rdaf087. doi: 10.1093/restud/rdaf087

Athey, S. & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, *31*(2), 3–32. doi: 10.1257/jep.31.2.3

Athey, S. & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*(Volume 11, 2019), 685-725. doi: https://doi.org/10.1146/annurev-economics-080217-053433

Athey, S., Tibshirani, J. & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148–1178. doi: 10.1214/18-AOS1709

Athey, S. & Wager, S. (2021). Policy learning with observational data. *Econometrica*, *89*(1), 133-161. doi: https://doi.org/10.3982/ECTA15732

Austern, M. & Zhou, W. (2025, November). Asymptotics of cross-validation. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, *61*(4), 2804–2865. doi: 10.1214/24-AIHP1488

Bai, J. & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, *70*(1), 191–221. doi: 10.1111/1468-0262.00273

Bai, J. & Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, *66*(1), 47–78.

Baker, S. R., Bloom, N. & Davis, S. J. (2016). Measuring economic policy

uncertainty. *The Quarterly Journal of Economics*, *131*(4), 1593–1636. doi: 10.1093/qje/qjw024

Belloni, A., Chernozhukov, V. & Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, *28*(2), 29–50. doi: 10.1257/jep.28.2.29

Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 acm conference on fairness, accountability, and transparency (facct '21)* (pp. 610–623). doi: 10.1145/3442188.3445922

Bengio, Y. & Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation. *Journal of Machine Learning Research*, *5*, 1089–1105.

Bergmeir, C., Hyndman, R. J. & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, *120*, 70–83. doi: 10.1016/j.csda.2017.11.003

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... Liang, P. (2021). *On the opportunities and risks of foundation models* (Tech. Rep.). Stanford Center for Research on Foundation Models (CRFM). Retrieved from https://crfm.stanford.edu/assets/report.pdf

Boser, B. E., Guyon, I. M. & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on computational learning theory (colt '92)* (pp. 144–152). doi: 10.1145/130385.130401

Bottou, L. & Bousquet, O. (2008). The tradeoffs of large scale learning. In *Advances in neural information processing systems 20 (nips 2007)* (pp. 161–168).

Box, G. E. P. & Jenkins, G. M. (1970). *Time series analysis: Forecasting and control*. San Francisco: Holden-Day.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. doi: 10.1007/BF00058655

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. doi: 10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth International Group.

Bruhin, A., Fehr-Duda, H. & Epper, T. (2010). Risk and rationality: Uncovering heterogeneity in probability distortion. *Econometrica*, *78*(4), 1375–1412. doi: 10.3982/ECTA7139

Bühlmann, P. & van de Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Heidelberg: Springer. doi: 10.1007/978-3-642-20192-9

Burnham, K. P. & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York: Springer.

Caldara, D. & Iacoviello, M. (2022). Measuring geopolitical risk. *American Economic Review*, *112*(4), 1194–1225. doi: 10.1257/aer.20191823

Camerer, C. F., Ho, T.-H. & Chong, J.-K. (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, *119*(3), 861–898. doi:

10.1162/0033553041502225

Caner, M. (2009). LASSO-Type GMM estimator. *Econometric Theory*, *25*(1), 270-290. Retrieved from https://www.jstor.org/stable/20532439

Caner, M. & Fan, Q. (2015). Hybrid Generalized Likelihood estiamtors: Instrument selection with adaptive LASSO. *Journal of Econometrics*, *187*(1), 256-274. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S030440761500069X

Caner, M., Maasoumi, E. & Riquelme, J. (2016). Moment and IV selection approaches: A comparative simulation study. *Econometric Reviews*, *35*(8-10), 1562-1581. doi: 10.1080/07474938.2015.1092804

Carner, M. & Medeiros, M. (2016). Model selection and shrinkage: An overview. *Econometric Reviews – Special Issue*, *35*(8-10), 1343-1346. Retrieved from https://www.tandfonline.com/doi/full/10.1080/07474938.2015.1071157

