CONTENTS

Chapter 1

INTRODUCTION

Prediction problems are central to asset pricing. To price stocks, investors must forecast firms' future cash flows. Investors seeking outperforming trading strategies search for signals that predict asset returns. Researchers testing asset pricing models look for predictor variables that can forecast return differences between assets or that capture forecastable variation in returns across time. Models of credit risk require predictors of default. Hedging and risk management models require forecasts of asset return comovement.

The number of predictor variables that are potentially relevant in these applications is enormous. Technological advances have led to an explosive growth in the amount of information that is available to investors and analysts. Even if we look just at the narrow slice of data that can be extracted from corporate financial reports, the growth in data availability has been staggering. Figure 1.1 provides some rough estimates. One hundred years ago, printed annual volumes like the *Moody's* manuals that summarized corporate financial reports represented much of what was readily available to the public. With the advent of electronic computing, databases like COMPUSTAT expanded coverage to perhaps hundreds or thousands of variables per firm. Today, there is an almost uncountable number of variables that one can construct from publicly available information. The SEC's Edgar database contains financial report data on the order of magnitude of terabytes. With textual analysis, one could probably construct a million variables for each firm from these files.

Corporate financial reports represent only a small fraction of what is potentially available to investors. Databases that record the past history of market prices and transactions contain a gigantic volume of data; sentiment measures can be extracted from social media; online reviews by customers and employees may contain valuable information; and many other data sources could be relevant.

This abundance of potential predictor variables gives rise to a statistical problem. As an example, consider the case of cross-sectional stock return prediction. Say there are $N = 5000$ stocks for which we can observe returns. The number of return predictors, J, that we might consider for forecasting differences in stocks' returns could easily exceed the number of stocks. Is it possible to estimate the relationship between so many

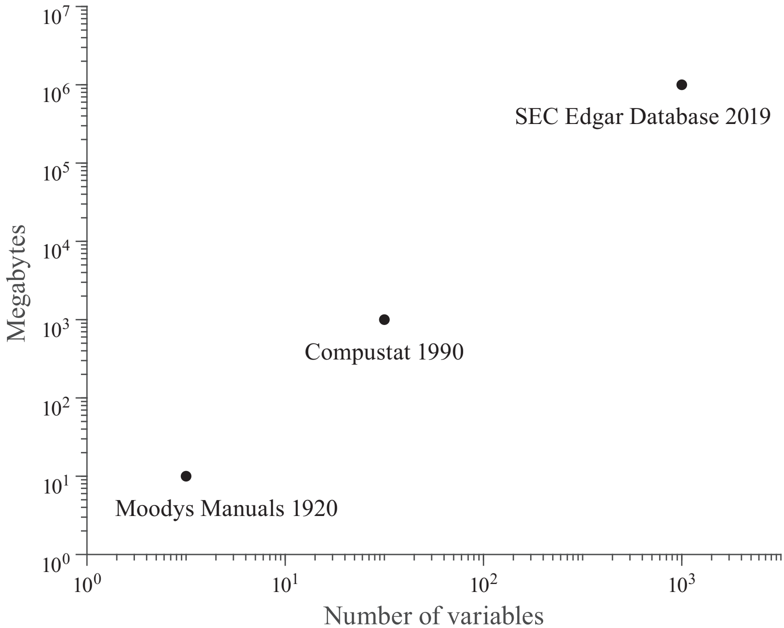

Figure 1.1. Corporate financial reports: 'Big data'

predictors and future returns in a way that delivers useful forecasts of returns?

Conventional statistical techniques like ordinary least squares regression (OLS) are not designed for such high-dimensional settings where J is big relative to N. When $J > N$, OLS regression doesn't have a unique solution. And even if $J < N$, but J is not much smaller than N, the OLS estimator often does not produce useful predictions. With such a high number of explanatory variables, the OLS regression overfits noise. This leads to a good in-sample fit, but poor out-of-sample forecasts.

