DOI: 10.1111/joes.12532

JOURNAL OF ECONOMIC SURVEYS WILEY

ARTICLE

Asset Pricing and Machine Learning: A critical review

Matteo Bagnara[1,2]

[1]Leibniz Institute for Financial Research SAFE, Frankfurt am Main, Germany

[2]Goethe University Frankfurt, Frankfurt am Main, Germany

Correspondence
Matteo Bagnara, Leibniz Institute for Financial Research SAFE, Theodor-W.-Adorno-Platz 3, 60323, Frankfurt am Main, Germany.
Email: bagnara@safe-frankfurt.de

Abstract

The latest development in empirical Asset Pricing is the use of Machine Learning methods to address the problem of the factor zoo. These techniques offer great flexibility and prediction accuracy but require special care as they strongly depart from traditional Econometrics. We review and critically assess the most recent and relevant contributions in the literature grouping them into five categories defined by the Machine Learning (ML) approach they employ: regularization, dimension reduction, regression trees/random forest (RF), neural networks (NNs), and comparative analyses. We summarize the empirical findings with particular attention to their economic interpretation providing hints for future developments.

KEYWORDS

empirical Asset Pricing, Machine Learning, risk premium, stochastic discount factor

1 | INTRODUCTION

In the last 40 years, the greatest endeavor in Asset Pricing has been documenting the properties of the stochastic discount factor (SDF) or pricing kernel (PK), which allows to price any asset with unknown future payoff, with the goal of understanding the determinants of asset returns. The well-documented empirical failures of the CAPM together with the work of Banz (1981), Rosenberg et al. (1985), Fama and French (1992), Carhart (1997) and others showing that a single economically motivated factor (market or aggregate consumption growth) is not enough to

explain differences in average stock returns (Mehra & Prescott, 1985), led us to focus on multi-factor models *à la* Fama and French (1993). The convenient shortcut adopted by the literature of treating unexplained portfolios as unknown risks and simply adding them to existing factors did not dwarf the search for new "anomalies," that is, assets whose risk-adjusted returns significantly spread over the distribution of one firm characteristic without a corresponding change in the exposure to some prespecified risk sources. Over time, a plethora of factors that supposedly predict cross-sectional variation in expected returns have been documented (Harvey et al., 2016; Hou et al., 2015). This "factor zoo" has been criticized either because of doubts on the actual usefulness of the factors proposed (McLean & Pontiff, 2016) or because of the struggles one faces when considering so many sources of risk in models which shall explain these empirical findings.

Besides these criticisms, a new strand of literature emerged, which tackles the factor zoo employing nonstandard statistical techniques known as Machine Learning (ML), which has the potential to recover complex patterns better than classical Econometrics thanks to its flexibility in high-dimensional settings (Karolyi & Van Nieuwerburgh, 2020). Some studies select only a few characteristics (Freyberger et al., 2020) or instrument factor loadings to summarize their information content (Kelly et al., 2019). Others combine these methods with economic restrictions (Lettau & Pelger, 2020b), while some others exploit them to circumvent daunting issues affecting traditional models (Giglio & Xiu, 2021). A growing level of attention has been paid to this topic as in several cases results achieved are outstanding, with findings that shed new light on the SDF structure. In view of these recent developments, practitioners and academics consider ML a powerful alternative to conventional approaches with tremendous potential ahead. A sound grasp of these models is imperative for those willing to keep abreast of the latest frontiers of research.

Applying ML to Asset Pricing involves econometric challenges the reader might not be familiar with considering its early stage. This paper critically summarizes the methods and the empirical discoveries of this promising new literature with special attention to their economic interpretation. It is organized as follows. Section 2 provides a general framework encompassing the goals pursued by the papers we examine and points out the econometric issues arising due to the factor zoo. Section 3 reviews and groups the most recent studies based on the main ML approach they use, dedicating a brief introduction to each of them. This serves both as high-level introduction to these techniques and as benchmark for more refined specifications adopted in current research. Section 4 summarizes the empirical findings assessing their contributions to our understanding of expected returns and attempts at giving guidelines for coordinated future efforts. Section 5 concludes.

2 | THE PLAYING FIELD AND THE ECONOMETRIC STRUGGLE

The studies we review belong to two main frameworks: predictions and factor models with associated SDF. The factor zoo results in different econometric issues for each of them. Since ML methods can be used to circumvent some of these, a brief overview is helpful.

2.1 | Predicting returns

One stream of literature tries to identify predictors for the cross-section of stock returns. Starting from a general additive prediction error model

$$r_{i,t+1} = \mathbb{E}_t\left[r_{i,t+1}\right] + \varepsilon_{i,t+1} \tag{1}$$

where $r_{i,t+1}$ is the return of stock i in excess of the risk-free rate between time t and $t+1$ and $\mathbb{E}_t[\varepsilon_{i,t+1}] = 0$ for $t = 1, \dots, T-1$ and $i = 1, \dots, N$, the conditional mean $\mathbb{E}_t[r_{i,t+1}]$ is often modeled as an unknown function $g(.)$ of some firm characteristics $z_{i,t}$ defined by the parameter vector θ. The model in Equation (1) becomes (Gu et al., 2020):

$$r_{i,t+1} = g(z_{i,t}; \theta) + \varepsilon_{i,t+1} \tag{2}$$

The vector $z_{i,t}$ contains P features supposed to forecast returns. In the factor zoo, P is in the order of hundreds (Harvey et al., 2016). With an ever-growing predictor space, researchers face issues arising in high-dimensional settings.[1] The first is *overfitting*: the richer the right-hand side, the harder the model fits the training sample (low bias), the less stable its performance with new data (high variance). While the in-sample R^2 benefits from new predictors, the out-of-sample (OOS) R^2 is low or even negative if these are not associated with the response, thus caution is needed when deciding what to include and what not. The second problem is *multicollinearity*: the risk of one regressor being expressible as linear combinations of others is remarkable with P large. Moreover, estimates of θ are likely imprecise due to *high collinearity*. Finally, traditional models soon become intractable including interactions and nonlinear transformations of all predictors.

2.2 | Factor models and stochastic discount factor

The second perspective adopted in many empirical studies is a multifactor model in the spirit of the arbitrage pricing theory (APT, Ross, 1976):

$$r_{i,t} = \alpha_{i,t} + f_t \beta'_{i,t} + \epsilon_{i,t} \tag{3}$$

where f_t is a vector of K risk factors with risk premia $\gamma_t = \mathbb{E}_{t-1}[f_t]$ and loadings $\beta_{i,t}$ and $\mathbb{E}_{t-1}[\epsilon_{i,t}] = \mathbb{E}_{t-1}[\epsilon_{i,t} f_t] = 0$. No-arbitrage implies $\alpha_{i,t} = 0 \forall (i,t)$, a condition equivalent to the existence of an SDF M_{t+1} satisfying the fundamental equation $\mathbb{E}_t[r_{i,t+1} M_{t+1}] = 0$. Such PK is an affine transformation of the tangency portfolio (Back, 2010), and is expressible as

$$M_t = 1 - (f_t - \mathbb{E}[f_t])b \tag{4}$$

where the weights b depend on factor mean $\mu = \mathbb{E}[f_t]$ and covariance matrix Σ, that is, $b = \Sigma^{-1}\mu'$. Equation (4) holds also for individual stock returns, but the large amount of noise in their time series makes the estimation of b imprecise. The same holds for $\alpha_{i,t}$ and $\beta_{i,t}$ in Equation (3). This issue led the profession to heavily rely on characteristics-based portfolio sorts, a tendency reinforced with the well-known approach of building new factors from long-short portfolio strategies (e.g., Fama & French, 1993, 1996). In a Fama–French world, unexplained alphas represent exposures to priced risk sources not yet included in a model for the cross-section. "There is no "alpha". There is just beta you understand and beta you do not understand" (Cochrane, 2011, p.1087). Add the omitted factors, the bias fades away and the fit improves. "[Fama & French (1993)] is an incomplete model for expected returns [...]. Motivated by this evidence [...] we add profitability and investment factors" (Fama and French, 2015, p.3). Appending relevant factors means increasing the Sharpe ratio (SR) attainable in the investment set spanned by the risk factors thus contributing to move closer to the multivariate mean–variance efficient portfolio (Merton, 1973; Fama, 1996).

Proxing risk factors through "hedge" portfolios (long-short portfolios based on firm characteristics generating anomalies) require two implicit assumptions, but none is innocuous. The first is that expected returns are constant over fixed parts of the characteristics distribution, that is, *within* the portfolios. The second one is that returns monotonically increase (decrease) in the characteristic of interest. As long as the difference between top and bottom portfolios shows significant spread in average returns, what happens in the middle is disregarded. This is especially concerning since portfolio means have troubles capturing the high variation of portfolio returns at the extremes of the distribution typically observed in the data (Fama & French, 2008). If sorts are too coarse, we might miss a significant relation or see one that is actually not there. Portfolio sorts further suffer from the *curse of dimensionality*: we simply cannot sort stocks along too many dimensions because we soon end up with more portfolios than stocks, and if they are not well-diversified they have small predictive power in Asset Pricing tests (Fama & French, 2015). Double sorts exacerbate these issues, because complex nonmonotonic relations can arise along each dimension.

Multifactor models like Equation (3) raise also misspecification concerns. First, factor exposures are usually not allowed to change over time, that is, $\beta_{i,t} = \beta_i$ like in Fama and French (2015). Loadings might instead depend on macroeconomic (Ferson & Harvey, 1991) or asset-specific variables (Kelly et al., 2019). Second, the PK functional form is likely complex and unknown (Chen et al., Forthcoming): leading theoretical contributions postulate nonlinearities between returns and state variables (e.g., Campbell & Cochrane, 1999; Bansal & Yaron, 2004), which impact the return factor structure, too. Third, Equation (3) suffers from omitted variable bias, the econometric bogeyman of all models. Fama and French (2015)'s *modus operandi* usually adds a new hedge portfolio to the existing factors if this results in roughly the same fit for benchmark test assets (e.g., 25 size-value-sorted portfolios) plus insignificant alphas for some small set of anomaly portfolios. This is equivalent to fit harder the data: the higher model complexity reduces the chance to perform well with yet unseen observations, hence portfolios sorted on newer characteristics are more likely to produce large pricing errors. It is not surprising that, even if the dimension of f keeps increasing, we find additional anomalies over time. We must take different paths, and ML is one of them.

2.3 | Pending questions

The factor zoo leaves some pending questions for the literature (Cochrane, 2011, p.1060).

"Which characteristics really provide independent information about average returns?" Although several anomalies are not robust phenomena as their returns decrease by two thirds after publication (McLean & Pontiff, 2016), a considerable number persist and we must deal with that. The question about how much *independent* information each one provides is still open, but some have recently addressed it (Freyberger et al., 2020; Feng et al., 2020).

"Does each new anomaly variable also correspond to a new factor formed on those anomalies?" This is crucial to find theoretical explanations for our empirical findings: do we accept anomalies as model failures (alphas) or are we just not properly accounting for them ("betas we do not know")? Alas, only few have tried to answer this question (Kelly et al., 2019).

"How many of these factors are really important?" This question raises the debate between those in favor and those against a *sparse* SDF. The crossroad is represented by: either accepting the existence of many factors and the need to account for them all (Kozak et al., 2020); or imposing a

low-dimensional SDF structure by regularization (Freyberger et al., 2020) or dimension reduction (Lettau & Pelger, 2020b).

Recently, some researchers have made excellent use of Cochrane (2011)'s advice that we need new methods for the factor zoo, and resorted to the ML field. Gu et al. (2020) provide a brilliant definition summarizing the main features of these techniques: they are generally used for high-dimensional predictions; they can be regularized to mitigate overfitting; their algorithms efficiently search and select among many model specifications. These attributes, often absent in standard methods, are well-suited to address and solve some of the problems mentioned above. Additionally, ML becomes a necessity when ordinary least squares (OLS) are not feasible (Karolyi & Van Nieuwerburgh, 2020). To keep up with the recent advances, one must understand which ML methods best fit Asset Pricing needs, how to employ them and what results they deliver. We address all these points reviewing state-of-the-art papers on the topic.

3 | A GUIDED LITERATURE REVIEW

We group recent studies into five categories according to the main ML approach they adopt: regularization, dimension reduction, regression trees/random forest (RF), neural networks (NNs), and comparative analyses. For every study, we indicate in Table 1 the new method proposed, data sample and covariates used, research question and main findings. We also report the Asset Pricing framework they belong to: prediction (Equation (2)) or factor models (Equation (3)). In the text, we dedicate a subsection to each ML area with a brief introduction to its mechanics.