Chahrour, R., Nimark, K. & Pitschner, S. (2021). Sectoral media focus and aggregate fluctuations. *American Economic Review*, *111*(12), 3872–3922. doi: 10.1257/aer.20191895

Chan, F., Harris, M. N., Ranjodh, B. S. & Wei, E. Y. (2022). Nonlinear econometric models with machine learning. In F. Chan & L. Mátyás (Eds.), *Econometrics with machine learning* (p. 41-78). Springer Nature. https://link.springer.com/chapter/10.1007/978-3-031-15149-1_2.

Chan, F. & Mátyás, L. (Eds.). (2022a). *Econometrics with machine learning*. Springer Nature. https://link.springer.com/book/10.1007/978-3-031-15149-1.

Chan, F. & Mátyás, L. (2022b). Linear econometric models with machine learning. In F. Chan & L. Mátyás (Eds.), *Econometrics with machine learning* (p. 1-39). Springer Nature. https://doi.org/10.1007/978-3-031-15149-1_1.

Chang, N.-C. (2020). Double/debiased machine learning for difference-in-differences models. *The Econometrics Journal*, *23*(2), 177–191. doi: 10.1093/ectj/utaa001

Chen, T., SKornblith, S., Norouzi, M. & Hinton, E., Geoffrey. (2020). A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th international conference on machine learning, ICML 2020, 13-18 july 2020, virtual event* (pp. 1597–1607). PMLR.

Chen, Y., Fang, H., Zhao, Y. & Zhao, Z. (2024, April). *Recovering overlooked information in categorical variables with llms: An application to labor market mismatch* (Working Paper No. 32327). National Bureau of Economic Research. Retrieved from https://www.nber.org/papers/w32327 doi: 10.3386/w32327

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68. doi: 10.1111/ectj.12097

Chernozhukov, V., Demirer, M., Duflo, E. & Fernández-Val, I. (2025). Fisher–schultz lecture: Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. *Econometrica*, *93*(4), 1121-1164. doi: https://doi.org/10.3982/ECTA19303

Chernozhukov, V., Hansen, C., Kallus, N., Spindler, M. & Syrgkanis, V. (2024). Applied causal inference powered by ml and ai. *arXiv preprint arXiv:2403.02467*.

Retrieved from https://causalml-book.org/ doi: 10.48550/arXiv.2403.02467

Chudik, A., Kapetanios, G. & Pesaran, H. M. (2018). A one covariate at a time, multiple testing approach to variable selection in high-dimensional linear regression models. *Econometrica*, *86*(4), 1479-1512.

Cinelli, C., Forney, A. & Pearl, J. (2024). A crash course in good and bad controls. *Sociological Methods & Research*, *53*(3), 1071–1104.

Claeskens, G. & Hjort, N. L. (2003). The focused information criterion. *Journal of the American Statistical Association*, *98*(464), 900–916.

Clayton, C., Coppola, A., Maggiori, M. & Schreger, J. (2025, July). *Geoeconomic pressure* (Working Paper No. 34020). National Bureau of Economic Research. Retrieved from https://www.nber.org/papers/w34020 doi: 10.3386/w34020

Coulom, R. (2007). Efficient selectivity and backup operators in monte-carlo tree search. In *Computers and games* (Vol. 4630, pp. 72–83). Springer. (Implemented in the 9x9 Go program Crazy Stone; reports tournament success.)

Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, *EC-14*(3), 326–334. doi: 10.1109/PGEC.1965.264137

Cover, T. M. & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27. doi: 10.1109/TIT.1967.1053964

Dean, J. & Ghemawat, S. (2004). Mapreduce: Simplified data processing on large clusters. In *Proceedings of the 6th symposium on operating systems design and implementation (osdi '04)* (pp. 137–150). San Francisco, CA, USA: USENIX Association.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–38. doi: 10.2307/2984875

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics. doi: 10.18653/v1/N19-1423

Diebold, F. X. & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263. doi: 10.1080/07350015.1995.10524599

Dimakopoulou, M., Zhou, Z., Athey, S. & Imbens, G. (2019). Balanced linear contextual bandits. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 3445–3453). doi: 10.1609/aaai.v33i01.33013445

Doornik, J. A. & Hendry, D. F. (1998). Modelling dynamic systems: Pcgive, volume ii [Computer software manual]. Retrieved from https://www.oxrun.org/doc/PcGive/PcGive_vol2.pdf

Dovonon, P. & Gospodinov, N. (2024). Specification testing for conditional moment restrictions under local identification failure. *Quantitative Economics*, *15*(3), 849-891. doi: 10.3982/QE2242

Drukker, D. M. & Liu, D. (2022). Finite-sample results for LASSO and stepwise Neyman-orthogonal Poisson estimators. *Econometrics Reviews*, *41*(9), 1047-1076. Retrieved from https://www.researchgate.net/publication/356136180_Finite-sample_results_for_lasso_and_stepwise_Neyman-orthogonal_Poisson_estimators

Duda, R. O. & Hart, P. E. (1973). *Pattern classification and scene analysis: Theory and practice*. New York: Wiley.

EC^2. (1990). *(ec)^2 conference series: European conferences of the econom[etr]ics community.* https://sites.google.com/site/ecpower2/home-1.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, *32*(2), 407–451. doi: 10.1214/009053604000000067

Efron, B. & Tibshirani, R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, *92*(438), 548–560. doi: 10.1080/01621459.1997.10474007

Fehr, E. & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, *90*(4), 980–994. doi: 10.1257/aer.90.4.980

Fernández, C., Ley, E. & Steel, M. F. J. (2001). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics*, *16*(5), 563–576. doi: 10.1002/jae.623

Fischbacher, U., Gächter, S. & Fehr, E. (2001). Are people conditionally cooperative? evidence from a public goods experiment. *Economics Letters*, *71*(3), 397–404. doi: 10.1016/S0165-1765(01)00394-9

Flach, P. A. (2011). The machine learning journal: 25 years young. *Machine Learning*, *82*(3), 273–274. doi: 10.1007/s10994-011-5241-z

Freund, Y. & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *Proceedings of the thirteenth international conference on machine learning (icml 1996)* (pp. 148–156). Morgan Kaufmann.

Freund, Y. & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139. doi: 10.1006/jcss.1997.1504

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, *19*(1), 1–67. doi: 10.1214/aos/1176347963

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. doi: 10.1214/aos/1013203451

Friedman, J. H., Hastie, T. & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. doi: 10.18637/jss.v033.i01

Friedman, J. H. & Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, *76*(376), 817–823. doi: 10.1080/01621459.1981.10477729

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, *70*(350), 320–328. doi: 10.1080/01621459.1975.10479865

Gentzkow, M., Kelly, B. & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535–574. doi: 10.1257/jel.20181020

Gentzkow, M. & Shapiro, J. M. (2010). What drives media slant? evidence from U.S. daily newspapers. *Econometrica*, *78*(1), 35–71. doi: 10.3982/ECTA7195

Ghemawat, S., Gobioff, H. & Leung, S.-T. A. (2003). The google file system. In *Proceedings of the nineteenth acm symposium on operating systems principles (sosp '03)* (pp. 29–43). doi: 10.1145/945445.945450

Goldberger, A. S. (1964). *Econometric theory*. New York: John Wiley & Sons.

Goldstein, A., Kapelner, A., Bleich, J. & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *arXiv preprint*. Retrieved from https://arxiv.org/abs/1309.6392 (arXiv:1309.6392)

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems (neurips)*.

