Inductive Learning and Theory Testing: Applications in Finance

Inductive Learning and Theory Testing: Applications in Finance

A dissertation presented

by

Tom Zimmermann

to

The Department of Economics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Economics

Harvard University

Cambridge, Massachusetts

May 2015

Dissertation Advisor:
Professor Andrei Shleifer

Author:
Tom Zimmermann

Inductive Learning and Theory Testing: Applications in Finance

Abstract

This thesis explores the opportunities for economic research that arise from importing empirical methods from the field of machine learning.

Chapter 1 applies inductive learning to cross-sectional asset pricing. Researchers have documented over three hundred variables that can explain differences in cross-sectional stock returns. But which ones contain independent information? Chapter 1 develops a framework, deep conditional portfolio sorts, that can be used to answer this question and that is based on ideas from the machine learning literature, tailored to an asset-pricing application. The method is applied to predicting future stock returns based on past stock returns at different horizons, and short-term returns (i.e. the past six months of returns) rather than medium- or long-term returns are recovered as the variables that convey almost all information about future returns.

Chapter 2 argues that machine learning techniques, although focusing on predictions, can be used to test theories. In most theory tests, researchers control for known theories. In contrast, chapter 2 develops a simple model that illustrates how machine learning can be used to conduct an *inductive* test that allows to control for some unknown theories, as long as they are covered in some way by the data. The method is applied to the theory that realization utility and nominal loss aversion lead to the disposition effect (the propensity to sell winners rather than losers). An inductive test finds that short-term price trends and other features of the price history are more important to predict selling decisions than returns relative to purchase price.

Chapter 3 provides another perspective on the disposition effect in the more traditional

spirit of behavioral finance. It assesses the implications of different theories for an investor's probability to sell a stock as a function of the stock's return and then tests those implications empirically. Three different approaches that have been used in the literature are shown to lead to the, at first sight, contradictory findings that the probability to sell a stock is either V-shaped or inverted V-shaped in the stock's return. Since these approaches compute different conditional probabilities, they can be reconciled, however, when the conditioning set is taken into account.

Contents

List of Tables

List of Figures

Acknowledgments

I thank John Campbell, Sendhil Mullainathan and Andrei Shleifer for extensive advice and support, and I thank Harvard University and the Mensa Education & Research Foundation for financial support. I thank my co-authors Jon Kleinberg, Benjamin Moritz, Sendhil Mullainathan, Frank Schilbach and Chenhao Tan. It has been a pleasure to work with them.

I would like to thank everyone who has given my co-authors and me helpful feedback on one or more of the papers included in this dissertation: Nick Barberis, Stefano DellaVigna, Ed Glaeser, Robin Greenwood, Larry Katz, Owen Lamont, Danial Lashkari, Stefan Mittnik, Lasse Pedersen, Daniel Pollmann, Matthew Rabin, Neil Shephard and Jeremy Stein.

Wenn sich, nach über zwanzig Jahren, die formale Ausbildung des Nachwuchses schließlich dem Ende zuneigt, sind sicher die Eltern am glücklichsten. Meinen Eltern danke ich für die stete Unterstützung, auf die ich mich immer verlassen kann.

Mein größter Dank geht an meine Frau Anna, die mit viel Geduld auf mich gewartet hat. Ich freue mich auf unser nächstes gemeinsames Kapitel.

Bold ideas, unjustified anticipations, and speculative thought, are our only means for interpreting nature: our only organon, our only instrument, for grasping her. And we must hazard them to win our prize. Those among us who are unwilling to expose their ideas to the hazard of refutation do not take part in the scientific game.

Popper (1934)

Introduction

In 1997, NASA researchers Michael Cox and David Ellsworth described the problem of big data for the field of computational fluid dynamics. Scientific visualization had become difficult because data sets had become too large to store them in the memory of even the largest graphics workstations. To the best of my knowledge, this marked the first article to use the phrase "big data". Almost twenty years later, the sentiment about big data has shifted remarkably: Big data is widely regarded as an opportunity rather than a problem, for businesses, policy and scientific research.

The market for so-called big data analytics has been estimated to grow to $ 125 billion by the end of 2015,[1] that are split across infrastructure, software and services. Big data sets have been used to predict election outcomes or the spread of diseases, or to track economic activity in real-time. Research that involves big data is conducted in a variety of fields, including economics: For example, economists have successfully analyzed large data sets to estimate teacher value-added effects or the effects of health care expansion in the US.[2]

There is a slightly more subtle way in which big data has started to affect economics: It has opened the door for a constructive dialogue between computer scientists and economists about the empirical methods that are used in either field. At first glance, it appears that the respective methods are quite different: Empirical economics is particularly interested in estimating causal effects or parameters of structural models and inference procedures

[1] http://www.forbes.com/sites/gilpress/2014/12/11/6-predictions-for-the-125-billion-big-data-analytics-market-in-2015/

[2] For a recent overview of applications in economic research, see Einav and Levin (2014).

typically target coefficient estimates. In contrast, the goal of machine learning (the empirical arm of computer science that economics mostly interacts with) is usually to come up with prediction functions, often estimated non-parametrically, and statistical guarantees commonly solely target the accuracy of predictions. The methods of machine learning are philosophically close to the principle of *inductive reasoning*, and the terms "machine learning" and "inductive learning" are often used interchangeably.[3]

On the other hand, there are also similarities between the approaches: In many problems in economics one is in fact interested in making accurate predictions rather than in assessing parameter estimates (e.g. predicting firm defaults or future economic activity). Furthermore, some problems in economics can be turned into prediction problems, even though traditionally they were viewed differently (e.g. instrumental variable estimation or finding an optimal set of controls).

I took this thesis as an opportunity to study the empirical approach of machine learning in more detail and to think about whether it can beneficially be applied to economic questions. The empirical applications in this thesis are in the field of finance but my conclusion is that inductive learning can find fruitful applications in many areas of economics.