We do not include in the table all the papers we mention for reasons of space but focus on the most relevant ones. We exclude papers using mainly Bayesian methods such as Bryzgalova et al. (Forthcoming). Although they are connected to regularization (Kozak et al., 2020), they do not belong to ML per se.[2] Bayesian inference is a statistical paradigm, which represent beliefs through prior and posterior *probability distributions*. ML denotes instead computational algorithms that learn complex data patterns in largely unstructured manners. Bayesian methods are neither in contrast nor in competition with ML. They are rather an alternative approach to inference that can be nested in other ML techniques (e.g., Bayesian neural networks, Neal (2012)), whose discussion goes beyond the scope of this review.

3.1 | Regularization

The simplest formulation of Equation (2) assumes that returns are linear in characteristics: $g(z_{i,t};\theta) = z'_{i,t}\theta$. Financial markets have weak signal-to-noise ratios, which means that there is a high degree of randomness resulting from forces of competition and profit maximization that wipe out most of predictability. Only unanticipated shocks (noise) really move markets (Israel et al., 2020). In these environments, OLS tend to overfit noise rather than extract the true signals if the right-hand side contains hundreds of characteristics. To get the "big picture" sacrificing some details to prioritize the strongest predictors, reducing the number of parameters is vital. This can be done through *regularization* (or penalization), a key concept in ML and "one of the first signs of the existence of intelligent inference" (Vapnik, 1998, p.9). It involves estimating model parameters adding a penalty term $\phi(\theta;\cdot)$ to a loss function $\mathcal{L}(\theta)$ to favor more parsimonious specifications, similarly to Akaike or Bayes information criteria. With OLS, we have a penalized sum of squared

TABLE 1 Studies grouped by ML method used

Authors (year)	ML area	Method	Framework	Test assets	Predictors/ Factors	Research question	Findings
Chinco et al. (2021)	Regularization	LASSO	Prediction	Individual stocks, 2005–2012	Lagged returns	Can variable selection support human intuition to predict returns?	High-frequency returns are forecasted by few short-lived predictors associated with news about fundamentals
Freyberger et al. (2020)	Regularization	Adaptive group LASSO	Prediction	Individual stocks, 1965–2014	62 firm characteristics	Which characteristics give independent information for the cross-section?	Past-return-based predictors stand out; Conditional and unconditional predictive power might substantially differ
Martin and Nagel (2022)	Regularization	Ridge	Prediction	Individual stocks, 1926–2019	Lagged returns	What happens to market efficiency in the "age of big data"?	In high-dimensional settings, there is predictability in-sample but none out-of-sample
Feng et al. (2020)	Regularization	Double-selection LASSO	Factor models	750 characteristics portfolios, 1976–2017	150 risk factors	What factors matter given a large set of existing ones?	Only few factors have explanatory power within the factor zoo (e.g., CMA, RMW, BAB, ROE, Quality-minus-Junk)
Kozak et al. (2020)	Regularization	Elastic net	Factor models	80 characteristics portfolios, 1973–2017	— (build SDF)	Is the SDF sparse?	Not enough redundancy in the factor zoo: SDF is inevitably dense
Kelly et al. (2019)	Dimension reduction	IPCA (Instrumented PCA)	Factor models	Individual stocks, 73 characteristics portfolios, 1962–2014	Latent factors	Are characteristics anomalies or compensation for risk?	Characteristics proxy for risk exposures; Most relevant: accounting and pat returns; Instrumenting betas is beneficial
Kelly et al. (Forthcoming)	Dimension reduction	IPCA (Instrumented PCA)	Factor models	Corporate bonds, 30 characteristics portfolios, 1999–2020	Latent factors	Which factors explain corporate bond returns?	Spread, duration, volatility, momentum, and spread-to-D2D explain corporate bond returns; Evidence of equity and bond markets integration
Büchner and Kelly (2022)	Dimension reduction	IPCA (Instrumented PCA)	Factor models	Index options, 15 characteristics portfolios, 1996–2017	Latent factors	Which factors explain index option returns?	Level, slope, and moneyness skew of volatility surface explain option returns
Lettau & Pelger (2020b)	Dimension reduction	RP-PCA (Risk-Premium PCA)	Factor models	37 characteristics portfolios, 1963–2017	Latent factors	Do weak factors matter for times series and cross-section?	Weak factors help explaining risk premia; Most relevant factors: market, trading frictions; past returns
Giglio and Xiu (2021)	Dimension reduction	PCA-based risk premia estimation	Factor models	647 portfolios (with bonds and currencies), 1976–2010	21 tradable and nontradable factors	Which factors are priced in the cross-section, given the others?	Most relevant factors: market, momentum, profitability, BAB, liquidity, intermediary, macro factors, stockholder's consumption growth

(Continues)

TABLE 1 Studies grouped by ML method used (continued)

Authors (year)	ML area	Method	Framework	Test assets	Predictors/ Factors	Research question	Findings
Pelger (2020)	Dimension reduction	PCA	Factor models	Individual stocks, 2004–2016	Latent factors	Which continuous and jump factors explain high-frequency returns?	High-frequency returns are explained by market, oil, finance, and electricity industry; Characteristics overnight, industry intraday
Moritz and Zimmermann (2016)	Regression trees	Random forest	Prediction	Individual stocks, 1963–2012	Lagged-returns-deciles	Which past returns are informative for the cross-section?	Short-term past returns explain more than longer ones, but interactions among the two matter
Bryzgalova et al. (2021)	Regression trees	Asset-Pricing Trees	Factor models	36 triple-sorted portfolios, 1964–2016	10 firm characteristics	Are large cross-sections of conventional sorts meaningful to build factors?	Large cross-sections of test assets are often redundant; Characteristics interactions matter
Chen et al. (Forthcoming)	Neural networks	GAN (Generative Adversarial Network)	Factor models	46 anomaly portfolios, Individual stocks, 1967–2016	46 firm characteristics, 178 macro variables	What is the SDF that best explains the hardest test assets to price?	Nonlinearities, test assets and macroeconomic information jointly matter; Most relevant: trading frictions, past returns
Gu et al. (2021)	Neural networks	Conditional Autoencoder	Factor models	94 chracteristics portfolios, 1957–2016	94 firm characteristics	Are characteristics anomalies or compensation for risk?	Nonlinear factor loadings-characteristics relationships give large empirical gains; Past returns, liquidity, risk measures
Gu et al. (2020)	Comparative-analyses	Several	Prediction	Individual stocks, 1957–2016	94 characteristics, 74 macro dummies, 8 macro variables	What is the best ML method to predict stock returns?	Best performing methods: NNs, boosted trees; Most relevant characteristics: past returns, liquidity, risk measures, valuation ratios
Bianchi et al. (2021)	Comparative-analyses	Several	Prediction	Treasury bonds, 1971–2018	Treasury yields; forward rates, 128 macro variables	Can ML methods help predicting bond returns?	Best performing method: group-ensembled NN; Most relevant predictors: inflation, money, credit
Avramov et al. (2022)	Comparative-analyses	Several	Prediction	Depending on replicated study	Depending on replicated study	Are ML investment good after transaction costs and in specific subsamples?	ML strategies mitigate downside risk, hedge during crises, do not decline recently; Most relevant: traditional anomalies

residuals (SSR):

$$\mathcal{L}^{OLS}_{regularized}(\theta; z_{i,t}) = \underbrace{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(r_{i,t+1} - z'_{i,t}\theta\right)^2}_{\text{OLS loss function}} + \underbrace{\lambda(1-\rho)\sum_{j=1}^{P}|\theta_j| + \frac{1}{2}\lambda\rho\sum_{j=1}^{P}(\theta_j)^2}_{\text{penalty } \phi(\theta;\lambda,\rho)} \quad (5)$$

The model in Equation (5) is known as *elastic net* (Zou & Hastie, 2005). The terms λ and ρ are *hyperparameters* or *tuning parameters* to be determined through *Cross-Validation* (*CV*, Hastie et al., 2009). The elastic net incorporates two models depending on the value of ρ. The *Ridge regression* (Hoerl & Kennard, 1970, $\rho = 1$) penalizes the estimator in the "l_2-space" for picking large values for θ, regardless their sign. Ridge performs *shrinkage*: it draws all coefficient estimates towards zero, but it still retains them all. λ controls the amount of shrinkage. When $\lambda = 0$, there is no penalization, that is, we have standard OLS. As λ increases, the model flexibility decreases, leading to higher bias but smaller variance. The *Least Absolute Shrinkage and Selection Operator (LASSO)* (Tibshirani, 1996, $\rho = 0$) does not shrink coefficients towards zero, rather it forces some to zero imposing an l_1-penalization. It achieves *selection* among predictors based on the value of λ maximizing the OOS performance. For each value of λ in Equation (5), there exists a unique γ that solves the following equivalent constrained minimization problem:

$$\min_{\theta} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(r_{i,t+1} - z'_{i,t}\theta\right)^2 \text{ subject to } \sum_{j=1}^{P}(\theta_j)^2 \leq \gamma\left(\sum_{j=1}^{P}|\theta_j| \leq \gamma\right) \quad (6)$$

for $\rho = 1$ ($\rho = 0$). To understand the differences between the two cases, consider $P = 2$. The constraint region is a disk for Ridge and a diamond for LASSO. Unlike the disk, the diamond has corners and if the first tangency point between the SSR contour lines and the constraint lies there, one parameter is set to zero. When $P > 2$, we have a sphere (Ridge) and a rhomboid (LASSO) with many corners, flat edges and faces (Hastie et al., 2009): there are many chances that several coefficients are set to zero with such an "edgy" constraint region. This is why LASSO delivers sparse solutions.

Ridge and LASSO estimates can be interpreted as the posterior distribution moments for given Bayesian priors. Ridge gives posterior mean and mode under a Normal model for the distribution of returns given the predictors and a Normal prior for the parameters. LASSO gives the posterior mode given Laplace priors (Athey & Imbens, 2019). Some studies exploit this link to map the models they first develop following Bayesian approaches into penalized estimators (e.g., Kozak et al., 2020). With ML, however, the optimal amount of regularization is chosen by CV without specific priors. CV carries out data resampling to tune hyperparameters on subsets of data not contained in the training set (*validation set*). For example, *K-fold CV* splits the data into K subsamples (*folds*), and each time it fits the model omitting the kth subsample, for $k = 1, 2, \ldots K$. The optimal value for the tuning parameter minimizes the prediction error across all the K validation subsamples initially excluded.[3] Model performance is evaluated in the test set. Besides this, both methods have pros and cons depending on the situation. For instance, LASSO struggles with strongly correlated variables, as it is indifferent between including one, the correlated one, or even both covariates as long as they get nonzero coefficients (Nagel, 2021). Elastic nets encourage both shrinkage and selection to skirt the limitations of each of its nested models.

Regularization is often embedded in the loss function of other ML techniques to mitigate overfitting. In Asset Pricing, special penalization specifications enhance traditional methods aimed at predicting returns or identifying risk factors, as we review below.

3.1.1 | Regularization in asset pricing

Chinco et al. (2019) propose LASSO as off-the-shelf method to improve 1-min-ahead return forecasts when the space of candidate predictors is large. Differently from Equation (2), here the predictors are not firm characteristics but rather the lagged returns of the entire NYSE universe during the previous 3 min. "Betting on sparsity" (Hastie et al., 2009) with LASSO regressions over 30-min rolling windows results in an OOS adjusted R^2 of 2.47% with significant increases in explained variation according to the formal (Giacomini & White, 2006) test for alternative conventional models including several AR(p) models, Fama & French (1993), and combinations of them augmented with firm characteristics. A trading strategy based on the model predictions yields a 1.79 OOS annualized SR and a significant alpha with respect to a four-factor model with market, size, value, and momentum without significant loadings on any of them except for the market. The authors also calculate the probability of LASSO selecting a certain predictor with a simple logit model, and find that size, liquidity, volume, and industry do not influence it, whereas they are relevant at weekly or monthly horizons. In this sense, the selected predictors are "unexpected". LASSO selects the same lagged return for no longer than 15 min, that is, predictors are short-lived, and retains mainly returns of stocks with recent news about their fundamentals, in line with economic intuition. The conclusion is that plain-vanilla selection methods like LASSO can support human intuition towards economically meaningful choices.