1.1 Ad hoc Sparsity in Empirical Asset Pricing

Research in asset pricing has, until recently, side-stepped this high-dimensionality problem by focusing on low-dimensional models. Work on cross-sectional stock return prediction, for example, has focused on regressions with a small number of firm characteristics. Collectively, researchers have investigated the predictive power of a large number of firm characteristics, but in any individual study, the number of predictors considered by researchers is typically small. Similarly, researchers looking to summarize the investment opportunities in the cross-section of stock returns with factor models have focused on models with a very small number of factors. For example, Hou, Xue, and Zhang (2015) and Fama and French (2015) include only three or four factors in addition to the value-weighted market portfolio excess return in their factor models. These factors are portfolios constructed based on firm characteristics such as firm size, profitability, investment, or the ratio of the firm's book equity to market equity.

Given the background of an enormously large number of variables that could potentially be relevant for predicting returns and for constructing characteristics-based factor portfolios, focusing on such a small number of factors effectively means that the researchers are imposing a very high degree of *sparsity* on these models. Among the hundreds, thousands, or more potential factors, researchers have chosen a specification that sets the effect of almost all of them to zero.

Imposing such extreme sparsity on the model ensures that conventional statistical methods are well behaved. But the imposition of sparsity is ad hoc. The researchers proposing these models have tested their low-dimensional factor models only against a small subset of the universe of factors that one could potentially construct based on firm-level variables. So we do not really know how much these models miss, in terms of predictive power, relative to the joint effect of this large number of omitted factors. In this regard, it is interesting to note that the number of "standard" factors that researchers view as necessary to adequately capture the cross-section of expected stock returns has been trending up over time. Fama and French (1993) started with three, then came four- and five-factor versions, and Barillas and Shanken (2018) suggest that six are necessary. One interpretation of this expansion in the number of factors is that the literature is slowly adjusting to the fact that there are, indeed, relevant omitted factors.

1.2 Ad hoc Sparsity in Theoretical Asset Pricing

These issues are not only relevant for empirical research in asset pricing, but they also raise questions about theoretical modeling of investor decision making. Asset prices reflect investors' expectations of future asset payoffs. But how do investors come up with these expectations? Real-world investors face the same problem that empirical asset pricing researchers face: there is an enormous number of potentially relevant predictor variables. Distilling them into a good forecasting model is a high-dimensional problem that conventional statistical methods are not well suited for.

Most theoretical asset pricing models assume rational expectations. This assumption is much stronger than just rationality of expectations. In these models, investors do not have to estimate the forecasting model—they already know it. More precisely, investors know perfectly the functional relationship between any relevant predictor variables and the variables they would like to forecast. Given the values of the predictor variables, investors are assumed to be able to calculate the conditional expectations of the forecasted variables. This assumption is often motivated with the idea that the model is meant to represent an equilibrium that would be reached after investors have had time to learn these functional relationships in a stable environment. Even in a low-dimensional setting, the assumption that the learning process has reached an end is questionable. Indeed, Timmermann (1993), Lewellen and Shanken (2002), Collin-Dufresne, Johannes, and Lochstoer (2017), and Nagel and Xu (2019) have argued that investor learning about parameters of the data-generating process is important for understanding asset prices. In a more realistic high-dimensional setting in which investors have to extract the predictive information from thousands of observable variables, the investors-have-already-learned argument is even less convincing.

Arguably, therefore, we should have theoretical models in which investors struggle with high dimensionality in the same way as econometricians do when they study asset price data. Existing models in which investors learn about forecasting models and their parameters typically assume that investors condition their forecasts on a small number of predictors. This sparsity is imposed ad hoc. It seems difficult to make the case that such a sparse representation adequately reflects the forecasting environment faced by real-world investors. Possibly, this mismatch between the difficulty of the prediction problem faced by investors in theoretical models and the difficulty of prediction problems in the real world could be a cause of the empirical shortcomings of existing theoretical asset pricing models.