Chapter 1 is an attempt to see how far one can push the paradigm of inductive learning in an application to cross-sectional asset pricing. Over the past forty years, researchers have documented over three hundred variables that can explain differences in cross-sectional stock returns. But how many of these variables contain independent information? Standard techniques like portfolio sorts and Fama-MacBeth regressions cannot easily answer this question when the number of candidate variables is large and when cross-terms might be important as well. Chapter 1 develops a framework, deep conditional portfolio sorts, that can be used in this context and that is based on ideas from the machine learning

[3]Incidentally, this is also close to Sherlock Holmes' modus operandi that he describes in Doyle (1892) as follows:

> Let me run over the principal steps. We approached the case, you remember, with an absolutely blank mind, which is always an advantage. We had formed no theories. We were simply there to observe and to draw inferences from our observations. [...]

literature, tailored to an asset-pricing application. The method is purely inductive in the sense that we let an algorithm choose the most important variables from a large set of candidate variables based on past data without imposing any prior knowledge. While this may sound somewhat extreme, I think of it as a useful benchmark: The algorithm effectively exploits all systematic variation that we could hope to explain with theory. The method is then applied to predicting future stock returns based on past stock returns at different horizons, and short-term returns (i.e. the past six months of returns) rather than medium- or long-term returns are recovered as the variables that convey almost all information about future returns.[4] A trading strategy based on these findings has Sharpe and information ratios that are about twice as high as in a Fama-MacBeth framework that accounts for two-way interactions. The second part of the paper goes beyond the prediction exercise and tries to shed light on how the black box model combines these past return variables to come up with superior predictions.

Chapter 2 argues that machine learning techniques, although focusing on predictions, can be used to test theories. In most theory tests, which we call *deductive*, researchers control for the known theories. In contrast, chapter 2 develops a simple model that illustrates how machine learning can be used to conduct an *inductive* test. These tests allow for one to control not just for known theories but some unknown theories, as long as they are covered in some way by the data. The method is then applied to a classic theory in behavioral finance: that realization utility and nominal loss aversion lead to the disposition effect. A deductive test replicates the original finding that investors are more likely to sell winners than losers (relative to purchase price) in their portfolios (the disposition effect). An inductive test, however, finds that other features of the price pattern since purchase are more important to predict selling decisions; short-term trends seem more important, as does the current price relative to the distribution of past prices. This suggests that realization utility and nominal loss aversion are not the most relevant motives of investment choices but that the

[4]While this chapter outlines the method and illustrates it using past return variables, a larger set of predictors (currently one hundred) is considered in ongoing work.

disposition effect could proxy for some other theory.

Chapter 3 provides another perspective on the disposition effect in the more traditional spirit of behavioral finance. It assesses the implications of different theories for an investor's probability to sell a stock as a function of the stock's return and then tests those implications empirically. Three different approaches that have been used in the literature are shown to lead to the, at first sight, contradictory findings that the probability to sell a stock is either V-shaped or inverted V-shaped in the stock's return. Since these approaches compute different conditional probabilities, they can be reconciled, however, when the conditioning set is taken into account. The conclusion is that realization utility is rejected as an explanation of the disposition effect whereas a particular version of prospect theory appears to be consistent with it.

The extent to which machine learning methods and big data will leave a footprint in (economic) research is yet unclear. Expectations run high and some have already declared the end of traditional science.[5] My own view is more pragmatic: Machine learning methods can assist researchers to find structure in data sets but they do not replace the frequent need for a good research design. Big data sets provide potentially exciting opportunities but they do not replace the need to define an interesting research question.[6] The process of scientific discovery will almost certainly evolve differently from what we imagine but expanding the toolkit and the data basis should be universally good (at least, weakly).

[5] http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory

[6] In conversations with classmates, we referred to that as the *big data curse*: The difficulty of turning access to large and promising data sets into well-defined research questions.

Chapter 1

Deep conditional portfolio sorts: The relation between past and future stock returns[1]

1.1 Introduction

Consider the challenge of a portfolio manager who wants to utilize past information to estimate expected returns at the firm level. He has at his disposal an overwhelming set of potentially correlated predictor variables as documented by a number of recent survey papers. Subrahmanyam (2010) surveys 50 earnings-based return predictive signals, McLean and Pontiff (2012) document 82, Harvey *et al.* (2013) and Green *et al.* (2013) both extend the list to around 330. These variables range from classic accounting-based variables like book-to-market to return-based variables like the stock return over the previous year to more exotic ones like the creativity of a stock's ticker. Figure 1.1 shows two graphs from Harvey *et al.* (2013) and Green *et al.* (2013) that illustrate the rate of discovery of predictor variables

[1]Co-authored with Benjamin Moritz

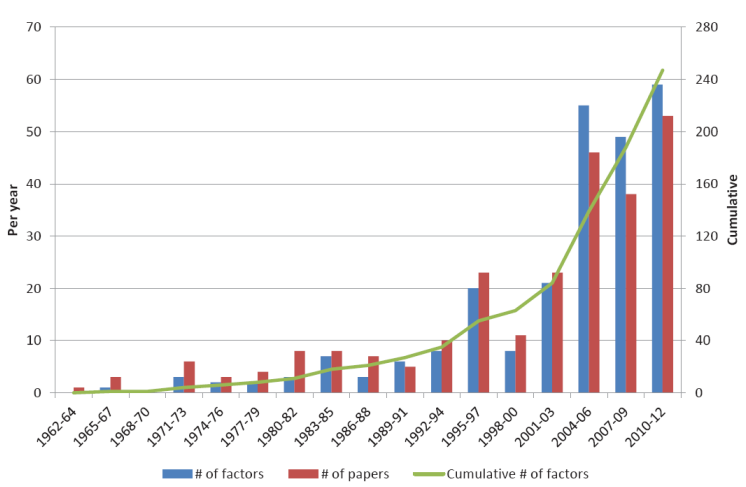
over time.[2] Both panels show a strong upward trend in the number of published (Harvey *et al.*) or publicly available (Green *et al.*) articles that report new predictor variables of returns, particularly in the last decade. Many of these variables might interact in non-trivial ways, which increases the set even more. In addition, the literature suggests a number of stand-ins for many variables (e.g. value or quality); which one should the manager pick? On top of these questions lurks the risk of overfitting the data with any estimation method that the manager might use, rendering the analysis worthless for new observations. How should one then go about estimating expected returns while taking all of these issues into account?