Gospodinov, N., Kan, R. & Robotti, C. (2014). Misspecification-robust inference in linear asset-pricing models with irrelevant risk factors. *The Review of Financial Studies*, *27*(7), 2139-2170.

Gospodinov, N. & Maasoumi, E. (2021). Generalized aggregation of misspecified models: With an application to asset pricing. *Journal of Econometrics*, *222*(1), 451-467. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0304407620302001

Granger, C. W. J. & Newbold, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics*, *2*(2), 111–120. doi: 10.1016/0304-4076(74)90034-7

Griewank, A. (2012). Who invented the reverse mode of differentiation? *Documenta Mathematica*.

Griliches, Z. & Intriligator, M. D. (Eds.). (1983). *Handbook of econometrics* (Vol. 1). Amsterdam: North-Holland. (Handbooks in Economics, Vol. 2)

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., ... Valko, M. (2020). Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in neural information processing systems (neurips)*.

Grimmer, J. & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297. doi: 10.1093/pan/mps028

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica*, *12*(0), 1-115. Retrieved from https://onlinelibrary.wiley.com/doi/abs/0012-9682(194407)12<iii:TPAIE>2.0.CO;2-F

Hahn, P. R., Murray, J. S. & Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, *15*(3), 965–1056. doi: 10.1214/19-BA1195

Hansen, B. E. (2005). Challenges for econometric model selection. *Econometric Theory*, *21*(1), 60–68. doi: 10.1017/S0266466605050048

Hansen, B. E. (2007). Least squares model averaging. *Econometrica*, *75*(4), 1175–1189. doi: 10.1111/j.1468-0262.2007.00785.x

Hansen, B. E. (2016). The risk if James-Stein and LASSO shrinkage. *Econometric Reviews*, *35*(8-10), 1456-1470. Retrieved from https://www.tandfonline.com/doi/full/10.1080/07474938.2015.1092799

Hartigan, J. A. & Wong, M. A. (1979). Algorithm as 136: A K-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *28*(1), 100–108. doi: 10.2307/2346830

Hassan, T. A., Hollander, S., van Lent, L. & Tahoun, A. (2019). Firm-level political risk: Measurement and effects. *The Quarterly Journal of Economics*, *134*(4), 2135–2202. doi: 10.1093/qje/qjz021

Hastie, T. & Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, *1*(3), 297–310. doi: 10.1214/ss/1177013604

Hastie, T., Tibshirani, R. & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer.

Hayashi, F. (2000). *Econometrics*. Princeton, NJ: Princeton University Press.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P. & Girshick, R. (2021). Masked autoencoders are scalable vision learners. In *arxiv preprint.*

He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. (2020, June). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr).*

Heckerman, D. (1995, March). *A tutorial on learning with bayesian networks* (Tech. Rep. No. MSR-TR-95-06). Microsoft Research.

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

Hennig, C. (2007, September). Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis*, *52*(1), 258–271.

Hirschber, J. G., Maasoumi, E. & Slotte, D. J. (2001). Clusters of attributes and well-being in the U.S.A. *Journal of Applied Econometrics*, *16*(3), 445-460.

Ho, J., Jain, A. & Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in neural information processing systems (neurips).*

Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. doi: 10.1080/00401706.1970.10488634

Jha, M., Qian, J., Weber, M. & Yang, B. (2024, February). *Chatgpt and corporate policies* (Working Paper No. 32161). National Bureau of Economic Research. Retrieved from https://www.nber.org/papers/w32161 doi: 10.3386/w32161

Johnston, J. (1963). *Econometric methods*. New York: McGraw–Hill.

Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). Springer. doi: 10.1007/b98835

Judge, G. & Harris, R. I. D. (1995, 05). Pcgive professional 8.0 and pcgive student 8.0. *The Economic Journal*, *105*(430), 776-786. doi: 10.2307/2235053

Kallus, N. & Mao, X. (2025). On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *Journal of the Royal Statistical Society Series B*, *87*(2), 480–509. doi: 10.1093/jrsssb/qkae099

Kingma, D. P. & Welling, M. (2013). Auto-encoding variational bayes. In *arxiv preprint.* Retrieved from https://arxiv.org/abs/1312.6114 (arXiv:1312.6114)

Kitagawa, T. & Tetenov, A. (2018). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, *86*(2), 591–616. doi: 10.3982/ECTA13288

Kleinberg, J. M. (2002). An impossibility theorem for clustering. In *Advances in neural information processing systems* (Vol. 15).

Kocsis, L. & Szepesvári, C. (2006). Bandit based monte-carlo planning. In *Machine learning: Ecml 2006* (Vol. 4212, pp. 282–293). Springer. doi: 10.1007/11871842_29

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95)*, *2*, 1137–1145.

Konishi, S. & Kitagawa, G. (2008). Various model evaluation criteria. In *Information criteria and statistical modeling* (pp. 239–254). Springer New York. doi: 10.1007/978-0-387-71887-3_10

Koopmans, T. C. (Ed.). (1950). *Statistical inference in dynamic economic models* (No. 10). New York: John Wiley & Sons.

Krolzig, H.-M. & Hendry, D. F. (2000). *Computer automation of general-to-specific model selection procedures* (Tech. Rep.). Institute of Economics and Statistics and Nuffield College, Oxford. Retrieved from http://fmwww.bc.edu/RePEc/es2000/0411.pdf

Künzel, S. R., Sekhon, J. S., Bickel, P. J. & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, *116*(10), 4156–4165. doi: 10.1073/pnas.1804597116

Lachenbruch, P. A. & Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics*, *10*(1), 1–11. doi: 10.1080/00401706.1968.10490530

Lai, T. L. & Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, *6*(1), 4–22. doi: 10.1016/0196-8858(85)90002-8

Langley, P. (2011). The changing science of machine learning. *Machine Learning*. doi: 10.1007/s10994-011-5242-y

Leamer, E. E. (1978). *Specification searches: Ad hoc inference with nonexperimental data*. New York: John Wiley & Sons.

Leamer, E. E. (1983a). Let's take the con out of econometrics. *The American Economic Review*, *73*(1), 31–43.

Leamer, E. E. (1983b). Model choice and specification analysis. In Z. Griliches & M. D. Intriligator (Eds.), *Handbook of econometrics* (Vol. 1, pp. 285–330). Amsterdam: North-Holland.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*(4), 541–551. doi: 10.1162/neco.1989.1.4.541

LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. doi: 10.1109/5.726791

Lee, D. D. & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791. doi: 10.1038/44565

Leeb, H. & Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, *21*(1), 21–59. doi: 10.1017/S0266466605050036

Li, Q. & Racine, J. (2004). Cross-validated local linear nonparametric regression. *Statistica Sinica*, *14*(2), 485–512.

Liao, Z. (2013). Adaptive GMM shrinkage estimation with consistent moment selection. *Econometric Theory*, *29*(5), 857-904. Retrieved from https://www.jstor.org/stable/24534478

Lieli, R. P., Hsu, Y.-C. & Reguly, Á. (2022). The use of machine learning in treatment effect estimation. In F. Chan & L. Mátyás (Eds.), *Econometrics with machine learning* (pp. 79–109). Cham: Springer International Publishing. doi: 10.1007/978-3-031-15149-1_3

Lighthill, J. (1973). Artificial intelligence: A general survey. In *Artificial intelligence: A paper symposium.* London: Science Research Council.

Lovell, M. C. (1983). Data mining. *The Review of Economics and Statistics*, *65*(1), 1–12.

Lundberg, S. M. & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems.* Retrieved from https://arxiv.org/abs/1705.07874 (arXiv:1705.07874)

Lütkepohl, H. (1991). *Introduction to multiple time series analysis*. Berlin and New York: Springer-Verlag.