1.3 Machine Learning

Machine Learning (ML) offers tools to tackle high-dimensional prediction problems. Broadly, ML involves algorithms that allow computers to learn from data. The computer is fed training data to learn, and then the trained algorithm can be used to make predictions. For example, in image recognition, an algorithm could be fed data on image *features* (numerical color values for each image pixel) from a large number of images that are *labeled* into categories. To take an extremely simple case, say we are interested in classifying images of food into ones showing hot dogs and

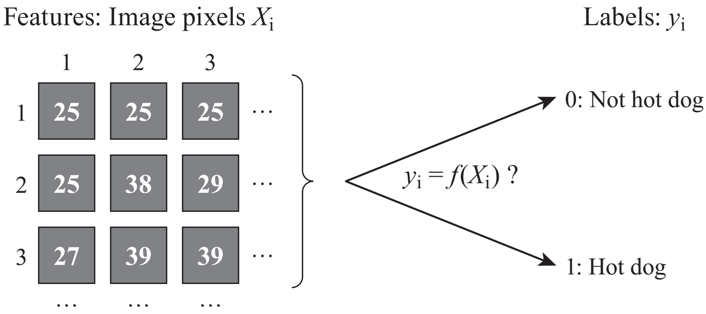

Figure 1.2. Image classification example

ones that do not show hot dogs. From the training data set of already-labeled images, the ML algorithm learns the relationship between image pixels and the classification as a hot dog or not hot dog. Figure 1.2 provides a stylized illustration. Once trained, the algorithm can then be used to predict, for not-yet-classified images, whether they show a hot dog or not a hot dog. In other applications, trained ML algorithms may classify email as spam based on email content, predict tumors based on gene expression data, or interpret sensor data in autonomous driving.

In many of these ML applications, the number of features is extremely large, and often larger than the number of observations that are available to train the algorithm. Conventional statistical tools like ordinary least-squares (OLS) regression would not work in such a setting. Much of the success of ML in practice is due to the development of effective methods to discipline the estimation such that the estimated model (or trained algorithm) produces useful out-of-sample forecasts.

The ML literature therefore offers a rich toolbox to tackle asset pricing prediction problems in high-dimensional settings. Many of these methods are not fundamentally new to the statistics literature, but the ML community has pushed them very successfully into applications. By experimenting heavily and focusing on methods that “work” rather than on understanding the theoretical properties of estimators, the ML literature has assembled an impressive array of methods that have proven to be useful in practical prediction problems. The aim of this book is to survey some of the first steps that asset pricing research has taken to bring these tools into asset pricing, highlight current challenges, and sketch some paths that researchers could take going forward.

The ML toolbox offers the opportunity to analyze asset prices without imposing extreme ad hoc sparsity on prediction problems. In empirical work, ML tools allow an econometrician to take into account the joint effect of a large number of predictor variables. In theoretical

Table 1.1
Terminology in ML and Statistics

ML	Statistics
Training, Learning	Estimation
Learner, Algorithm	Model, Estimator
Features	Covariates, Explanatory Variables, Independent Variables, Predictors
Target, Label, Output	Dependent Variable
Example, Instance	Data Point, Observation

work, investors can be modeled as machine learners in a realistic high-dimensional environment.

A recurrent theme throughout this book is that even though ML methods have been impressively successful in a wide variety of applications, using these tools off the shelf in asset pricing is not necessarily going to work well. The properties of data in asset pricing applications are often substantially different from those in technology, medicine, and other scientific fields. Successful application of ML methods in asset pricing therefore will often require some adaptation. To develop appropriate adaptations, we need to bring in some prior economic knowledge about the environment that generates the data. The idea, sometimes associated with ML, that one could make predictions in an entirely data-driven automated fashion is too good to be true. Much of this book is devoted to the question of how we can use economic reasoning to make ML tools effective in asset pricing.

Bayesian statistics offers a framework to incorporate prior knowledge into statistical estimation. The Bayesian framework therefore allows us to build a bridge between economic theories of asset pricing and ML. At various points throughout the book I draw on Bayesian statistics to give an interpretation of ML methods and to suggest economically motivated adaptations of these methods.

1.4 Terminology

Coming from computer science, the machine learning literature has developed its own terminology. The concepts and methods often overlap with similar ones in the statistics literature, but they are named differently. This can be confusing. Table 1.1 lists some common terms that will appear throughout various parts of this book. I will use the ML and statistics terminology interchangeably.