The literature in empirical asset pricing provides a few methods to assist the manager in his decision-making. As we will show, however, two prominent methods, portfolios sorts and Fama-MacBeth regressions, can only deal with a subset of the questions posed above. We suggest an alternative approach that is motivated by the method of *conditional* portfolio sorts but that extends easily to large sets of predictor variables and flexibly deals with their interactions. In contrast to how conditional portfolio sorts are usually applied, we estimate both the optimal conditioning variables and associated optimal thresholds from the data.

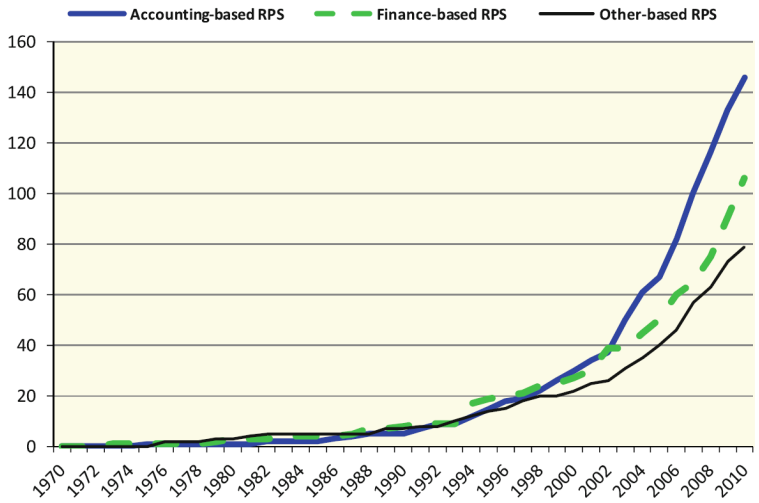
Our contribution to the literature is threefold: First, we provide a framework that can be used to organize different methods of estimating expected returns. The framework illustrates that these methods can be thought of as different approximations of a conditional expectation and it can be used to evaluate the relative merits of different techniques on simple metrics. We argue that, within this framework, portfolio sorts and Fama-MacBeth regressions, are not suited to evaluate the independent information in the entirety of many cross-sectional predictor variables and their potential interactions.

Second, we import ideas from the machine learning literature and tailor them to a financial application in order to produce a model that works in this context. While our method is data-driven in nature, we are careful to develop valid out-of-sample validations

[2]Note that these papers use a different terminology: Predictor variables are *factors* in Harvey *et al.* (2013) and *return-predictive signals* in Green *et al.* (2013).



(a) *Harvey et al. (2013)*



(b) *Green et al. (2013)*

Figure 1.1: *Time trends in the discovery and publication of return predictive signals*

of the model. As the machine learning literature is often criticized for producing black-box predictions, we put particular emphasis on new measures to extract interpretable information about the structure of the estimated prediction function.

Third, we apply our methodology to past-return based prediction of future returns, and we recover short-term returns (i.e. the past six most recent one-month return) as the most important predictors. Implementable trading strategies based on our findings have a risk-adjusted monthly return of around 2 percent per month, with an information ratio that is about three times as high as the information ratio that can be achieved in a linear framework that does not account for non-linearities and variable interactions and twice as high as in a Fama-MacBeth framework that accounts for two-way interactions. Transaction costs cannot account for our results.

While this paper focuses on a particular application, the methodology can be applied more generally and it has interesting implications for the analysis of cross-sectional predictor variables going forward that we discuss in the conclusion.

We start by documenting some results that are based on standard methodologies in finance. We show that, if the investor above had used those methodologies to estimate future returns from past returns, he could have made reasonable returns of around 1 percent per month (after controlling for risk factors) with information ratios of about 1. We also show that, had the investor taken potential two-way interactions between past returns into account, he could have earned similar monthly returns at an information ratio of 1.3, that is, at much reduced risk. Similarly, when we repeat Fama and French (2008)'s exercise and extend it to a number of other variables, we show that there are important interactions between past returns and firm fundamentals.

These results pose a challenge for existing methodologies when the goal is to evaluate many variables in a joint framework. The portfolio sort methodology, a dominant method in analyzing cross-sectional predictor variables,[3] each month (or year) sorts stocks into three to ten portfolios based on the value of a particular variable. In the next step, subsequent

[3]See the survey of Green *et al.* (2013).

returns for each portfolio are calculated and it is checked whether there is a monotone relation between the sorting variable and these subsequent portfolio returns. In addition, researchers often compute the equal- or value-weighted hedge return of going long (short) the highest quantile portfolio and going short (long) the lowest quantile portfolio. The relevance of the sorting variable is then assessed by comparing the hedge return to some equilibrium model of asset prices (e.g. the capital asset pricing model) and/or by assessing the monotonicity of the returns over deciles. With regard to the former, a sorting variable is considered relevant if the hedge return strategy makes abnormal returns that are statistically different from zero. With regard to the latter, Patton and Timmermann (2010) provide a test for monotonicity in one- or two-variable sorts. The portfolio sort methodology is a powerful, non-parametric, tool that works best in low dimensional cases. Problems arise if returns are to be sorted on more than two or three predictor variables as there will be typically be few stocks in each portfolio. But this makes it challenging to control for information contained in other variables or, as Fama and French (2008) put it, "sorts are awkward for drawing inference about which anomaly variables have unique information about average returns."