Freyberger et al. (2020) use a variation of LASSO named *adaptive group LASSO* paired with a nonparametric method (quadratic splines) to establish which firm characteristics provide independent information for the cross-section of expected returns. The conditional expectation in Equation (1) is a potentially nonlinear function $m_{t+1}(.)$ of the characteristics: $m_{t+1}(z_1,\dots,z_P) := \mathbb{E}[r^i_{t+1}|z_{1,i,t}=z_1,\dots,z_{P,i,t}=z_P]$. Once characteristics are normalized into the unit interval with a rank-transformation, as usual in ML to eliminate sensitivity to outliers, there exists an equivalent formulation with these new features $\tilde{z}_{i,t}$, that is, $\tilde{m}_{t+1}(\tilde{z}_1,\dots,\tilde{z}_P)$. To sidestep the shortcomings of portfolio sorts mentioned above, the authors use the additive model $\tilde{m}_{t+1}(\tilde{z}_1,\dots,\tilde{z}_P) = \sum_{j=1}^{P}\tilde{m}_{t+1,j}(\tilde{z}_j)$ offering a rate of convergence independent from P, where $\tilde{m}_{t+1,j}$ are unknown functions to estimate through adaptive group LASSO. The method first splits the support of every characteristic into L intervals similar to portfolios and fits quadratic splines in each of them. In this way, $\tilde{m}_{t+1,j}(\tilde{z}_j) \approx \sum_{\kappa=1}^{L+2} b_{t,j,\kappa}s_\kappa(\tilde{z}_j)$, with $s_\kappa(\tilde{z})$ basis functions and $b_{t,j,\kappa}$ must be estimated. Second, it employs the group LASSO (Huang et al., 2010), which drops not one but all the coefficients associated with a given $\tilde{z}_j$ in $\tilde{m}_{t+1,j}(\tilde{z}_j)$ if it does not help predicting returns:

$$\hat{b}_t = \underset{b_{j,\kappa}:\, j=1,\dots,P;\kappa=1,\dots,L+2}{\arg\min} \sum_{i=1}^{N}\frac{1}{N}\left(r_{i,t+1} - \sum_{j=1}^{P}\sum_{\kappa=1}^{L+2} b_{j,\kappa}s_\kappa(\tilde{z}_{j,i,t})\right)^2 + \lambda_1\sum_{j=1}^{P}\left(\sum_{\kappa=1}^{L+2} b^2_{j,\kappa}\right)^{1/2} \tag{7}$$

where $\hat{b}_t$ is an $(L+2)\times P$ vector. The *adaptive* step consists in minimizing Equation (7) again, this time multiplying the parameters $b^2_{j,\kappa}$ by weights, which nullify characteristics with a zero squared contribution over all intervals that might have been selected in the first step. This is one of the few

models able to disentangle the *incremental* predictive power of individual characteristics beyond their relevance due to the correlation with other measures. Out of 62 predictors from the anomaly literature, several of them like idiosyncratic volatility are not associated with expected returns after conditioning for other covariates. Even if the set of selected predictors varies depending on interpolation points, data subsamples and over time, 41 are never chosen. The group surviving the selection in most cases mainly includes past-return-based predictors, except for change in shares outstanding and standardized unexplained volume (SUV). Some of these characteristics are easily identified as proxies for risk exposures, while others require mispricing-based explanations. In contrast to their formulation, the main results presented are obtained with time-invariant mean functions $\hat{m}_j(\tilde{z})$ estimated over the entire sample to get more precise estimates. Rolling estimations show that the relation between stock returns and some predictors might vary substantially over time: predictability might disappear and reappear in different periods.

Return predictability in high-dimensional settings is investigated also in Martin and Nagel (2022). In the "age of big data," investors observe a great number of signals and might incorporate ML when learning about cross-sectional anomalies. Owing to the large amount of noise in past cash flows, their beliefs update imperfectly and slowly, even with Bayesian learners with correct priors about the data-generating process. As a result, an econometrician who looks at the data *ex-post* finds predictability, but sees none OOS. Exploiting the connection between regularization and Bayesian methods explained earlier, the model implications are corroborated running Ridge regressions with a large set of standard predictors to mimic investors' learning process. Predictability results indeed much weaker OOS than in-sample and there is substantial decay over time. In the past, investors could not process information as effectively as today, as if they had only bounded rationality, which induced excessive shrinkage or sparsity in their forecasting models, hence anomalies would occur. The study concludes that in-sample predictability tests are ill-suited to uncover risk premia and suggests to focus on OOS periods to test the efficient market hypothesis.

Feng et al. (2020) introduce regularized two-pass cross-sectional regressions based on the *double-selection LASSO* of Belloni et al. (2014) to establish the marginal contribution of a new factor controlling for high-dimensional sets of existing risk sources. A distinctive feature is the focus on model selection: factors that do not predict the cross-section might still be useful to reduce omitted variable bias. Consider a linear model like Equation (3), where factors f_t are split into g_t, the new factors to be tested, and h_t, the known factors at time t. In standard two-pass regressions (Fama & MacBeth, 1973), controlling for the entire factor zoo is inefficient or even infeasible due to the *curse of dimensionality*, but "to cherry-pick a handful of control factors" (Feng et al., 2020, p.1336) is an ad hoc solution that omits potentially relevant factors biasing the risk premia estimates. Furthermore, one needs to look at the SDF loadings to understand whether a factor is useful or not (Cochrane, 2009): investors would still pay a nonzero risk premium to hedge against risk sources correlated with the "true" ones regardless their contribution to the PK. Hence, a double-selection procedure is introduced, which modifies the first step using covariances instead of betas and avoids omitted variable bias. First, a cross-sectional LASSO regression of average returns on sample covariances between h_t and r_t selects the factors. Then, the covariance between returns and each factor in g_t is regressed on the covariance between r_t and all h_t with LASSO. This step seeks factors potentially missed in the first stage because they have low explanatory power but that may induce a bias if excluded. Finally, post-selection SDF weights are estimated regressing r_t on covariances of returns with both h_t and the selected g_t.

To mimick the discovery process over time, 150 risk factors proposed between 2012 and 2016 are tested conditional on the ones known up to that point, using several bivariate sorts based

on traditional characteristics to build test assets. Due to the search for omitted factors, first- and second-step LASSO produce strikingly different results, selecting usually 4 and 20 up to 80 factors, respectively. As in other studies, most new factors turn out redundant, given the others. Among those that matter are Hou et al. (2015)'s investment factor, intermediary capital (He et al., 2017), Quality Minus Junk (Asness et al., 2019), Betting Against Beta (BAB, Frazzini and Pedersen, 2014), Conservative Minus Aggressive and Robust Minus Weak (CMA and RMW, Fama and French (2015)), and Return On Equity (ROE, Hou et al. (2015)). Past-return variables are less important than in Freyberger et al. (2020), although both methods rely on LASSO. Room for further research is given by the possibility of replacing LASSO with another technique in the first stage.

Kozak et al. (2020) is a major contribution in this review. The paper estimates the SDF weights b in Equation (4) through a Bayesian approach with economically motivated priors that can be mapped into regularized parameter estimation along the lines of what discussed in Section 3.1, hence its inclusion here. The authors acknowledge the multidimensional challenge of the factor zoo but argue that "[…] the multi-decade quest to summarize the cross-section of stock returns with sparse characteristics-based factor models […] is ultimately futile. There is simply not enough redundancy" (p.21). We lack redundancy because we often need several observable proxies to capture the effect of factors justified by theory that we cannot directly measure, like expected profitability and investment in the q-theory model (Lin & Zhang, 2013). Each proxy gains at least some predictive power only thanks to its correlation with the unobservable factors entering the true SDF.

A naïve SDF-weight estimator using sample moments $\hat{b} = \bar{\Sigma}^{-1}\bar{\mu}'$ overfits the data in high-dimensional settings and thus needs regularization. In a Bayesian setting, first and second factor moments should be linked together in investors' beliefs since, intuitively, factors with high returns should be themselves relevant risk sources or be highly exposed to the true ones. The authors choose a prior from a family of corresponding shrinkage estimators including those used in Pástor (2000) and Pástor and Stambaugh (2000), which leads to an estimator $\hat{b}$ for the posterior mean of b shrinking the SDF coefficients towards zero based on factor volatilities. Alternatively, one finds the same $\hat{b}$ with Ridge-style regressions, either minimizing the model cross-sectional R^2 with an l_2-penalty on the maximum SR, or minimizing the HJ-distance (Hansen & Jagannathan, 1991) subject to a squared penalty on b:

$$\hat{b} = \arg\min_{b} (\bar{\mu}' - \bar{\Sigma}b)'\bar{\Sigma}^{-1}(\bar{\mu}' - \bar{\Sigma}b) + \lambda b'b \tag{8}$$

Despite being fierceful advocates for density, Kozak et al. (2020) do not exclude sparsity a priori and include an additional l_1-penalization in Equation (8) hoping to set to zero some redundant factors, obtaining a model similar to an elastic net. The method is tested on a cross-section of 80 standard anomaly based factors to ensure that it contains more redundancy than typical size-value-sorted portfolios spanned by few factors (Lewellen et al., 2010). The "misbelief" in sparsity is justified by the findings: there is almost no redundancy in the test assets, as the elastic net essentially switches off the l_1-penalization to maximize the OOS R^2. This happens because the l_2-penalization already pushes many coefficients close to zero and almost no selection takes place with many strongly correlated covariates. Therefore, we can only say which factors are *the most* relevant looking at their SDF weight: industry momentum and relative reversals, seasonality, earnings surprises, market, ROE, and momentum among others. Introducing interactions among factors reveals that most of them are superfluous, in contrast to what others find (e.g., Bryzgalova et al., 2021). The only result speaking in favor of sparsity is the good OOS performance of a

low-dimensional SDF built using the factor *Principal Components (PCs)*. However, since these are still built rotating all factors, the main message remains that "Sparsity is generally elusive" (p.21).

3.2 | Dimension reduction

Both those who "bet on sparsity" and those who *impose* it to avoid high-dimensional problems would like to reduce the dimension of the factor zoo without losing valuable information for the cross-section. *Dimension reduction* methods compress the data into a lower-dimensional space. Prominent examples are *cluster analysis*, *principal component analysis (PCA)*, and *partial least squares* (*PLS*, Gu et al. (2020)). We focus here on PCA, the method that has been applied most often in Asset Pricing based on our knowledge.

PCA can be employed both for forecasts and in factor models. In predictive regressions like Equation (2), PCA first summarizes a large set of covariates into few linear combinations, which best describe returns, and then uses them as predictors. A classical example of such *principal component regression (PCR)* is in Gu et al. (2020). PCA for factor models like Equation (3) is described next referring to the specification used in Lettau & Pelger (2020b), who build on Connor and Korajczyk (1986, 1988, 1993) and Bai and Ng (2002).

We start from an "approximate" multifactor model which follows Ross (1976), where by assumption there is no mispricing and therefore no alpha:

$$\begin{aligned} r_{i,t} &= f_t \beta_i' + \epsilon_{i,t} \\ \underbrace{R}_{T\times N} &= \underbrace{F}_{T\times K}\underbrace{B'}_{K\times N} + \underbrace{\epsilon}_{T\times N} \end{aligned} \tag{9}$$

where in the second line, we stack the time series altogether and matrix dimensions are analogous to Lettau & Pelger (2020b). PCA applies singular value decomposition (SVD) to R, or, alternatively, *eigen*-decomposition to its sample covariance matrix $\bar{\Sigma}_R := \frac{1}{T}R'R = VD^2V'$ when returns are demeaned.[4] The eigenvalues are the elements d_j^2 in D^2 with associated eigenvectors v_j from V, for $j = 1, 2, \dots, N$. The PCs are linear combinations of returns with eigenvectors as weights, that is, Rv_j. Since $Var(Rv_j) = d_j^2/N$, eigenvalues are a natural metric to sort PCs: the first PC has the highest eigenvalue and thus the largest variance among all combinations of the columns of R. Each subsequent kth PC has the kth largest variance, conditional on being orthogonal to all previous PCs, for $k = 2, 3, \dots, K$, where K is the desired number of components. PCs constitute the closest subspace (e.g., surface) to the data cloud, thereby providing a K-dimensional approximation in terms of Euclidean distance (Hastie et al., 2009). In latent factor models like Equation (9), the first K eigenvectors form the $K \times N$ loading matrix $\hat{B}$ on the K orthogonal factors with the largest variance, that are estimated as $\hat{F} = R\hat{B}(\hat{B}'\hat{B})^{-1}$. Minimizing the quadratic loss function corresponding to Equation (9) choosing jointly factors and loadings produces the same results (Stock & Watson, 2002).

How does one choose the number of components K? The most straightforward way is using the *proportion of variance explained (PVE)*. The PVE for the kth PC is the ratio of its variance to the total variance of the data. One keeps estimating PCs until their cumulative PVE reaches a user-defined target. This procedure can guide the choice of the number of factors in theoretical models, as it happens for the term structure of interest rates, where the first three PCs usually

representing its level, slope, and curvature capture most of the variation. Alternatively, one can stop when the contribution of the kth PC is negligible (Hastie et al., 2009) or resort to statistical estimation (Pelger, 2019). Lastly, some studies treat K as a hyperparameter to be found minimizing the squared prediction error in the validation set (Gu et al., 2020). This approach is more OOS-oriented and fits well the typical Machine Learning paradigm that learns patterns in the data without relying just on in-sample information.