1.5 Supervised and Unsupervised Learning

ML methods can be broadly classified into two categories. *Supervised* learning basically refers to regression methods. The training data that is used to train the algorithm comes with features $\boldsymbol{x}_i$ and labels y_i. The goal is to find a function $y_i = f(\boldsymbol{x}_i)$ that maps features into labels. These methods are called supervised learning because one can view the training of the algorithm as a learning process supervised by a "teacher." The learner makes predictions based on $\boldsymbol{x}_i$ in the training data. By revealing the correct labels y_i, the teacher tells the learner whether the prediction was correct. This information about correct or incorrect predictions is used by the learner to tweak the estimate of the function. Once the training is completed, this learned function can then be used to predict labels out-of-sample in data sets where we only have features, but not labels. The image classification example we discussed earlier belongs to this supervised learning category, but there is a large number of other methods that belong into this category as well. Chapter 2 reviews some of the most important ones.

In *unsupervised* learning, the data that is used to train the algorithm only has features, not labels. The goal in unsupervised learning is to find a compressed summary of the data that captures its essential properties. One simple example of a method in this category is principal component analysis (PCA). In PCA, a set of variables is approximated with a smaller number of underlying factors that capture a large amount of the common variation among the variables.

Methods in both categories have useful applications in asset pricing. In this book, I focus largely on supervised learning. One of the fundamental problems in asset pricing—both for financial economists studying asset prices and for investment practitioners—is the estimation of expected asset returns conditional on a set of predictor variables. This is, effectively, a supervised learning problem. Similarly, estimation of cash flow forecasting models in asset valuation is a supervised learning problem. Unsupervised learning methods often play a secondary role in asset pricing applications, for example, in an initial dimension-reduction step that summarizes data before it is fed into a supervised learning algorithm. The distinction between supervised and unsupervised learning in asset pricing applications is not as sharp as it may seem, though. As we will see, some supervised learning approaches effectively have built-in dimension-reduction elements that are similar to those in unsupervised learning approaches.

1.6 Limitations of This Book

This book has some limitations that I would like to clarify at the outset. First, this book is primarily a book about asset pricing. I discuss the application of ML techniques in asset pricing, but this book is not the place to look for information on the latest new developments in ML. Moreover, I do not devote much space to computational questions. This is not because computational issues are unimportant. Quite to the contrary. The success of ML in analyzing huge, high-dimensional data sets is founded on many clever computational advances. But these topics are covered well in the ML literature. Conceptual questions about the suitability of ML tools for asset pricing problems have received comparatively less attention. The focus of this book is on closing this gap.

Second, this book is not an exhaustive survey of machine learning methods in asset pricing. There are many exciting new approaches in current working papers and recently published studies, but I will be able to discuss only a small number of these. Rather than attempting to provide a comprehensive review of the literature, my objective is to highlight the opportunities that exist and the generic challenges that arise when we apply ML methods in asset pricing. The array of available ML tools is vast. I hope to provide some useful thoughts on issues that we need to consider when we bring them into asset pricing. One theme that I return to throughout the book is that economic restrictions are important. To reap the full benefit of ML methods in asset pricing, we need to bring in a limited dose of economic reasoning when we pick from the ML toolbox and make specification choices. Off-the-shelf application of ML techniques without thoughtful adaptation to the specific properties of data in asset pricing is unlikely to work well.

Third, within the area of asset pricing, I focus mostly on cross-sectional return prediction applications. There are certainly other areas in asset pricing where ML methods can be useful, too. For example, valuation models require predictions of asset fundamentals, credit risk models require prediction of credit risk realizations, and risk management applications require predictions of codependencies between asset prices. For all of these, ML techniques can be useful for bringing in high-dimensional predictive information and for handling nonlinear relationships. Yet, a short book like this one necessarily has to be selective. The focus on cross-sectional return prediction in this book simply reflects what I have been working on in my own research. I nevertheless hope that by using these specific applications for illustration, I can provide some insights that generalize beyond these specific settings.

Finally, throughout this book, I often highlight open questions rather than providing definite answers. This book therefore does not offer cookbook recipes for applying ML techniques in asset pricing. Instead, by pointing out interesting unresolved issues, I hope to provide some inspiration for future research that addresses these issues. In this spirit, the last chapter summarizes a number of open research questions that are particularly important for further progress in this area of asset pricing.