Multivariate Fama-MacBeth regressions are able to address this concern by showing the marginal effect of each predictor variable once all others are controlled for. The methodology is based on estimating a cross-sectional regression in each period and averaging the coefficient estimates over time. This works well with a larger number of predictor variables. But when we include interactions between predictor variables, this methodology reaches its limit, too: Even if only fifty variables are considered jointly, the total number of regression coefficients that include all two-way interactions (and no higher-order interactions) is 1275, higher than the number of companies in early months of the sample, and higher than the number of companies throughout the entire sample if the sample is split by firm size first as in Fama and French (2008). Second, as Fama and French note, results can be vulnerable to influential observations of extreme returns. With this in mind, Green *et al.* (2014) "view it as infeasible to examine non-linearities in RPS-returns relations in the manner undertaken in Fama and French (2008)."

We suggest a method that is based on the well-known idea of conditional portfolio sorts that is designed to address the aforementioned challenges and that can account for arbitrary interaction terms.[4] Conditional portfolio sorts arrange firms into groups based on the value of some variable (e.g. book-to-market). Within each group, stocks are then sorted again based on the value of some other variable. Sorting variables and sorting values are typically chosen based on a-priori reasoning. We start from the assumption that neither the sorting variable nor the sorting value are known and need to be estimated. Furthermore, conditional sorts are typically conducted for no more than two levels (that is, stocks are sorted twice) and the same sorting variable is used in all branches on the second level. We estimate sorts at deeper levels (motivating the method's name in the title) and allow for flexible variable selection at each branch.

The optimization problem is computationally challenging but can be solved with insights from the machine learning literature. The solution follows a simple algorithm that, for each portfolio of firms and starting from the portfolio of all firms (the entire data set), splits the firms in the portfolio into two new groups. The algorithm finds the sorting variable and associated sorting value that minimize a loss function over the data in the two resulting groups. The optimization is repeated at every non-terminal node using the remaining observations as long as that number is not too small and there is still a split of the data that significantly improves upon the value of the loss function.

There are two well-known and related problems with this approach. First, since the optimization proceeds stepwise, the variables and sorting values that are selected at each point need not be globally optimal. But since the sorting rule is discrete, any error in the estimation of the sorting variable and sorting value could have a large impact on the model's predictions. Second, the approach is data-driven and easily overfits the data. We, therefore, need to take great care to make sure that the estimates are valid out-of-sample.

The solution that we employ is based on model-averaging. We estimate deep conditional

[4]The finance literature is somewhat imprecise about the distinction between interaction terms and non-linearities, and often uses both terms interchangeably. We reserve "interactions" for cross-products between two variables, and we think of "non-linearities" as higher-order polynomial terms with respect to a single variable.

portfolio sorts many times, with different subsets of regressors and on different subsets of the data, and combine the estimates from all models into a final prediction. The rationale is that by averaging estimates that come from models that are de-correlated in this manner, one can obtain different but related signals about the true underlying process, even if the simple underlying models are not entirely correct. At the same time, model-averaging helps with the overfitting problem because only subsets of the data and predictor variables are used in each model. The idea is grounded in the computer science literature and has been successfully applied in many contexts. We find that deep conditional portfolio sorts combined with model-averaging produces very accurate predictions of expected returns.

The main drawback of averaging over many models is that results are not as easy to interpret as a single deep conditional sort. In order to shed light on the mechanism, we suggest a number of evaluation measures. We compute a measure of predictor variable importance that can be interpreted similarly to t-statistics in regressions. In addition, we develop a way to compute partial derivatives for each predictor variable so that we are able to talk about directional effects of specific variables. We also run diagnostic checks to see whether the predictions from the model can be explained by a simple linear regression on our predictor variables (which would speak against the importance of interaction effects).

The method takes into account that a predictor variable's influence might vary over time.[5] We set up the out-of-sample tests in such a way that they lend themselves naturally to investigate time-variation of the importance of particular variables. In each year, we estimate the model with data over the past years. For the next twelve months, one-month expected returns are then projected by fixing the model estimates and making predictions based on the new data that were reported only after the estimation period. Not only are our trading results below strong in this exercise, but the approach also enables us to look at which variables come out as important in the search procedure in which period.

We apply our method to contribute to the debate about whether past returns contain

[5] As Harvey *et al.* (2013) note "it is possible that a particular factor is very important in certain economic environments and not important in other environments. The unconditional test might conclude the factor is marginal."

information about future returns and, if so, which past returns matter the most. This debate has recently regained interest after Novy-Marx (2012) found that medium-term momentum, that is a stock's return over the twelve to seven months prior to portfolio formation, can be a better indicator of future return than momentum calculated over the entire previous year (excluding the most recent month). Goyal and Wahal (2013) cannot find this effect in 37 other markets outside the US. Other recent articles have looked at a moving average strategy derived from past prices (Han *et al.* (2011)) or construct a trend factor from daily to annual returns that outperforms the standard momentum factor (Han and Zhou (2013)). We, therefore, regard the relation between past and future returns as a natural laboratory for our method.

As predictor variables, we construct a set of decile rankings for the non-overlapping one-month returns over the two years before portfolio formation. This yields a set of twenty-five predictors and it is ex-ante unclear how to combine them optimally to forecast next period's returns.

We use standard methods to derive forecasts and benchmark them against forecasts from deep conditional portfolio sorts. A strategy based on deep conditional portfolio sorts yields abnormal returns (relative to the four-factor model) of 2-2.3 percent per month, depending on the exact specification, with information ratios of around 2.8. Our preferred specification has an abnormal monthly return at the lower end of that range. Although the strategy has high turnover, transaction costs do not dwarf the abnormal return. This compares to results from a Fama-MacBeth regression framework with abnormal returns of 1-1.4 percent per month with an information ratio of 1-1.5, depending on whether two-way interactions between past returns are included. We conclude that deep conditional portfolio sorts perform better via producing a moderate increase in average abnormal returns at much reduced variance.