As PCA combines all regressors into the linear combinations best describing the data, it is particularly useful in situation with high collinearity, where shrinkage methods might lead to sub-optimal forecasts (Gu et al., 2020). PCA has strong explanatory power in many applications, but it is prone to overfitting. Another shortcoming is its instability to data structure. If one covariate captures a disproportionate amount of the data variation, their first PC roughly coincides with it. Now, if that covariate was split into L "smaller covariates," then the first L PCs would approximately correspond to those L parts of the original covariate. Data do not change, but results differ (one versus many PCs). When identifying dynamic factors with high-dimensional data, PCA is more efficient than maximum likelihood estimation (MLE), which requires estimating many parameters (Stock & Watson, 2002). There exists also nonlinear PCA (Hastie et al., 2009) and Bayesian PCA (Bishop, 1999), which have not been explored in Asset Pricing yet.

3.2.1 | Dimension reduction in asset pricing

As mentioned above, we focus on PCA. Chamberlain and Rothschild (1983) and Connor and Korajczyk (1986, 1988) provide early contributions. More recently, Kozak et al. (2018) show that a low-dimensional SDF built using PCs of the test assets covariance matrix can easily price several anomaly portfolios. However, this tells us nothing about whether the economy is rational: mispricing can still persist since arbitrageurs exploit near-arbitrage opportunities only insofar as they are orthogonal to common risk sources and are reluctant to trade aggressively against anomalies that exposes them to factor risk.

Kelly et al. (2019) adapt standard PCA allowing time variation in factor loadings. In the IPCA (*instrumented* PCA), these are linear functions of characteristics with predictive power for the cross-section of returns:

$$\begin{aligned} r_{i,t} &= \alpha_{i,t-1} + f_t \beta'_{i,t-1} + \epsilon_{i,t} \\ \alpha_{i,t-1} &= z'_{i,t-1}\Gamma_\alpha + \nu_{\alpha,i,t-1} \\ \beta_{i,t-1} &= z'_{i,t-1}\Gamma_\beta + \nu_{\beta,i,t-1} \end{aligned} \tag{10}$$

where $z_{i,t-1}$ is a vector with $P = 36$ firm characteristics. "Instrumenting" betas enables IPCA to incorporate the information contained in $z_{i,t-1}$ to describe returns, overcoming a major shortcomings of PCA, namely that it is a static model ill-suited to estimate conditional factor models. IPCA achieves dimension reduction through Γ_β, which maps a large number of features into K factors in f_t without resorting to portfolio sorts. In this way, it requires much fewer parameters compared to static-beta models. For example, Fama and French (2015)'s model plus momentum estimates 68712 parameters for individual stocks in the period considered, while IPCA needs 3816. Hence, this technique works well in high-dimensional environments. Operationally, IPCA finds

Γ_β, Γ_α, and f_t to minimize

$$\min_{\Gamma_\beta,\Gamma_\alpha,f_t} \sum_{t=1}^{T} (r_t - Z_{t-1}\Gamma_\alpha - Z_{t-1}\Gamma_\beta f_t')'(r_t - Z_{t-1}\Gamma_\alpha - Z_{t-1}\Gamma_\beta f_t') \tag{11}$$

where r_t is a $N \times 1$ vector and Z_{t-1} is an $N \times P$ matrix. The method is easily understood considering the application of SVD to "characteristic-managed portfolios" $x_t = Z_{t-1}' r_t / N_t$. From this perspective, the factors are linear combinations, or "portfolios of portfolios," that best explain the covariation among managed portfolios. The IPCA allows three types of tests. First, testing $H_0 : \Gamma_\alpha = 0$ comparing a restricted and an unrestricted model ($\Gamma_\alpha = 0$ and $\Gamma_\alpha \neq 0$, respectively), one examines whether characteristics give rise to "genuine" risk factors because they proxy for exposures to common risks commanding a premium, or if they are mere anomalies. The investigation of this point is hard to find elsewhere in the literature. Second, adding observable factors in the spirit of Fama and French (2015) in Equation (10), one can test whether latent factors have explanatory power beyond the rest. Third, one can verify which characteristics significantly contribute to the loadings $\beta_{i,t-1}$, controlling for the others. All these tests utilize Wald-type statistics backed by robust bootstrapping procedures.

IPCA is applied to data from Freyberger et al. (2020). Model performance is measured both with total and predictive R^2. The latter represents the variation explained by model-implied *conditional* returns when factors are fixed at their average. The main finding is that "characteristics are covariances," which means they proxy for risk exposures: for $K > 1$, the unrestricted model consistently outperforms the restricted one. IPCA performs particularly good for conditional returns, as its predictive R^2 is more than three times that of Fama and French (2015) plus momentum for individual stocks, and it explains almost 99% of the in-sample variation of managed portfolios with six factors. Its winning ingredients are both the use of unspecified statistical factors and dynamic betas, which enable good results also OOS. The authors interpret IPCA factors looking at the weights Γ_β in the first K factor loadings. Factor 1 is a firm-value factor dominated by size and book assets. Factor 2 is mostly determined by market beta, factors 3 and 4 by momentum and Short-Term Reversal (STR) and factor 5 is a mixture of many "less important" characteristics. Except for book assets, these ones are also the most relevant characteristics according to the contribution to the model R^2 in several robustness tests. In sum, accounting and past-return-based characteristics predominantly contribute to the stock return variation, in line with the cross-sectional Asset Pricing tradition.

IPCA has proved successful also in other contexts. Kelly et al. (Forthcoming) show that the method captures the bulk of momentum and long-term reversal anomalies as it recovers the time-varying exposures to priced latent factors that these characteristics pick up. STR still generates mispricing, however, probably because its effect is partially driven by the illiquidity of small firms (Asness et al., 2014) and does not stem from systematic risk compensations. Kelly et al. (Forthcoming) employ IPCA to extract latent factors from corporate bond returns. They use 30 bond characteristics (including a constant) with predictive power from the literature to model conditional loadings of individual assets and characteristic-managed portfolios. One factor alone explains 41% and an impressive 93% of their variation, respectively. With $K = 5$, the model surpasses (Bai et al., 2019)' and Fama and French (2015)'s models both in-sample and OOS and cannot reject the hypothesis of zero-pricing error. The risk factors are dominated by option-adjusted bond spread, duration, bond volatility, equity momentum, and spread-to-distance-to-default ratio. These characteristics together with few others mostly contribute to the model fit, too. Even if IPCA factors differ for equity and bonds, they produce remarkably similar bond risk premia estimates, hence the authors claim that the integration between these two markets is stronger than

previously documented. Finally, Büchner and Kelly (2022) use IPCA on delta-hedged returns of S&P500 index options. Due to their short lifespan and rapidly changing risk attributes, options represent hard challenges for traditional models. Instrumenting betas with option time-to-maturity, implied volatility, embedded leverage, and Black–Scholes–Merton "Greeks" (delta, theta, gamma, vega, volga) recovers statistical factors explaining between 72% ($K = 1$) and 90% ($K = 5$) of the movements in individual index options. Static PCA and observable factor models augmented with option-specific risk factors such as BAB and straddle factor (Coval & Shumway, 2001) are outmatched. Using the bootstrapping procedures of Kelly et al. (2019), the study concludes that option characteristics are not just anomalies but compensation for risk exposures to mainly three factors, interpretable as level, maturity slope, and moneyness skew of the implied volatility surface. The characteristics contributing the most to explain option returns are implied volatility, vega, and gamma.

Lettau & Pelger (2020b) modify traditional PCA to recover latent factors depending on their ability to simultaneously explain time series and cross-section of stock returns. Statistical factor analysis in the spirit of Connor and Korajczyk (1986, 1988) is a second-order method, which identifies factors depending only on their comovements with the data, disregarding the information about the first moments. With low signal-to-noise ratios only "strong" factors affecting numerous assets and/or having large variance are picked, while "weak" factors are not recovered. This is problematic for Asset Pricing: even if they do not generate much comovement among stocks, they can still foster the proliferation of anomalies and capture differences in risk premia. To circumvent this limitation, *risk-premium PCA (RP-PCA)* augments the standard PCA problem of minimizing time-series pricing errors with a penalty term on cross-sectional pricing errors:

$$\widehat{F}_{\text{RP-PCA}},\ \widehat{B}_{\text{RP-PCA}} = \underset{B,F}{\arg\min}\ \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(r_{i,t} - f_t\beta_i')^2 + \lambda\frac{1}{N}\sum_{i=1}^{N}(\bar{r}_i - \bar{f}\beta_i')^2 \tag{12}$$

where $\bar{r}_i$ is the sample mean of excess returns, which must not be restricted to zero: thanks to the penalty term, in RP-PCA a factor signal strength depends also on its first moment besides its volatility. If the signal strength is high enough relative to noise, the factor is identified; otherwise it is discarded as a weak one. Alternatively, RP-PCA extracts factors and loadings through an SVD on a modified covariance matrix $\Sigma_{RP} := \frac{1}{T}R'R + \lambda\overline{RR}'$ where λ determines the weight of cross-sectional errors as in Equation (12). When $\lambda = -1$, the model reduces to standard PCA. The model-implied SDF is then obtained from Equation (4) with the estimated factors. Lettau and Pelger (2020a) provide statistical properties and derivations.

Empirically, RP-PCA is employed to extract factors from single-sorted anomaly portfolios from Kozak et al. (2020) and other double-sorted portfolios. CV sets $K = 5$: RP-PCA delivers a sparse SDF. The method is superior to standard PCA and Fama & French (1993)'s model in terms of SDF SR, pricing errors and residual variance, both in-sample and OOS. Similarly to Kelly et al. (2019), we can attach economic meaning to the extracted factors inspecting the weights of the test portfolios. Unsurprisingly for latent factor models, the first one is the market. Portfolios in all groups contribute to the second factor, while the third one is a value factor. The most important portfolios for the fourth factor are related to momentum, interactions with value and trading frictions and those belonging to the reversal category dominate the fifth one. The rest of the factor zoo is largely redundant, as other studies claim. While the market affects both time-series and cross-sectional fit, factors 2 and 5 impact only the cross-section and factors 3 and 4 only the time series. In comparison, standard PCA factors are tougher to interpret as the factor weights do not provide a clear-cut grouping. The study also finds that sparse mean–variance optimization produces higher

SRs than shrinkage, in contrast to Kozak et al. (2020). RP-PCA can be combined with either IPCA or Kozak et al. (2020)'s approach to improve model performance. It is therefore flexible enough to support other methods accounting for important factors that otherwise might be missed, making it a noteworthy contribution to the literature.

Giglio and Xiu (2021) provide a *three*-pass methodology to estimate risk premia valid even with omitted factors and measurement error. The factors f_t in Equation (3) can always be split into potentially correlated and unobservable innovations ν_t and their risk premia γ:

$$r_t = \beta\gamma' + \beta\nu_t' + \epsilon_t \tag{13}$$

where β is an $N \times K$ matrix. Estimation of γ for a set of $d < K$ factors controlling for the remaining ones typically occurs through two-pass regressions (Fama & MacBeth, 1973) and the mimicking portfolio approach. In the first method, omitting some of the K components of ν_t entails a "compunded" bias for γ: the bias in time-series regression loadings further distorts the risk premia estimated with cross-sectional regressions. In the second method, the bias results from missing assets in the maximally correlated portfolio. Measurement errors arise when proxying for an unobserved factor with an observable variable g_t. Assume that $g_t = \delta + \eta\nu_t + \xi_t$, where ξ_t introduces the error in measuring ν_t through g_t. Even if $\text{Cov}(\xi_t, \nu_t) = 0$, potential correlation between ξ_t and ϵ_t in Equation (13) biases the risk premium estimate for the proxy g_t, that is, $\gamma_g = \eta\gamma$. To bypass both kinds of bias, the authors exploit a property of risk premia called *rotation invariance*: the product $\eta\gamma$ can be identified up to an arbitrary full-rank factor rotation of the type $\hat{\nu}_t = \nu_t H$, which means the risk premium of a single component of g_t is invariant to how the remaining $K-1$ are rotated, as long as they span the same risks as the true model. Hence, we need neither H nor ν_t explicitly but only $\hat{\nu}_t$, which can be consistently identified by PCA if factors are sufficiently strong (Bai, 2003). Following this logic, the three-pass methodology first extracts factors $\hat{\nu}_t$ and loadings from a large panel of test assets with PCA. Second, it computes their risk premia regressing average returns on estimated betas. Third, a mimicking portfolio procedure recovers the relation between g_t and ν_t removing the effect of measurement error. The risk premium of g_t is the product between third-step loadings and second-step risk premia. The three-pass method can be interpreted both as a PCA-augmented cross-sectional regression and as a regularized mimicking portfolio, where the PCs of the test assets play the role of controls and the space where to project factors onto, respectively, instead of using arbitrary assets.

The method is used to test both tradable (e.g., Fama and French, 2015) and nontradable factors (e.g., Pástor & Stambaugh, 2003). A model with seven factors delivers a 59% R^2 on a cross-section of 647 portfolios including treasuries, currencies, and corporate bonds. The three-pass approach produces risk premia estimates that are very close to average returns of tradable factors. The other standard methods give instead unstable estimates due to omitted variables. The most important factors are market, momentum, profitability, and BAB. Among nontradable ones, only liquidity and intermediary factors together with some macro variables and stockholders' consumption growth have economic and statistic significance. Results are robust to including RP-PCA or IPCA in the first step.