1.7 Organization of This Book

The rest of this book is organized as follows. Chapter 2 provides a brief overview of a number of basic supervised learning methods. The chapter begins by reviewing regression methods that are designed to predict continuous variables, including ridge regression, lasso, trees and random forests, and neural networks. Several of these learning algorithms involve hyperparameters that need to be set in advance, before the actual estimation. Chapter 2 discusses data-driven methods of tuning these hyperparameters to optimize the predictive performance of the learning algorithm. Hyperparameters often control the degree of regularization imposed on the estimation. Chapter 2 ends with a brief illustration of a Bayesian interpretation of regularization. In this interpretation, regularization corresponds to imposing certain prior distributions on model parameters. We use this Bayesian interpretation in subsequent chapters inject economic reasoning and prior knowledge into the design of ML approaches in asset pricing.

Chapter 3 explores challenges that arise when applying ML methods in asset pricing. The chapter starts by outlining some key differences between the properties of data that most ML algorithms have been developed for and the properties of data that are typical in asset pricing. Throughout the chapter, I illustrate some of these issues with a concrete empirical example of cross-sectional stock return prediction using each stock's own price history as source of predictive information. The chapter highlights that while ML methods are well suited for the prediction problems that arise in asset pricing, these techniques require significant adaptation if they are to deliver on their full potential. What works in typical ML applications, say in the technology sector or biostatistics, does not necessarily work well in asset pricing. Among other things, the low signal-to-noise ratio in asset pricing applications means that attempts to simply let the data speak within an extremely flexible framework are unlikely to yield good results. We will have to impose some structure on the learning algorithm. To do so, we need to connect the ML methods to basic economic frameworks of asset pricing and portfolio choice. Some principles from Bayesian statistics are useful for making this connection.

Chapter 4 describes an approach that makes some progress in this direction. It starts from a basic mean-variance portfolio choice problem—or, equivalently, the problem of finding a stochastic discount factor expressed as a linear combination of asset returns—and the notion that near-arbitrage opportunities are unlikely to exist in the stock market. By formulating Bayesian prior beliefs about the risk-return opportunities in the market, this framework allows us to impose economically motivated constraints on the return prediction problem. The estimator that emerges from this approach is similar to the elastic net estimator common in many ML applications, but with some important differences that come from taking into account the fact that prediction error covariances are an important determinant of a portfolios risk-return profile. The empirical application of this estimator uses a broad set of firm characteristics as return predictors, as well as nonlinear transformations of those, including pairwise interactions between characteristics. The empirical findings suggest that imposing economically motivated prior beliefs is important for obtaining good out-of-sample predictive performance.

Chapter 5 takes a theoretical perspective. The earlier chapters show that statistical analysis of financial market data must address the fact that the environment is high-dimensional and ML methods provide a good toolbox for this purpose. But what about the investors whose investment decisions determine the asset prices that feed into these statistical analyses? If the information environment in financial markets is such that the prediction problems are high-dimensional, investors in theoretical models of financial markets should presumably face this high dimensionality, too. Machine learning methods therefore provide an attractive blueprint for modeling investor belief formation in theoretical models. Chapter 5 pursues this approach. To focus on fundamental issues, we consider a simple environment in which investors must learn from historical data the relationship between stocks' cash flows and a large set of firm characteristics that serve as predictor variables. Investors are Bayesian and shrink their prediction model estimates toward objectively correct prior beliefs. In equilibrium, stocks are priced such that returns are unpredictable out-of-sample. However, returns are strongly predictable in-sample in ex-post statistical analyses of returns. This is due to the fact that an econometrician analyzing data ex post with in-sample analyses has the advantage of hindsight knowledge that investors in real time do not have. In a low-dimensional environment this implicit advantage of the econometrician can be small, but it is large in a high-dimensional environment. In-sample regressions are therefore ill suited for inferring risk premia or the effects of behavioral biases of investors from asset price data.

Chapter 6 concludes the book by outlining a research agenda for future research on ML in asset pricing.

INDEX