What is the structure of the prediction function that we estimate? While it cannot be summarized as a simple linear equation, we can use our suggested evaluation measures to shed light on the black box: Intriguingly, the most important predictor variables are

short-term return functions and returns appear to become less important when they are in the more distant past. In particular, we show that the most recent six months of past returns capture almost all the information that is contained in more distant past returns.

We then show how past returns are related to future returns in the deep conditional portfolio sorts. While we recover some standard results like short-term reversal over the most recent past month or momentum over the previous twelve months of returns, we also find evidence for the relevance of non-linear effects (e.g. both high and low returns over the month before the most recent one predict lower returns) and interactions (e.g. the one-month return over the second-to-last month is negatively related to returns for stocks with low returns last month, but is positively related to returns with high returns last month).

The results hold in a variety of alternative settings. We construct another set of predictor variables that includes many possible past returns with different horizons and gaps to the portfolio formation date to see how our methodology performs when many of the predictor variables (a total of 126) are highly correlated. In this setting, abnormal returns are again high and a similar return structure, with similar partial derivatives for specific predictor variables, is estimated. Our results are also unaffected by including eighty-six additional firm characteristics from the literature. Here, results for abnormal returns are actually a bit stronger because of the additional information in accounting variables and other characteristics, and the return structure results still hold. We then make sure that our results are not entirely driven by illiquid stocks by re-doing all computations for large, small and micro firms (in the terminology of Fama and French (2008)) separately. While we find that results are stronger in small stocks and strongest in micro stocks, our main conclusions hold throughout all size categories. We conclude that more recent past returns are more relevant than intermediate past returns for prediction of future returns and, more generally, past returns are related to future returns in a more complex way than can be captured by any single one past return.

Before we continue, we provide a short overview of the related literature. In his

presidential address, Cochrane (2011) describes the "factor zoo" of stock market anomalies and how it has developed over the years. Subrahmanyam (2010), Goyal (2011), Green *et al.* (2013) and Harvey *et al.* (2013) review as many as 330 anomalies that have been found by academic research and call for a synthesis of the existing literature. While early attempts in this direction where undertaken by Haugen and Baker (1996), Daniel and Titman (1997) and Brennan *et al.* (1998) who focus on smaller sets of characteristics, Cochrane (2011) argues that different methods might be required to find the independent information for average returns in the entirety of documented predictor variables. Our paper can be read as an attempt to provide just such a new method.

Green *et al.* (2014) investigate the mutual information in 100 signals, and find that up to 24 of them have predictive power for returns when used jointly. They suggest an alternative to the standard three factor model by Fama and French (1992) that is based on 10 different characteristics. The paper notes the potential relevance of interactions but does not investigate them in detail.[6] Lewellen (2013) investigates the power of 15 different firm characteristics to predict variation in the cross section. He finds that expected stock returns derived from the model are strongly predictive of actual stock returns for as much as 12 months.

Fama and French (2013) follow an alternative approach that attempts to capture variation in returns by a (small) factor model. They extend the three factor model by proxies for profitability and investment which appears to capture contemporaneous variation in cross-sectional returns well, except for small stocks. The paper uses a quadruple sorting strategy to address interactions between size, value, profitability and investment opportunities. Kogan and Tian (2012) construct all combinations of three and four factor models from a set of 27 firm characteristics. They find that the best performing models are unstable across time periods.

The literature on momentum and reversal is too large to review it here but we note a view

[6]They write, "fundamental valuation type measures and market trading type measures appear to matter across firm size. In large-cap firms the important RPS can be broadly classified as fundamental valuation measures or trading type measures. For mid-cap and small-cap firms the themes appear slightly different."

key articles. If stock prices systematically over- or underreact, future stock returns should be predictable from past returns data alone. de Bondt and Thaler test overreaction by sorting stocks based on the return in the previous three years (the portfolio formation period). They find that losers (the bottom decile of returns in the formation period) outperforms winners by about 25% over three years. They hint at the fact that there is some interaction with the January effect. A similar "reversal effect" has been found by Jegadeesh (1990) and Lehmann (1990) for portfolios that are formed based on short-term (one week to one month) prior returns. Jegadeesh and Titman (1993), on the other hand, find evidence for a "momentum effect" when portfolios are sorted on medium-term (3 to 12 months) prior return. "Momentum" means that past winners continue to outperform past losers for up to 12 months (with an apparent reversal effect after 12 months). The momentum finding survives the analysis in Fama and French (1996) who use the three-factor model as a model of equilibrium returns. Long-term reversal disappears as an anomaly once normal returns are approximated by the three factor model. For much more on momentum, we refer to Asness *et al.* (2014) who use simple analysis and survey published studies to show that momentum returns are (among other things) not too volatile, not only a small firm phenomenon and not dwarfed by tax considerations or transaction costs.

This essay is organized as follows. Section 1.2 discusses the data, sets up a motivating framework and investigates two standard methods, portfolio sorts and simple Fama-MacBeth regressions, that a portfolio manager could employ to predict future returns. Section 1.3 explains deep conditional portfolio sorts in detail. Section 1.4 applies the method to past return predictor variables and section 1.5 has further results on transaction costs and a risk factor vs characteristics interpretation. Section A.3 in the appendix illustrates robustness of our results along several dimensions. Section 1.6 concludes.

1.2 Data, motivating framework and standard methods

Before we introduce deep conditional portfolio sorts, we analyze a few standard approaches that an investor might try. These are: Single variable selection, i.e. investing based on the

single best-performing variable in historical data over a certain time window; standard Fama-MacBeth regressions, i.e. a multivariate prediction that combines historically important signals; and Fama-MacBeth regressions that include variable interactions.

1.2.1 Data

Since we will use the relation between past returns and future returns as a running example throughout the article, we start by describing the data and the variable construction first.