Pelger (2020) identifies continuous, jump, and overnight factors in high-frequency data with PCA. Changes in the $N \times 1$ vector of log prices $P(t)$ have the following factor structure:

$$dP(t) = B^C dF^C(t) + B^D dF^D(t) + B^N dF^N(t) + de(t) \tag{14}$$

where B^C, B^D, and B^N are continuous, jump, and overnight loadings on the corresponding K-dimensional factor processes F^C, F^D, and F^N, respectively. The return matrix R has elements

$R_{i,j} = P_i(t_{j+1}) - P_i(t_j)$ for $i = 1, \dots, N,\ j = 1, \dots, T$. Hence, the model for high-frequency returns becomes $R = \Delta F B' + \Delta\epsilon$, where the Δ operator represents changes from t_j to t_{j+1}. If returns at time t_j exceed a threshold number of standard deviations above the estimated stochastic volatility, they are considered jumps R^D of the price process; otherwise they are continuous components R^D. Accordingly, a factor model holds for each R, R^C, and R^D. Factor changes and loadings are estimated following Pelger (2019), who essentially applies PCA to a volatility and jump covariance matrix considering only factors whose perturbated eigenvalue-ratio statistics cluster far away from 1 to identify their number K. Continuous and jump components add up to intraday returns.

The dataset used consists of 5-min and daily returns together with characteristics for individual stocks. Overnight returns are the difference between intraday and daily returns. The method identifies four high-frequency and continuous factors for most of the sample, whereas jump factors play a negligible role. Factors are given economic meaning following Pelger and Xiong (2021). The first factor is a market portfolio, while the others represent industry factors, that is, oil, finance, and electricity, and they have low generalized correlation (Bai & Ng, 2006) with Fama–French–Carhart factors. This is the only study we review where industries play such an important role: high-frequency returns may load on different risk sources than those at other frequencies. Overnight and daily data show indeed weaker links to industry portfolios. The major finding is a reversal pattern between intraday and overnight returns. Statistical factors earn a much higher SR intraday than overnight, when it often turns negative. Exactly the opposite happens for Fama–French–Carhart factors. A potential explanation comes from Lou et al. (2019), who observe that large trades from institutional investors tend to occur near the close, while small trades on part of individuals usually take place near the open. In few words: firm characteristics matter overnight and industries matter intraday, such that studying only the cross-section of daily returns might neglect risk-return trade-offs with overnight reversals. Statistical factors capture in fact variation in industry portfolios better than observable ones, which miss this phenomenon.

3.3 | Regression trees/random forest

Interactions among firm characteristics may have predictive power for stock returns (Bryzgalova et al., 2021), but traditional methods struggle to account for all of them in the factor zoo. *Regression trees* are powerful tools to capture interactions and efficiently group stocks into homogeneous groups similar to portfolios. Intuitively, they are nonparametric techniques that partition the feature space into H rectangular regions called *leaves* according to some distance metric, and then fit a simple model in each of them to predict the outcome variable. We can interpret a regression tree as a set of nearest neighbors (Athey & Imbens, 2019) approximating the function of interest $g(z_{i,t}; \theta)$ with a multidimensional step function (Breiman et al., 2017). Figure 1 illustrates a tree partitioning the predictor space into three regions based on size and value according to the final nodes. We can express the prediction for $r_{i,t+1}$ in a tree with L subsequent vertical nodes (or *depth*) and H leaves as

$$r_{i,t+1} = g(z_{i,t}; \theta, H, L) = \sum_{h=1}^{H} \theta_h \mathbf{1}_{\{z_{i,t} \in C_h(L)\}} \tag{15}$$

where $C_h(L)$ is the region at depth L of the tree and θ_h is a constant parameter coinciding with the average return of the node. Since each stock i falls in just one of the final nodes, its predicted return is the node average.

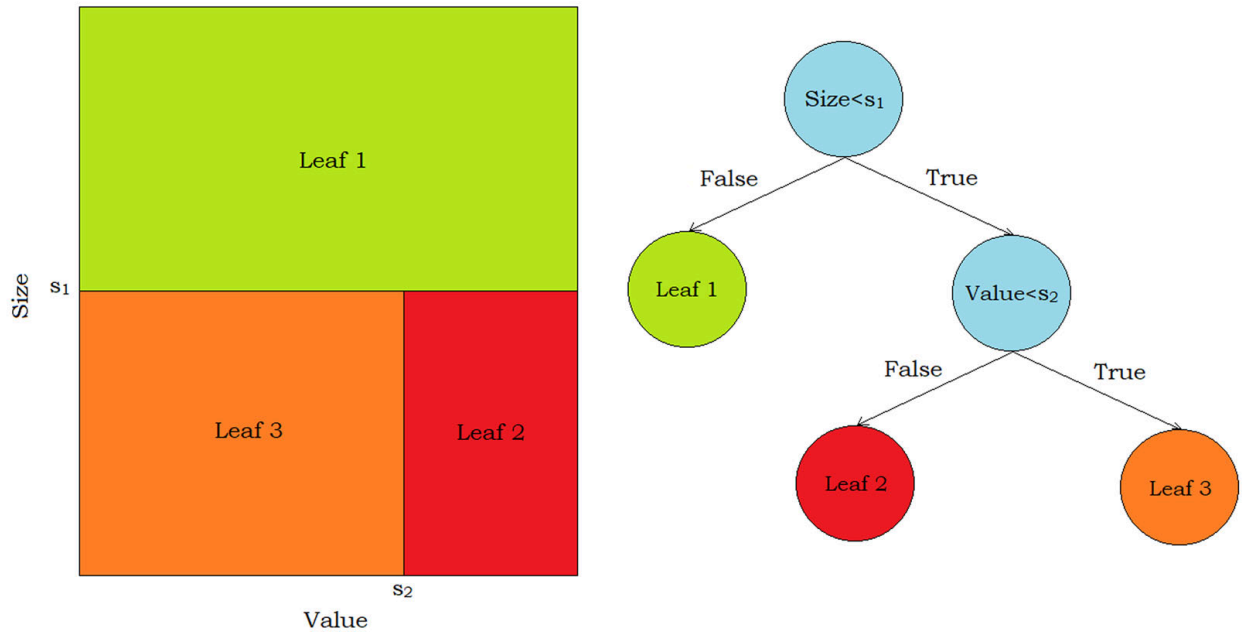

FIGURE 1 Example of regression tree diagram using size and value with $L = 2$ and regressor space partition into leaves according to splitting points s_1 and s_2 [Colour figure can be viewed at wileyonlinelibrary.com]

How do we determine the partitions to "grow" a tree? As in other ML methods, the decision metric is the forecast error. Since finding the globally optimal partition minimizing it is often computationally infeasible (Nagel, 2021), we need to resort to an algorithm to approximate the solution. The *greedy* algorithm (Hastie et al., 2009) is a local method focusing on a single branch every time disregarding the rest. For each splitting variable $j = 1, \dots, P$, a splitting point s creates two half-planes of the covariates Z, that is, $\{Z|z_j \leq s\}$ and $\{Z|z_j > s\}$. We seek both j and s to minimize the local measure

$$\min_{j,s}\left[\min_{c_1}\sum_{z_i\in\{Z|z_j\leq s\}}(r_{i,t+1}-c_1)^2+\min_{c_2}\sum_{z_i\in\{Z|z_j>s\}}(r_{i,t+1}-c_2)^2\right] \tag{16}$$

One can show that the constants c_1 and c_2 are the average return in each subregion. Hence, locating the best splitting point for each variable is quick. We keep partitioning the feature space into half-planes every time we find the best split until we get H leaves. This process is known as *recursive binary splitting*. The greedy algorithm is efficient, but it does not ensure that the resulting "path" of thresholds and splits is the correct one (Gu et al., 2020).

Regression trees are invariant to monotonic transformations of the data and capture up to $(L-1)$ interactions with a depth L. Depth and leaves are tuning parameter governing the model complexity and they should be determined with techniques reducing overfitting. *Cost-complexity pruning* consists in growing a large tree with H leaves first, and then in "pruning" it back to a smaller one by collapsing some leaves into bigger ones to minimize a Ridge- or LASSO-penalized SSR over the terminal nodes. This creates a sequence of smaller trees and the one with the smallest cost becomes the final tree. Regularization can also be done through *ensemble* methods, which simply denote the practice of aggregating results of different hyperparameter choices through some form of average. *Boosting* recursively merges forecasts from oversimplified ("shallow") trees into a single "stronger" tree according to some shrinkage factor. This leads to *gradient-boosted regression trees* (*GBRT*, Gu et al. (2020)). *Bagging* or *bootstrap aggregation* averages predictions

over several bootstrapped samples (Hastie et al., 2009). *RF* (Breiman, 2001) modify bagging combining large collections of decorrelated trees. One first bootstraps B samples of the same size of the data. Then, for each sample, each node in the tree is split considering only a random subset $\upsilon < P$ of the characteristics, and the selection process is repeated until obtaining H leaves. The RF predictor is the average of the B predictions. This method reduces correlation among trees in different bootstrap samples to avoid problems arising from covariates with disproportionate effects compared to the rest. CV can be used to choose υ and B.

3.3.1 | Regression trees/random forest in asset pricing

Moritz and Zimmermann (2016) propose *tree-based conditional portfolio sorts* to sidestep the limitations of traditional linear models and account for interactions among many predictors. The method follows the greedy algorithm explained above to predict a company's stock returns using decile rankings based on its 1-month returns over the past 2 years. To reduce overfitting and decorrelate results, trees are aggregated into RF with $\upsilon = 8$ random past-return decile rankings in each tree. A prediction-based long-short portfolio delivers anomalous returns with respect to Fama & French (1993)'s model plus momentum and does not significantly load on size and value. The study provides two insightful approaches to interpret ensemble methods like RF. The first one assesses variable importance mechanically perturbating the value of a candidate predictor in all the trees and then measuring the resulting mean-squared error relative to the original one across trees. From this perspective, recent returns (STR) contain more predictive power than more distant ones (intermediate and standard momentum) because they generate larger reductions in the forecasting error. The second method builds "average partial derivatives" of monthly returns with respect to the decile ranking of one past month looking at average differences over time, bootstrapped samples and firms for predictions obtained artificially setting its value to 1,2,…,10, *ceteribus paribus*. Using this procedure for two-way interaction terms shows that the interplay between short- and longer-term past-returns significantly affects future returns. A limitation of the study is the use of decile rankings in the RF instead of continuous variables, which means they depend on univariate sorts with all their shortcomings after all.

Bryzgalova et al. (2021) prune RF based on penalized mean–variance optimization to build test assets representing higher hurdles for Asset Pricing models. A well-fitting model for test assets, which do not span the underlying SDF, could easily miss relevant risk factors. Hence, constructing meaningful cross-sections of returns is crucial.

Conventional portfolio sorts suffer from several problems, such as *curse of dimensionality* and noisy repackaging of the same economic risks.[5] The former affects standard regression trees, too: with M sorting variables, we have $M^L \cdot 2^L$ (overlapping) portfolios with $N/2^L$ stocks each. The contribution of the paper consists in providing sound economic intuition to prune trees back to a feasible number of leaves spanning the SDF as closely as possible. Greedy algorithms are ill-suited to maximize the OOS SR because one must consider the complete asset covariance matrix to find the tangency portfolio weights. *Asset Pricing Trees (AP-Trees)* achieve pruning by employing instead a global criterion based on the typical mean–variance optimization that seeks SDF weights b, augmented with an elastic net:

$$\min_b \frac{1}{2} b' \bar{\Sigma}_{AP} b + \lambda_1 |b|' \mathbf{1} + \frac{1}{2} \lambda_2 b' b$$
$$\text{subject to } b' \mathbf{1} = 1;\ b' \bar{\mu}_{AP} \geq \mu_0 \tag{17}$$

where $\bar{\mu}_{AP}$ and $\bar{\Sigma}_{AP}$ denote AP-Trees portfolios sample moments, $\mathbf{1}$ is an $N \times 1$ vector of ones and $\mu_0, \lambda_1, \lambda_2$ are tuning parameters to choose by CV maximizing the SDF SR. Portfolio return weights damp the role of "deeper" nodes to compensate for their higher diversification. The robust mean–variance optimization in Equation (17) intuitively shrinks the contribution of assets that do not contribute to span the PK collapsing high-level nodes (i.e., portfolios) into bigger ones, thereby reducing overfitting, and generalizes the SDF recovery of Kozak et al. (2020). A caveat of AP-Trees is that there is no guarantee that the resulting portfolios span the whole mean–variance space covered by individual stocks.