Maasoumi, E. (1978). A modified Stein-like estimator for the reduced form coefficients of simultaneous equations. *Econometrica*, *43*(3), 695-703. Retrieved from www.jstor.org/stable/1914241

Maasoumi, E. (1980). A ridge-like method for simultaneous estimation of simultaneous equations. *Journal of Econometrics*, *12*(2), 161-176. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/0304407680900044

Maasoumi, E. (1993). A compendium to information theory in economics and econometrics. *Econometric Reviews*, *12*(2), 137–181. doi: 10.1080/07474939308800260

Maasoumi, E. & Phillips, C. B. (1982). On the behavior of inconsistent instrumental variable estimators. *Journal of Econometrics*, *19*(2-3), 183-201.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 281–297). Berkeley, CA: University of California Press.

Maddala, G. S. (1988). *Introduction to econometrics*. New York: MacMillan.

Mallows, C. L. (1973). Some comments on c_p. *Technometrics*, *15*(4), 661–675. doi: 10.2307/1267380

McCulloch, W. S. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133. doi: 10.1007/BF02478259

McInnes, L., Healy, J. & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint*. Retrieved from

https://arxiv.org/abs/1802.03426 (arXiv:1802.03426)

Meinshausen, N. & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(4), 417–473. doi: 10.1111/j.1467-9868.2010.00740.x

Minsky, M. L. & Papert, S. A. (1969). *Perceptrons: An introduction to computational geometry*. Cambridge, MA: MIT Press.

Molnar, C. (2019). *Interpretable machine learning: A guide for making black box models explainable.* https://christophm.github.io/interpretable-ml-book/pdp.html.

Molnar, C., Freiesleben, T., König, G., Casalicchio, G., Wright, M. N. & Bischl, B. (2021). Relating the partial dependence plot and permutation feature importance to the data generating process. *arXiv preprint*. Retrieved from https://arxiv.org/pdf/2109.01433 (arXiv:2109.01433)

Mullainathan, S. & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106. doi: 10.1257/jep.31.2.87

Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, *26*(4), 354–359. doi: 10.1093/comjnl/26.4.354

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and Its Applications*, *9*(1), 141–142. doi: 10.1137/1109020

Nadeau, C. & Bengio, Y. (2003). Inference for the generalization error. *Machine Learning*, *52*(3), 239–281. doi: 10.1023/A:1024068626366

Ng, A. Y., Jordan, M. I. & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, *14*, 849–856.

Nie, X. & Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, *108*(2), 299–319. doi: 10.1093/biomet/asaa076

Oprescu, M., Syrgkanis, V. & Wu, Z. S. (2019). Orthogonal random forest for causal inference. In *Proceedings of the 36th international conference on machine learning* (Vol. 97, pp. 4932–4941). PMLR.

Pesaran, M. H. & Smith, R. J. (1994). A generalized R^2 criterion for regression models estimated by the instrumental variables method. *Econometrica*, *62*(3), 705–710.

Phillips, P. C. B. (1986). Understanding spurious regressions in econometrics. *Journal of Econometrics*, *33*(3), 311–340. doi: 10.1016/0304-4076(86)90001-1

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106. doi: 10.1007/BF00116251

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286. doi: 10.1109/5.18626

Racine, J. & Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, *119*(1), 99–130. doi: 10.1016/S0304-4076(03)00157-X

Reguly, A. (2025). *Discovering heterogeneous treatment effects in regression discontinuity designs.* Retrieved from https://arxiv.org/abs/2106.11640

Robbins, H. & Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, 400–407.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., ... Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929. doi: 10.1111/ecog.02881

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. doi: 10.1037/h0042519

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. doi: 10.1038/s42256-019-0048-x

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. doi: 10.1038/323533a0

Schölkopf, B. & Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/4175.001.0001

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. doi: 10.1214/aos/1176344136

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, *7*(2), 221–264.

Shawe-Taylor, J. & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge: Cambridge University Press.