The basis for our analysis is the universe of monthly US stock returns from the Center for Research in Security Prices (CRSP) from 1963 to 2012. Since we use firm characteristics from Compustat and IBES in some robustness checks, we match stock price data to those data sets first, and focus our analysis on those firms that can be linked in all datasets. Firm characteristics include traditional variables like size, book-to-market, dividend yield, gross profitability and eighty-two others that are described in more detail in appendix A.3.1. The number of firms in our sample varies over time between 1182 and 6626. Size, value, momentum factors and the risk-free interest rate are taken from Kenneth French's data library.[7]

Figure 1.2 illustrates how return-based predictor variables are constructed. Suppose that the investor wants to form a portfolios at the formation time t_f. Return-based predictor variables can be defined by two parameters; the *gap* between the time of portfolio formation and the most recent month that is included in the return calculation, and the *length* of the return computation horizon. We denote the former by g, the latter by l and a return function by $R_{i,t_f}(g,l)$ maps returns into cross-sectional decile ranks. For example, $R_{i,t_f}(1,11) = 10$ implies that firm i is in the highest decile of returns at time t_f for the return that is computed over the previous twelve months and leaves out the most recent one.

Our benchmark set of predictors contains all one-month returns over the two years before portfolio formation, that is, $R_{i,t}(g,1), g = 0,\ldots,24$. Much of the related literature is based on sorting firms into one of ten deciles depending on the values of a sorting variable.

[7] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

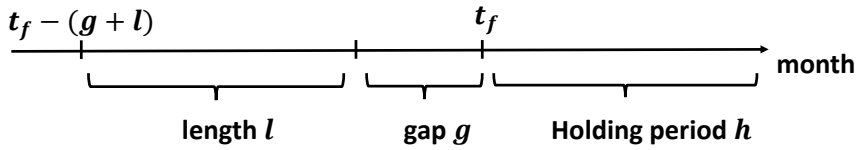


Figure 1.2: *Construction of past return-based characteristics*

Notes: The investor forms a portfolio at time t_f. Return-based predictor variables can be defined by two parameters; the gap *between the time of portfolio formation and the most recent month that is included in the return calculation, and the* length *of the return computation horizon. We denote the former by g, the latter by l and a return function by $R_{i,t_f}(g,l)$ maps returns into cross-sectional decile ranks.*

When we consider return-based strategies below, we refer to buying the upper decile and selling the lower decile based on $R_{i,t}(g,l)$. As in Novy-Marx (2012), we will use the notation $R_{i,t}(g,l)$ to denote both the return for portfolio formation, and the strategy return based on that simple sorting strategy.[8]

The problem of predicting future returns based on past returns has the ingredients that make it difficult for an investor to find the relevant signals: Should momentum be measured over the most recent six or twelve months? What if the signals go in opposite directions? Should one leave out the most recent month? Or the most recent six (Novy-Marx (2012))? Degrees-of-freedom in choosing the gap and length parameters above contribute to the fact that these questions do not have a definitive answer yet.

1.2.2 Motivating framework

In each time period, an investor has access to information Θ_{it} about firm i to model the conditional expectation of next period's return as in equation (1.1), a general version of the

[8]We have checked that results are robust when future returns are computed over the next future month, but skip a day to make sure that the return would actually be implementable.

model[9] that is typically estimated in the literature.

$$E_t[r_{i,t+1}|\Theta_{it}] = f_t(\Theta_{it}), \tag{1.1}$$

Here, the expectation of $r_{i,t+1}$ is formed at time t (we take a period to be one month in what follows), and the function $f_t()$ that maps the information set into expected returns can be time-varying. The information set Θ_{it} can contain data on the firm's past earnings, balance sheet information, past stock return movements and many other variables. Since we will focus on the relation between past and future returns in this paper, and in line with the sorting-based literature, we assume that the information set consists of decile rankings of companies over the past two years, that is, $\Theta_{it} = \{R_{i,t}(0,1), \ldots, R_{i,t}(24,1)\}$. In other words, we consider decile rankings for each of the most recent twenty-five one-month returns.

With that information set, adding an additive error term and choosing the common specification of a linear form (see e.g. Haugen and Baker (1996), Daniel and Titman (1997) or Brennan *et al.* (1998)) for the function $f_t()$, equation (1.1) can be written as

$$r_{i,t+1} = a + \sum_{g=0}^{24} \beta_g^t R_{i,t}(g,1) + \epsilon_{i,t}, \tag{1.2}$$

which is usually estimated via a Fama-MacBeth procedure or by a cross-sectional regression. In general, the model can be viewed as a joint test of the relevance of characteristics and of the linearity assumption. We first illustrate how an investor could go about predicting returns using standard methods.

1.2.3 Existing methods

In our running example, the investor faces the problem of predicting returns based on one-month returns over the previous two years. We consider two possible solutions to that

[9] At a greater level of generality, one could write the model as

$$E_t[r_{i,t+1}|\Theta_{it}] = f_t(z_{i,t}, z_{i,t-1}, \ldots, \lambda_t, \lambda_{t-1}, \ldots),$$

which would also include risk factors, and z_{it} and λ_t and their histories are subsumed in the information set $\Theta_{it} = \{z_{i,t}, \ldots, \lambda_t, \ldots\}$ at time t. We disregard this aspect for now but note that our framework easily extends to the case where all returns are interpreted as excess returns over risk factors.

problem that are employed in the existing literature.

Portfolio sort

The potentially simplest strategy is to evaluate one variable at a time, and then base forecasts on the single variable that has performed best in the past. More specifically, we suggest the following simple strategy: In each month, compute the m month trailing average return for each sorting variable, pick the one with the best performance (in terms of the Sharpe ratio), and base the subsequent long and short orders on values of that variable.