The SDF resulting from Equation (17) is used to compare standard 32 and 64 triple-sorted portfolios based on size, investment, and profitability against 40 AP-Trees. The latter deliver much higher SRs and traditional Fama–French models capture only about 60% of their variation versus an R^2 above 80% for the former. The study finds remarkable redundancy in many test assets. By choosing λ_1 appropriately, the authors show that one retains roughly 90% of the SR of an SDF built using only 10 portfolios instead of 40 for many sets of AP-Trees. Another relevant result is that interactions matter. This is cleverly shown removing nodes containing interactions among characteristics. The resulting PK SR is only half as large as the original one in many cases.

3.4 | Neural networks

Neural networks find parameterized mappings of inputs into outputs to approximate highly nonlinear functions of the data in flexible ways. We describe here one of the simplest types, the *feedforward network (FFN)*, henceforth just NN unless otherwise specified. Its *architecture* is defined by its input, hidden, and output layers, as illustrated in Figure 2. In terms of Equation (2), inputs are the characteristics z and outputs are the returns r. The hidden layers interact and nonlinearly transform the inputs estimating the unknown function $g(z;\theta)$ to obtain accurate predictions. Each layer is composed of *nodes* (*units*, or *neurons*). While the dimension of the inputs determines the nodes in the input layer, the number of nodes Q in each hidden layer as well as the number of hidden layers L are hyperparameters.

To train a NN, one starts from a semi-affine transformation of the input vector $z^{(0)}$ according to some parameter matrix $\theta^{(0)}$. Dropping indices i and t for simplicity, $z^{(1)} = \theta_0^{(0)} + \sum_{j=1}^{P} z_j^{(0)}\theta_j^{(0)} = z^{(0)\prime}\theta^{(0)}$. The "bias term" $\theta_0^{(0)}$ plays the role of intercept in a linear model. Then, at each unit of the first hidden layer, a nonlinear *activation function* is applied element-wise to $z^{(1)}$ obtaining $z^{(2)} = f(z^{(1)};\theta^{(1)})$. Popular choices for f are the sigmoid and the rectified linear unit (ReLU) function, which is $f = max(0, z)$. $z^{(2)}$ becomes the input of the next hidden layer, and the process repeats for each hidden layer. The return forecast is the linear combination of the outputs of each node: $r = \theta_0^{(2)} + \sum_{q=1}^{Q} z_j^{(2)}\theta_j^{(2)}$ if $L = 1$ as in Figure 2. A *deep* NN (when $L \geq 2$) can be expressed recursively with its intermediate outputs $z_q^{(l)}$ for neuron $q = 1, \ldots, Q$ in each hidden layer $l = 1, \ldots, L$, and the final output r:

$$\begin{aligned} z_q^{(l)} &= f(z^{(l-1)\prime}\theta_q^{(l-1)}) \\ r &= g(z;\theta) = z^{(L-1)\prime}\theta^{(L-1)} \end{aligned} \tag{18}$$

which shows that predictions are nothing but linear combinations of a composition of simpler nonlinear functions. The complete parameter matrix θ (*network weights*) is estimated minimizing

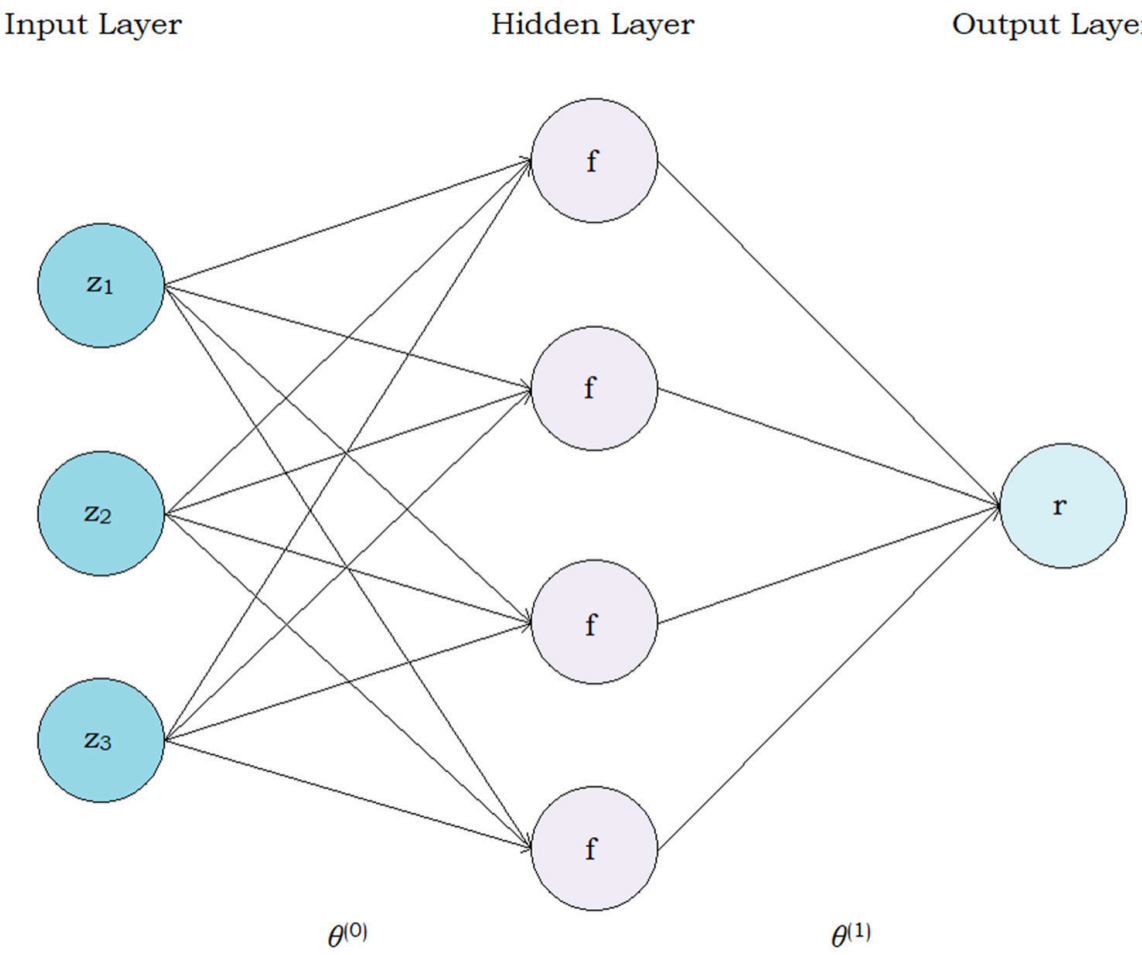

FIGURE 2 Example of FFN with three inputs (z_1, z_2, z_3), one hidden layer with $Q = 4$ units and activation function f, and output r. Arrows are associated with network weight parameters. [Colour figure can be viewed at wileyonlinelibrary.com]

the SSR, just like with OLS: these ones are indeed equivalent to a *shallow* NN ($L = 0$) with linear activations. In practice, there are several algorithms to do this. The popular *stochastic gradient descent* evaluates the gradient of the objective function on small random subsets of data instead of using the entire sample at every iteration, enabling tremendous reductions of computation times. *Back propagation* iteratively updates the gradient at each step controlling for the update size with a hyperparameter called *learning rate*.

NNs require estimating many parameters: $(1 + P)(Q)^L + (1 + P)$ if Q is the same in every hidden layer. Hence, one needs to restrict to simple architectures if the data available are not too large. But then why are NNs so popular? Because, under some conditions, they can approximate any function. According to the *universal approximation theorem* (Hornik et al., 1989), with sufficiently many hidden units, every continuous function over $\mathbb{R}^P$ can be approximated arbitrarily closely by NNs with as few as one hidden layer, for any input dimension P and any monotonically increasing activation function. But not all that glitters is gold: the theorem has important limitations (Dixon et al., 2020), and NNs face other difficulties. Their great flexibility is coupled with high risks of overfitting. Besides Ridge-style penalizations, which shrink the network weights in regions where $g(.)$ is roughly linear, NNs are regularized using small learning rates, with *early stopping* and *batch normalization* (Gu et al., 2020). Overfitting is also curtailed with *ensemble learning*, which reduces both the effect of randomness and the output sensitivity to the initial network weights, and with *dropout*, which averages out the output of multiple networks randomly removing some inputs.

There exists several other types of NNs. For example, *recurrent neural networks (RNNs)* extract hidden states from time-series data with nonlinear dependencies. They are basically flexible $AR(p)$ processes (Dixon et al., 2020). Unknown and long autocorrelation in the data can be captured adding *long-short-term-memory (LSTM)* cells. *Autoencoders* are self-supervised networks that approximate the mapping of a target on itself. Dixon et al. (2020) describe further architectures with applications in finance.

3.4.1 | Neural networks in asset pricing

Chen et al. (Forthcoming) estimate the SDF starting from new test assets that are hard to price combining no-arbitrage together with three types of NNs. With individual demeaned returns, Equation (4) becomes $M_{t+1} = 1 - b_t' r_{t+1}$. The authors assume the SDF weights b_i and the corresponding loadings β_i are functions of both macroeconomic (I_t) and firm-specific information ($I_{i,t}$): $b_{t,i} = b(I_t, I_{i,t})$ and $\beta_{i,t} = \beta(I_t, I_{i,t})$. To find them, they solve the method of moments problem $\mathbb{E}[M_{t+1} r_{i,t+1} g(I_t, I_{i,t})] = 0$ implied by the analogous no-arbitrage condition using an *adversarial* approach which, in loose sense, can be summarized by a min-max problem over the model mispricing:

$$\min_b \max_g \frac{1}{N} \sum_{j=1}^{N} \left(\mathbb{E}\left[\left(1 - \sum_{i=1}^{N} b(I_t, I_{i,t}) r_{i,t+1} \right) r_{j,t+1} g(I_t, I_{j,t}) \right] \right)^2 \tag{19}$$

The flexible nonparametric function $g(.)$ is determined through an iterative zero-sum game where first the "adversary" chooses it to form the portfolios that are the hardest to price, and then the asset pricer corrects the model to price such assets choosing b, until all information is accounted for. Operationally, the *generative adversarial network (GAN)* looks for b and g to minimize Equation (19) allocating one NN each. The "SDF network" seeks b_t, given an initial choice for g. First, a RNN with LSTM cells extracts hidden state variables h_t relevant for pricing discarding the cross-sectional redundancies in the macro aggregates I_t. Second, a FFN estimates b_t using h_t and firm-specific characteristics $I_{i,t}$ as inputs with the objective of approximating the sample counterpart of Equation (19). In similar fashion, the "conditional network" looks for g, given b_t from the SDF network: a RNN finds the hidden states h_t^g, which then serve as inputs for the FFN, which finds g this time *maximizing* (19). This step builds the assets producing the biggest mispricing possible, such that the resulting SDF is more robust than one obtained using assets with strong factor structure like Fama–French portfolios. The algorithm iterates through the two NNs until convergence: minimizing the largest loss among all possible conditioning functions, the loss for any g is small. Finally, factor loadings are obtained fitting a FFN to the moments $\mathbb{E}_t[b_t' r_{t+1} r_{i,t+1}] = \beta_{i,t}$. The optimal model employs dropout and ensemble learning to reduce sensitivity to specific hyperparameter choices.

The GAN is implemented using 46 firms-specific characteristics and 178 macro variables as inputs. It achieves a cross-sectional R^2 of at least 90% for each of the 46 corresponding anomaly portfolios and outperforms simpler models excluding nonlinearities or adversarial approach for both Fama–French portfolios and individual stocks. Average excess returns of portfolios sorted on SDF betas line up almost perfectly against their loadings: representing the security market line (SML) is a traditional Asset Pricing exercise, which shows that ML needs not to be a "black box" but can be given clear economic interpretation. To interpret the complex GAN structure,

the authors average the absolute derivative of the model-implied weights b with respect to the variable of interest over time and across assets. Based on this, the most relevant characteristics are trading frictions and past returns, in particular SUV, STR, and momentum, but other traditional accounting variables matter as well. Among the macro variables, median bid–ask spread and federal funds rate stand out. A final insight reveals that the PK weights are approximately linear in single characteristics, but flexible functional forms are necessary to capture interactions among covariates rather than single higher-order terms. The GAN can be combined with IPCA, with encouraging results.