Shi, Z. (2016). Estimation of sparse structural parameters with many endogenous variables. *Econometric Reviews*, *35*(8-10), 1582-1608.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310. doi: 10.1214/10-STS330

Stock, J. H. & Watson, M. W. (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, *20*(2), 147–162. doi: 10.1198/073500102317351921

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *36*(2), 111–133. doi: 10.1111/j.2517-6161.1974.tb00994.x

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 44–47.

Sutton, R. S. & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, *16*(4), 437–450. doi: 10.1016/S0169-2070(00)00065-0

Tchetgen Tchetgen, E. J., Ying, A., Cui, Y., Shi, X. & Miao, W. (2024). An introduction to proximal causal inference. *Statistical Science*, *39*(3), 375–390. doi: 10.1214/23-STS911

Tenenbaum, J. B., de Silva, V. & Langford, J. C. (2000). A global geometric framework

for nonlinear dimensionality reduction. *Science*, *290*(5500), 2319–2323. doi: 10.1126/science.290.5500.2319

Theil, H. (1961). *Economic forecasts and policy* (2nd ed.). Amsterdam: North-Holland. (2nd revised edition)

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, *25*(3/4), 285–294. doi: 10.2307/2332286

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Tibshirani, R., Walther, G. & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423. doi: 10.1111/1467-9868.00293

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460. doi: 10.1093/mind/LIX.236.433

van der Maaten, L. & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, *9*, 2579–2605.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York: Springer. doi: 10.1007/978-1-4757-2440-0

Varma, S. & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, *7*(1), 91. doi: 10.1186/1471-2105-7-91

von Luxburg, U. (2010). Clustering stability: An overview. *Foundations and Trends in Machine Learning*, *2*(3), 235–274. doi: 10.1561/2200000008

von Luxburg, U. & Ben-David, S. (2005). Towards a statistical theory of clustering. *Manuscript*. Retrieved from https://cs.uwaterloo.ca/~shai/LuxburgBendavid05.pdf

Wager, S. (2025). *Causal inference: A statistical learning approach.* Retrieved from https://web.stanford.edu/~swager/causal_inf_book.pdf (Draft monograph)

Wager, S. & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242. doi: 10.1080/01621459.2017.1319839

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, *26*(4), 359–372.

Wattenberg, M., Viégas, F. & Johnson, I. (2016). How to use t-sne effectively. *Distill*. doi: 10.23915/distill.00002

Werbos, P. J. (1974). *Beyond regression: New tools for prediction and analysis in the behavioral sciences* (Unpublished doctoral dissertation). Harvard University, Cambridge, MA. (Ph.D. dissertation)

Wooldridge, J. M. (1997). Multiplicative panel data models without the strict exogeneity assumption. *Econometric Theory*, *13*(5), 667–678.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Yu, H., Chapman, B., Di Florio, A., Eischen, E., Gotz, D., Jacob, M. & Blair, R. H. (2019). Bootstrapping estimates of stability for clusters, observations

and model selection. *Computational Statistics*, *34*, 349–372. doi: 10.1007/s00180-018-0830-y

Yule, G. U. (1926). Why do we sometimes get nonsense-correlations between time-series? a study in sampling and the nature of time-series. *Journal of the Royal Statistical Society*, *89*(1), 1–64. doi: 10.1111/j.2397-2335.1926.tb01829.x

Zellner, A. (1985). Bayesian econometrics. *Econometrica*, *53*(2), 253–269.

Zhang, K. W., Janson, L. & Murphy, S. A. (2020). Inference for batched bandits. *Advances in Neural Information Processing Systems*, *33*, 9818–9829.

Zhao, Q. & Hastie, T. (2019). Causal interpretations of black-box models. *Journal of Business & Economic Statistics*. doi: 10.1080/07350015.2019.1624293

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320. doi: 10.1111/j.1467-9868.2005.00503.x