Table 1.1 shows that the return to such a strategy is .71 percent per month with an information ratio (relative to the four factor model) of .89, when the trailing performance is computed over the sixty months that precede the portfolio formation date. While this is already a good result, each month's returns are based on the values of a single sorting variable. The question remains whether the investor can do even better by combining information from different variables. While a few more variables can be incorporated (e.g. double sorts), the number of observations in each portfolio decreases quickly such that estimates become unreliable.

Fama-MacBeth regressions

With that question, the investor turns to a multivariate regression setup that we describe in some detail. We suggest two approaches: A "kitchen-sink" Fama-MacBeth estimation that throws in all past return variables and that uses them for prediction regardless of their individual significance. On the other hand, it could be more appropriate to base predictions solely on the relevant variables where we define "relevant" as variables that are selected in a LASSO regression.[10] While we report results for the LASSO regression, we have tried other model selection methods (general-to-specific, specific-to-general) and obtained similar

[10]Least absolute shrinkage and selection operator (LASSO), originally introduced by Tibshirani (1996), is a method that regularizes regressions by putting a penalty on the size of regression coefficients. Due to the nature of the penalty term (the sum of the absolute values of individual coefficients), the optimum will typically set many coefficients to exact zeros, which is why the method can be viewed as a variable selection device.

Table 1.1: *Strategy factor loadings: Portfolio Sort*

	(1)	(2)	(3)	(4)
Intercept	0.71	0.74	0.71	0.72
	(6.45)	(6.77)	(6.60)	(6.67)
MKT		-0.03	-0.03	-0.03
		(-1.51)	(-1.28)	(-1.28)
SMB			0.03	0.03
			(0.73)	(0.73)
HML			0.04	0.04
			(1.17)	(1.16)
UMD				-0.00
				(-0.11)
R^2		0.00	0.01	0.01
IR		0.92	0.89	0.89
SR	0.88			
N	540	540	540	540

This table shows time-series regressions of strategy returns on factors. Returns are specified in percent per month. The strategy is to go long (short) the highest (lowest) decile of firms based on a single past return variable from the set of the most recent twenty-five past one-month returns. In each month, the past return that would have produced the highest strategy Sharpe ratio over the sixty preceding months is selected as the sorting variable. MKT is the market return, SMB and HML are the Fama-French factors for size and value, and UMD is the momentum factor. SR is the Sharpe ratio and IR is the information ratio. The sample period covers 1968 to 2012. T-statistics are in parentheses, and standard errors are clustered using Newey-West's adjustment for serial correlation.

results.

Our general implementation for the Fama-MacBeth-framework works as follows. In each cross-section, the investor fits the regression

$$r_{i,t+1} = \beta^t_{cons} + \sum_{g=0}^{24} \beta^t_g R_{it}(g,1) + \epsilon_{it} \tag{1.3}$$

and he keeps either all coefficients (kitchen sink) or uses LASSO to select the relevant variables.

His period $t+1$ forecast is computed based on the rolling average of the coefficient estimates up to period $t-1$ and then applying the linear model to $R_{it}(g,1)$, that is,

$$\hat{r}_{i,t+1} = \overline{\beta}^{t-1}_{cons} + \sum_{g=0}^{24} \overline{\beta}^{t-1}_g R_{it}(g,1), \tag{1.4}$$

where $\overline{\beta}^{t-1}_g = \frac{1}{m}\sum_{j=t-1-m}^{t-1} \hat{\beta}^t_g$. We initially use a rolling window of 120 months but, as Lewellen (2013), have found that results are robust to varying that parameter.

Lewellen (2013) uses a set of 15 predictor variables that are well-established in the literature. In contrast, we consider an investor who faces substantial uncertainty about which variables he should include and, therefore, has to cast a wide net. Consistent with our running example, the investor considers all one-month returns over the two years before portfolio formation. Each period, he computes return predictions based on past model estimates, and sorts predictions into ten deciles. He constructs an equal-weighted hedge portfolio that goes long the highest decile of predicted returns and that goes short the lowest decile of predicted returns, analogous to the strategies above.

Starting with the kitchen sink model, the first four columns of table 1.2 show the strategy's factor loadings from time-series regressions on the market, size, value and momentum factors. The strategy has a positive and significant average return of 1.51 percent per month, and loads mostly on the market and the momentum factor. The alpha relative to the four-factor model is about 1 percent per month, with an information ratio of about 1.

When we use the LASSO in the Fama-MacBeth framework as described above, results remain almost unchanged. The last four columns of table 1.2 show that the average strategy

Table 1.2: *Strategy factor loadings: Fama-MacBeth predictions using all variables*

	Kitchen sink regression				LASSO regression			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Intercept	1.51	1.33	1.30	1.00	1.50	1.31	1.28	1.00
	(8.93)	(7.87)	(8.06)	(6.23)	(8.63)	(7.60)	(7.79)	(6.16)
MKT		0.20	0.20	0.26		0.21	0.21	0.26
		(3.07)	(3.08)	(4.34)		(3.29)	(3.34)	(4.51)
SMB			0.05	0.06			0.05	0.05
			(0.58)	(0.51)			(0.53)	(0.47)
HML			0.05	0.14			0.05	0.14
			(0.40)	(1.35)			(0.44)	(1.31)
UMD				0.30				0.27
				(4.25)				(3.76)
R^2		0.06	0.06	0.18		0.06	0.07	0.16
IR		1.23	1.20	0.98		1.20	1.17	0.97
SR	1.36				1.33			
N	540	540	540	540	540	540	540	540

This table shows time-series regressions of strategy returns on factors. Returns are specified in percent per month. Strategies are based on the predictions of a Fama-MacBeth regressions of future returns on past decile sorts of returns. Past return sorts include decile rankings R(g,l) with length l equal to 1 and gap g between 0 and 24 months (i.e. all one-month returns over the two years before portfolio formation), that is, predictions are based on the equation

$$r_{i,t+1} = \beta^t_{cons} + \sum_{g=0}^{24} \beta^t_g R_{it}(g,1) + \epsilon_{it}.$$

The kitchen sink Fama-MacBeth model uses all variables in each period regardless of their significance, and the LASSO model selects a set of relevant variables each period based on a penalty function approach. Both procedures are described in section 1.2.3. Strategies go long the highest predicted return decile and go short the lowest predicted return decile. The sample period covers 1968 to 2012, and all results are based on rolling out-of-sample estimates of the models. MKT is the market return, SMB and HML are the Fama-French factors for size and value, and UMD is the momentum factor. SR is the Sharpe ratio and IR is the information ratio. T-statistics are in parentheses, and standard errors were clustered using Newey-West's adjustment for serial correlation.

return is again around 1.5 percent per month, and the four-factor alpha is 1 percent per month. The information ratio is close to 1, as in the kitchen sink regression. The reason that these results are very similar is that many irrelevant regressors have coefficients close to zero in the kitchen sink case.