Gu et al. (2021) extend IPCA to allow for nonlinearities between factor loadings and characteristics using autoencoders. Autoencoders are dimension reduction tools, which can be seen as nonlinear network counterparts to PCA, where returns are function of themselves and no additional information is required. First, autoencoders compress the inputs into a lower dimension passing them into fewer nodes in the hidden layers (*encoding*); then, they unpack the intermediate results into the outputs, which have the same dimension as the original inputs (*decoding*). With one hidden layer of K neurons and a linear activation function one gets a linear K-factor model similar to those studied in Bai and Ng (2002) and Bai (2003). Treasuring the insights from Kelly et al. (2019), a *conditional autoencoder (CA)* exploits asset-specific conditional information. The overall model is composed of two networks that estimate latent factors and loadings of the first row in Equation (10) minimizing a LASSO-penalized SSR. The first one is a FFN with L_β hidden layers fitting a flexible function between betas and firm characteristics through the parameters θ_β: $\beta_{i,t-1} = g_\beta(z_{t-1};\theta_\beta)$. The second network is a standard autoencoder with $L = 1$ and linear activation function recovering latent factors from managed portfolios x_t similar to Kelly et al. (2019) to circumvent the problem of unbalanced panels, such that $f_t = g_f(x_t;\theta_f)$. Predicted returns are the product between factors f_t and the stacked loading matrix B_{t-1}. The model reduces to IPCA using a linear activation and one hidden layer in both networks. Overfitting is further reduced through early stopping and ensemble learning.

The CA is trained on the dataset of Gu et al. (2020), used to build 94 managed portfolios. The baseline model described above (CA_0) is compared against PCA, IPCA, three richer formulations, each adding one hidden layer to the beta-network (CA_1,CA_2, and CA_3), and with observable factor models with up to six factors. Although traditional models are outperformed by far, the OOS total and predictive R^2 cannot point at a clear winner: IPCA performs the best in terms of total R^2, while CAs generate higher predictive R^2. The model-implied tangency portfolio earns an annualized OOS SR of 4.58 and 3.72 for CA_1 and IPCA, respectively. The hypothesis that the OOS pricing error is zero cannot be rejected for 37 test portfolios for Fama–French models, but only in eight cases with the best-performing CA (CA_2). The most relevant characteristics, identified through the reduction in total R^2 like in Kelly et al. (2019), belong to past returns (STR and several versions of momentum), liquidity variables (e.g., volume, Amihud, 2002's illiquidity and bid–ask spread) and risk measures (such as market beta and idiosyncratic volatility). A limit of the study is the lack of a clear criterion guiding the choice of a specific NN architecture over another one other than OOS performance. Furthermore, it is unclear in which circumstances and why IPCA performs better despite being a simpler method.

3.5 | Comparative analyses

A substantial push to the diffusion of ML for Asset Pricing comes from Volume 33, Issue 5 (2020) of the *Review of Financial Studies*, which collects several studies offering new methods for the

cross-section of stock returns, including three of those we review. In their introduction to this special issue, Karolyi & Van Nieuwerburgh (2020) identify ML as promising field to answer Cochrane (2011)'s pending questions. Nevertheless, they call for tools to consistently compare different methods among each other and stress the importance of clear economic interpretations, potentially with a view to incorporate them in equilibrium models, "the ambitious goal towards which we are all aiming" (Cochrane, 2011, p.1067). Some comparative analyses have been recently carried out following these guidelines.

Gu et al. (2020) compare a vast number of ML methods to predict returns according to Equation (2) including elastic nets, PCA, PLS, generalized linear models with penalization, GBRT, RF, and NNs. As stressed in Gu et al. (2021), the study focuses purely on forecasts without taking any stand on risk–return relationships. Nonetheless, it has become the benchmark for later papers employing ML, both for practices used (e.g., database, sample split) and results achieved (R^2, variable importance, etc.). For the data considered, NNs perform the best, closely followed by RF and GBRT. The methods generally agree regarding the most influential stock-level predictors, which belong to four categories: price trends, liquidity variables, risk measures, and valuation ratios. At the macro-level, the aggregate book-to-market ratio stands out. The prediction-based trading strategy generating the highest SR is an ensemble of all methods, but it is difficult to find the economic intuition behind it.

In similar vein, Bianchi et al. (2021) predict Treasury bond returns at different maturities with several ML methods. Differently from stock data, here the majority of them perform bad when using only yields as inputs following Campbell and Shiller (1991) and produce a negative OOS R^2. NNs, extreme regression trees and RF, however, still manage to explain around one fourth of the variation of 10-year Treasuries. When forward rates and macro variables are used as predictors, the best-performing method is a "group-ensembled" NN, that is, a collection of networks, one for each group of macroeconomic aggregates, where the interactions across groups are switched off. Based on these findings, the authors argue that it is the nonlinearities and interactions *within* a variable group that matter, not the interactions *across* groups. The same network is able to forecast both level, slope and curvature of interest rates with reasonable precision, and produces countercyclical bond risk premia. Variable importance is assessed through the average gradient of the network. Predictors linked to inflation, money, and credit matter regardless the maturity, while the role of other covariates is term-structure-dependent: some have either a level or a slope effect only, hence averaging bond returns across maturities for research purposes is not a good idea.

Avramov et al. (2022) evaluate the performance of trading strategies built using a three-layer NN from Gu et al. (2020) (NN3), the GAN of Chen et al. (Forthcoming) (CPZ), IPCA, the CA of Gu et al. (2021), and the robust SDF of Kozak et al. (2020) (KNS) after transaction costs and with restrictions on the sample used. Excluding microcaps, financially distressed firms or without credit rating from the training sample results in network-based models losing more than half of their Fama–French risk-adjusted returns, while IPCA offers more robust strategies. Concerns about net-of-fee performance arise due to the large turnovers ML strategies generate, and the rather extreme portfolio positions of the KNS-SDF cast doubts on its feasibility from a real-time investment perspective. On the positive side, the profitability with NN3 and CPZ increases in periods of high market sentiment and volatility and low market liquidity, as expected following previous research. Moreover, it does not decline in recent years, in contrast to traditional anomalies. Looking at the stocks assigned to the same prediction-sorted portfolios reveals that all ML techniques agree in investing long in traditional characteristics carrying positive risk premia, such as size, illiquidity, beta, momentum, and so forth. Furthermore, the authors decompose the payoff of a trading strategy into an intra-industry and an inter-industry component, specifically investing long (short) in the winners (losers) *within* an industry and in the winner (loser) *industries*, respectively. The

intra-industry strategy outperforms the inter-industry one substantially, and accounts for 84% of the original strategy. In other words, ML informs on stock selection rather than industry rotation: controlling for similar firm fundamentals within the same industry helps identifying the most profitable stocks.

4 | LESSONS FOR THE FUTURE

Does ML contribute to our understanding of expected returns? How does theory fit with the recent findings? What potential pitfalls should practitioners and academics be wary of when employing ML in Asset Pricing? What developments should we expect? In what follows, we highlight the most salient results that emerge from the papers we discussed pointing out prospective directions for future research.

4.1 | What did we learn about asset returns?

Despite the recent noticeable efforts to apply new methods in Asset Pricing, a major insight that transpires from this emerging literature is that the factor zoo issue is still unresolved. Every method we review detects indeed different groups of prominent factors for the cross-section and at this stage, it is complicated to tell which approach to prefer, especially considering that several well-known anomalies has become less relevant over time (Chordia et al., 2014) and that the predictive power of some characteristics varies after conditioning on the others (Freyberger et al., 2020). Notwithstanding, some factors tend to stick out more often than others, namely past returns (e.g., STR and momentum), liquidity factors (e.g., Pástor & Stambaugh, 2003), and trading frictions (e.g., SUV). Pinning down a small set of risk sources robust to various identification algorithms ultimately ameliorates the comprehension of what drives returns and provides an excellent starting point for further research. In particular, this information can assist researchers in the question about the sparsity/density of the SDF, which remains open. While majority of the papers find redundancies in the factor zoo (e.g., Feng et al., 2020), others claim that one needs to consider all anomalies to achieve good performance (Kozak et al., 2020). The point is far-reaching as it guides researchers' views in the process of building theoretical models to explain these empirical facts.

A second notable finding is that nonlinearities matter. Interactions among covariates play a big role in several studies, and matter more than nonlinearities in single characteristics (Bianchi et al., 2021), although not everybody agrees (Kozak et al., 2020).

Finally, new patterns have been discovered in specific applications thanks to ML. For example, high-frequency stock returns are largely driven by industry factors in contrast to traditional characteristics-based factors (Pelger, 2020). Returns of assets traditionally difficult to describe with factor models, like options, are easily captured by few economically interpretable factors (Giglio et al, 2022). Firm characteristics and first moments of returns contain valuable information to build low-dimensional but highly efficient statistical factor representations (Kelly et al., 2019; Lettau & Pelger, 2020b).

4.2 | Going forward

ML offers great prediction accuracy (Gu et al., 2020), with trading strategies that deliver hedge during crises, mitigate downside risk, and remain profitable in recent years (Avramov et al., 2022).

Additionally, some methods allow to side-step well-known problems of the field (Feng et al., 2020; Giglio and Xiu, 2021). However, there are also limitations we must bear in mind moving forward.

Probably the biggest downside of ML methods is their lack of interpretability. This partially occurs because they are mainly designed for predictions, where performance is often all that matters. Ensemble learning tends to outmatch other approaches (Bianchi et al., 2021), but it is difficult to attaching economic meaning to the results found. We need bridges between economic theory and ML: the latter is a tool to support the former, not a substitute for it. Tailoring ML to Asset Pricing purposes instead of using methods "off the shelf" has revealed the most effective practice, both for prediction purposes (Freyberger et al., 2020), mean–variance optimization (Bryzgalova et al., 2021), and for factor models (Lettau & Pelger, 2020b). Clear economic restrictions enable the interpretation of the most complex approaches (e.g., Chen et al., Forthcoming).

Adding economic structure to describe some aspects of the data before resorting to computational algorithms enhances intuition and indirectly mitigates the issues due to low signal-to-noise ratios in financial markets. Furthermore, it helps rationalizing the broad and diverse range of findings that recent papers brought about, which are often in contrast with each other. The patterns captured might not be robust features of the data, a situation in which embedding economic theory would be extremely valuable. Alternatively, the methods presented might still be sensitive to changes in the data despite all the regularization introduced. It is, therefore, critical to follow some standards with these methods.

An obvious starting point is using the same data. In line with this perspective, Chen and Zimmermann (2022) provide a large and freely accessible database of anomalies. Of course, this should not encourage *p*-hacking or data mining. Preprocessing choices such as rank-normalization should be standardized to avoid affecting the final results.

Common grounds should also be established with respect to evaluation metrics. Since risk factors can be combined into a PK (Back, 2010), any factor model can be evaluated using its SR analogously to trading strategies, although caution is necessary: some strategies can be tilted to "sharpen" the SR (Goetzmann et al., 2002). Moreover, the tangency portfolio may not be feasible in practice (Avramov et al., 2022) and transaction costs must always be accounted for. An SML-like representation as in Chen et al. (Forthcoming) is a good benchmark because it verifies a crucial Asset Pricing restriction. Likewise, comparing the behavior of SDFs from different methods could help recognize and isolate common components. Once again, attaching economic meaning to methods and evaluation measures becomes decisive.

Finally, better tools to quantify the contribution of inputs to model outcomes are pivotal to eradicate the skepticism that persists about ML. We need to dive deeper in the ML literature as it already offers valid instruments even for very complex methods (Dixon et al., 2020). Alternatively, one can simply perturbate the value of one covariate and see how the outcome varies, as in Moritz and Zimmermann (2016). Further still, the lack of standard fit metrics like *p*-values and the notion of statistical significance in ML should not refrain us from designing new evaluation metrics. For example, bootstrapping remains a powerful instrument to quantify the uncertainty of coefficient estimates also in this context.

Looking ahead, ML for Asset Pricing is a promising and fertile field. Other than employing further techniques and refining those already explored, there is room for research when considering additional asset classes. The stock market has drawn a disproportionate amount of attention, while applications for bonds and derivatives have played a limited role thus far. Apart from this, time is needed to better comprehend the interplay between ML and finance. After all, there is only one true OOS, without real shortcuts for it.

5 | CONCLUSION

ML methods have been recently exploited in empirical Asset Pricing with encouraging results opening up possibilities for further research. The studies we review show that pricing anomalies are not necessarily stable after conditioning for further covariates, such that there is a high degree of redundancy in the factor zoo, but it is not clear whether the SDF is sparse or dense yet. The most relevant variables belong to past returns, liquidity and trading frictions, and nonlinearities (especially interactions) matter for the cross-section. Risk of overfitting the data and difficult interpretation of the procedures employed are the price to pay for the flexibility and the performance of ML methods. Common grounds for data sample, evaluation metrics, and tools to identify the contributions of characteristics to expected returns are vital for future research.

ACKNOWLEDGMENTS

We gratefully acknowledge research support from the Leibniz Institute for Financial Research SAFE. We also thank two anonymous reviewers and the Associate Editor for their comments and suggestions.

ORCID

Matteo Bagnara https://orcid.org/0000-0002-7674-1812

ENDNOTES

[1] The literature often talks about high-dimensional problems referring to the factor zoo, but the number of observations per time series, T, can still be larger than P for each firm, depending on the data frequency. Nonetheless, issues arise even when $P < T$ but close to it (Hastie et al., 2009). We refer to high-dimensional problems in a relative sense to label issues which are *more likely* with many anomalies than in low-dimensional models derived, say, from first principles. Notice that N plays a negligible role here because of the limited incremental information contained in new cross-sectional observations (Gu et al., 2020).

[2] ML-specific reviews do not mention them (e.g., Athey and Imbens (2019)).

[3] Standard CV is not appropriate with sequential data, as excluding independent subsets in the middle of the data breaks the time-series dependence of predictors and returns (Bianchi et al., 2021). The studies surveyed here bypass this issue validating the model on subsamples which are consequential to an expanding training set. Gu et al. (2020) provide an excellent discussion of sample splitting schemes for forecasting purposes.

[4] Data must be standardized before applying PCA to avoid that arbitrary scaling choices affect the results: we can correctly assess their contribution to the total variance only if they are all on the same "unit." Accordingly, the studies employing dimension reduction we review apply standardization as initial step.

[5] To address (Lewellen et al., 2010)'s concerns, many studies stack together several traditional cross-sections, for example, 25 size-value-sorted portfolios together with industry and momentum portfolios. Such assets are known to have a strong factor structure, which means they are well explained by few factors. Pooling them together might artificially inflate the model performance giving an illusion of robustness to many test assets.

REFERENCES

Amihud, Y. (2002). Illiquidity and stock returns: Cross-section and time-series effects. *Journal of Financial Markets*, *5*(1), 31–56.

Asness, C., Frazzini, A., Israel, R., & Moskowitz, T. (2014). Fact, fiction, and momentum investing. *The Journal of Portfolio Management*, *40*(5), 75–92.

Asness, C. S., Frazzini, A., & Pedersen, L. H. (2019). Quality minus junk. *Review of Accounting Studies*, *24*(1), 34–112.

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*, 685–725.

Avramov, D., Cheng, S., & Metzker, L. (2022). Machine learning versus economic restrictions: Evidence from stock return predictability.*Management Science*, *00*, 1–33.
Back, K. (2010). *Asset pricing and portfolio choice theory*. Oxford University Press.
Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, *71*(1), 135–171.
Bai, J., Bali, T. G., & Wen, Q. (2019). Common risk factors in the cross-section of corporate bond returns. *Journal of Financial Economics*, *131*(3), 619–642.
Bai, J., & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, *70*(1), 191–221.
Bai, J., & Ng, S. (2006). Evaluating latent and observed factors in macroeconomics and finance. *Journal of Econometrics*, *131*(1-2), 507–537.
Bansal, R., & Yaron, A. (2004). Risks for the long run: A potential resolution of asset pricing puzzles. *The Journal of Finance*, *59*(4), 1481–1509.
Banz, R. W. (1981). The relationship between return and market value of common stocks. *Journal of Financial Economics*, *9*(1), 3–18.
Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, *81*(2), 608–650.
Bianchi, D., Büchner, M., & Tamoni, A. (2021). Bond risk premiums with machine learning. *The Review of Financial Studies*, *34*(2), 1046–1089.
Bishop, C. M. (1999). Bayesian PCA. In *Advances in Neural Information Processing Systems*, 382–388.
Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.
Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). *Classification and regression trees*. Routledge.
Bryzgalova, S., Huang, J., & Julliard, C. Bayesian solutions for the factor zoo: We just ran two quadrillion models. *Journal of Finance,* Forthcoming.
Bryzgalova, S., Pelger, M., & Zhu, J. (2021). Forest through the trees: Building cross-sections of stock returns. Tech. rep., London Business School and Standford University.
Büchner, M., & Kelly, B. (2022). A factor model for option returns. *Journal of Financial Economics*, *143*(3), 1140–1161.
Campbell, J. Y., & Cochrane, J. H. (1999). By force of habit: A consumption-based explanation of aggregate stock market behavior. *Journal of Political Economy*, *107*(2), 205–251.
Campbell, J. Y., & Shiller, R. J. (1991). Yield spreads and interest rate movements: A bird's eye view. *The Review of Economic Studies*, *58*(3), 495–514.
Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of Finance*, *52*(1), 57–82.
Chamberlain, G., & Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, *51*, 1281–1304.
Chen, A. Y., & Zimmermann, T. (2022). Open source cross-sectional asset pricing. *Critical Finance Review*, *27*(2), 207–264.
Chen, L., Pelger, M., & Zhu, J. Deep learning in asset pricing. *Management Science,* Forthcoming.
Chinco, A., Clark-Joseph, A. D., & Ye, M. (2019). Sparse signals in the cross-section of returns. *The Journal of Finance*, *74*(1), 449–492.
Chinco, A., Neuhierl, A., & Weber, M. (2021). Estimating the anomaly base rate. *Journal of Financial Economics*, *140*(1), 101–126.
Chordia, T., Subrahmanyam, A., & Tong, Q. (2014). Have capital market anomalies attenuated in the recent era of high liquidity and trading activity? *Journal of Accounting and Economics*, *58*(1), 41–58.
Cochrane, J. H. (2009). *Asset pricing: Revised edition*. Princeton University Press.
Cochrane, J. H. (2011). Presidential address: Discount rates. *The Journal of Finance*, *66*(4), 1047–1108.
Connor, G., & Korajczyk, R. A. (1986). Performance measurement with the arbitrage pricing theory: A new framework for analysis. *Journal of Financial Economics*, *15*(3), 373–394.
Connor, G., & Korajczyk, R. A. (1988). Risk and return in an equilibrium apt: Application of a new test methodology. *Journal of Financial Economics*, *21*(2), 255–289.
Connor, G., & Korajczyk, R. A. (1993). A test for the number of factors in an approximate factor model. *The Journal of Finance*, *48*(4), 1263–1291.
Coval, J. D., & Shumway, T. (2001). Expected option returns. *The Journal of Finance*, *56*(3), 983–1009.
Dixon, M. F., Halperin, I., & Bilokon, P. (2020). *Machine learning in finance*. Springer.

Fama, E., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, *33*(1), 3–56.

Fama, E. F. (1996). Multifactor portfolio efficiency and multifactor asset pricing. *Journal of financial and quantitative analysis*, *31*(4), 441–465.

Fama, E. F., & French, K. R. (1992). The cross-section of expected stock returns. *The Journal of Finance*, *47*(2), 427–465.

Fama, E. F., & French, K. R. (1996). Multifactor explanations of asset pricing anomalies. *The Journal of Finance*, *51*(1), 55–84.

Fama, E. F., & French, K. R. (2008). Dissecting anomalies. *The Journal of Finance*, *63*(4), 1653–1678.

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, *116*(1), 1–22.

Fama, E. F., & MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, *81*(3), 607–636.

Feng, G., Giglio, S., & Xiu, D. (2020). Taming the factor zoo: A test of new factors. *The Journal of Finance*, *75*(3), 1327–1370.

Ferson, W. E., & Harvey, C. R. (1991). The variation of economic risk premiums. *Journal of Political Economy*, *99*(2), 385–415.

Frazzini, A., & Pedersen, L. H. (2014). Betting against beta. *Journal of Financial Economics*, *111*(1), 1–25.

Freyberger, J., Neuhierl, A., & Weber, M. (2020). Dissecting characteristics nonparametrically. *The Review of Financial Studies*, *33*(5), 2326–2377.

Giacomini, R., & White, H. (2006). Tests of conditional predictive ability. *Econometrica*, *74*(6), 1545–1578.

Giglio, S., Kelly, B. T., & Xiu, D. (2022). Factor models, machine learning, and asset pricing. *Annual Review of Financial Economics*, *14*(1).

Giglio, S., & Xiu, D. (2021). Asset pricing with omitted factors. *Journal of Political Economy*, *129*(7), 1947–1990.

Goetzmann, W. N., Ingersoll, J., Spiegel, M., & Welch, I. (2002). *Sharpening Sharpe ratios*. NBER Working paper no. w9116.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223–2273.

Gu, S., Kelly, B., & Xiu, D. (2021). Autoencoder asset pricing models. *Journal of Econometrics*, *222*(1), 429–450.

Hansen, L. P., & Jagannathan, R. (1991). Implications of security market data for models of dynamic economies. *Journal of Political Economy*, *99*(2), 225–262.

Harvey, C. R., Liu, Y., & Zhu, H. (2016). … and the cross-section of expected returns. *The Review of Financial Studies*, *29*(1), 5–68.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

He, Z., Kelly, B. T., & Manela, A. (2017). Intermediary asset pricing: New evidence from many asset classes. *Journal of Financial Economics*, *126*(1), 1–35.

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

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366.

Hou, K., Xue, C., & Zhang, L. (2015). Digesting anomalies: An investment approach. *The Review of Financial Studies*, *28*(3), 650–705.

Huang, J., Horowitz, J. L., & Wei, F. (2010). Variable selection in nonparametric additive models. *Annals of Statistics*, *38*(4), 2282–2313.

Israel, R., Kelly, B. T., & Moskowitz, T. J. (2020). Can machines "learn" finance? *Journal of Investment Management*, *18*(2), 23–36.

Karolyi, G. A., & Van Nieuwerburgh, S. (2020). New methods for the cross-section of returns. *The Review of Financial Studies*, *33*(5), 1879–1890.

Kelly, B. T., Moskowitz, T. J., & Pruitt, S. (2021). Understanding momentum and reversal. *Journal of Financial Economics*, *140*(3), 726–743.

Kelly, B. T., Palhares, D., & Pruitt, S. Modeling corporate bond returns. *The Journal of Finance*, Forthcoming.

Kelly, B. T., Pruitt, S., & Su, Y. (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, *134*(3), 501–524.

Kozak, S., Nagel, S., & Santosh, S. (2018). Interpreting factor models. *The Journal of Finance*, *73*(3), 1183–1223.
Kozak, S., Nagel, S., & Santosh, S. (2020). Shrinking the cross-section. *Journal of Financial Economics*, *135*(2), 271–292.
Lettau, M., & Pelger, M. (2020a). Estimating latent asset-pricing factors. *Journal of Econometrics*, *218*(1), 1–31.
Lettau, M., & Pelger, M. (2020b). Factors that fit the time series and cross-section of stock returns. *The Review of Financial Studies*, *33*(5), 2274–2325.
Lewellen, J., Nagel, S., & Shanken, J. (2010). A skeptical appraisal of asset pricing tests. *Journal of Financial Economics*, *96*(2), 175–194.
Lin, X., & Zhang, L. (2013). The investment manifesto. *Journal of Monetary Economics*, *60*(3), 351–366.
Lou, D., Polk, C., & Skouras, S. (2019). A tug of war: Overnight versus intraday expected returns. *Journal of Financial Economics*, *134*(1), 192–213.
Martin, I. W., & Nagel, S. (2022). Market efficiency in the age of big data. *Journal of Financial Economics*, *145*, 154–177.
McLean, R. D., & Pontiff, J. (2016). Does academic research destroy stock return predictability? *The Journal of Finance*, *71*(1), 5–32.
Mehra, R., & Prescott, E. C. (1985). The equity premium: A puzzle. *Journal of Monetary Economics*, *15*(2), 145–161.
Merton, R. C. (1973). An intertemporal capital asset pricing model. *Econometrica: Journal of the Econometric Society*, *41*, 867–887.
Moritz, B., & Zimmermann, T. (2016). Tree-based conditional portfolio sorts: The relation between past and future stock returns. Tech. rep., Ludwig Maximilian Unversity Munich and Federal Reserve Boar.
Nagel, S. (2021). *Machine learning in asset pricing*. Princeton University Press.
Neal, R. M. (2012). *Bayesian learning for neural networks* (Vol. 118). Springer Science & Business Media.
Pástor, L. (2000). Portfolio selection and asset pricing models. *The Journal of Finance*, *55*(1), 179–223.
Pástor, L., & Stambaugh, R. F. (2000). Comparing asset pricing models: An investment perspective. *Journal of Financial Economics*, *56*(3), 335–381.
Pástor, L., & Stambaugh, R. F. (2003). Liquidity risk and expected stock returns. *Journal of Political Economy*, *111*(3), 642–685.
Pelger, M. (2019). Large-dimensional factor modeling based on high-frequency observations. *Journal of Econometrics*, *208*(1), 23–42.
Pelger, M. (2020). Understanding systematic risk: A high-frequency approach. *The Journal of Finance*, *75*(4), 2179–2220.
Pelger, M., & Xiong, R. (2021). Interpretable sparse proximate factors for large dimensions. *Journal of Business & Economic Statistics*, *00*, 1–23.
Rosenberg, B., Reid, K., & Lanstein, R. (1985). Persuasive evidence of market inefficiency. *The Journal of Portfolio Management*, *11*(3), 9–16.
Ross, S. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, *13*(3), 341–360.
Stock, J. H., & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, *97*(460), 1167–1179.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.
Vapnik, V. N. (1998). *The nature of statistical learning theory* (Vol. 1). Springer.
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.

How to cite this article: Bagnara, M. (2024). Asset Pricing and Machine Learning: A critical review. *Journal of Economic Surveys*, *38,* 27–56. https://doi.org/10.1111/joes.12532