Note that the approaches so far have not included variable interactions. The Fama-MacBeth regression framework lends itself to a simple implementation of additionally including interactions of predictor variables. Equation (1.5) shows the regression equation that adds all two-way interactions between past return rankings.

$$r_{i,t+1} = a + \sum_{g=0}^{24} \beta_g^t R_{i,t}(g,1) + \sum_{g=0}^{24} \sum_{j>g} \gamma_{gj}^t R_{i,t}(g,1) R_{i,t}(j,1) + \epsilon_{i,t}. \tag{1.5}$$

Table 1.3 shows strategy returns that are based on predictions from equation (1.5).[11] At 1.13 percent per month, the average excess return relative to the four factor model is slightly higher than in the levels-only version above. The information ratio, however, experiences a much stronger increase to 1.3.-1.4. Hence, the main benefit to including two-way interactions appears to be a reduction in variance rather than an improved mean return.

Of course, this begs the question whether we have now captured all information in past returns for future returns or whether we should estimate the prediction equation more flexibly. For instance, if we are interested in exploring all systematic variation, why would we stop at two-way interactions? Appealing as the Fama-MacBeth method might seem, it quickly becomes infeasible when we want to analyze the entirety of potential interactions. Considering only two-way interactions, the number of terms to include when p candidate predictors are included is $\frac{p(p+1)}{2}$ which starts to become greater than a thousand at a mere forty-five predictor variables. This prevents the use of Fama-MacBeth regressions in the early years of the sample (although LASSO would still be a feasible alternative) if all firms are considered, and over the entire sample if the sample is divided by, say size, first. With higher-order interactions, estimation becomes difficult for even fewer candidate

[11] Since the model with two-way interactions has 325 regressors, we focus on results based on variable selection.

Table 1.3: *Strategy factor loadings: Fama-MacBeth predictions using all variables and two-way interactions*

	(1)	(2)	(3)	(4)
Intercept	1.46	1.27	1.28	1.13
	(9.62)	(8.29)	(8.35)	(7.55)
MKT		0.21	0.18	0.21
		(4.38)	(3.82)	(4.40)
SMB			0.12	0.12
			(1.63)	(1.48)
HML			-0.03	0.02
			(-0.33)	(0.20)
UMD				0.15
				(2.91)
R^2		0.09	0.11	0.15
IR		1.43	1.46	1.32
SR	1.57			
N	540	540	540	540

This table shows time-series regressions of strategy returns on factors. Returns are specified in percent per month. Strategies are based on the predictions of a Fama-MacBeth regressions of future returns on past decile sorts of returns. Past return sorts include decile rankings R(g,l) with length l equal to 1 and gap g between 0 and 24 months (i.e. all one-month returns over the two years before portfolio formation) and their two-way interactions, that is, predictions are based on the equation

$$r_{i,t+1} = \beta^t_{cons} + \sum_{g=0}^{24} \beta^t_g R_{i,t}(g,1) + \sum_{g=0}^{24} \sum_{j>g} \gamma^t_{gj} R_{i,t}(g,1) R_{i,t}(j,1) + \epsilon_{i,t}.$$

LASSO estimation is applied to select relevant variables each period, described in more detail in section 1.2.3. Strategies go long the highest predicted return decile and go short the lowest predicted return decile. The sample period covers 1968 to 2012, and all results are based on rolling out-of-sample estimates of the models. MKT is the market return, SMB and HML are the Fama-French factors for size and value, and UMD is the momentum factor. SR is the Sharpe ratio and IR is the information ratio. T-statistics are in parentheses, and standard errors were clustered using Newey-West's adjustment for serial correlation.

predictors. In the next section, we import a method from the machine learning literature that is sufficiently flexible in this setting and tailor it to a finance application.

1.3 Estimation strategy

Returning to the general model for expected returns in equation (1.1), we briefly discuss the difficulties that arise when the set of firm characteristics gets large. More specifically, even when the set of characteristics appears manageable, the number of regressors can grow quickly if characteristics interact or are non-linearly related to returns.

Interactions between different anomalies can arise quite naturally from simple economic models. Chen *et al.* (2002) test the theory of gradual information diffusion to explain momentum. They argue that the rate of information diffusion could be different for firms which would result in different strength of momentum profits. They find that momentum interacts with firm size and with analyst coverage, and that the effect of analyst coverage on momentum profits is largest in small firms (a triple interaction). Vassalou and Xing (2004) illustrate a complex interaction between size and value and default risk. They show that small stocks earn higher returns than big stocks only if they have higher default risk and the same holds for the return of value over growth stocks. Complementary, high default risk firms earn higher returns than low default risk firms if they are small or value stocks. Expected return-relevant two-way interactions have been demonstrated between size and value (Fama and French (1992)), between size and seasonal effects (Daniel and Titman (1997)), or between stock exchange and volume (Brennan *et al.* (1998)). Some authors have also considered interactions between past-returns and firm fundamentals (see e.g. Asness (1997) for the interaction between value and momentum, or (Lee and Swaminathan (2000)) for the interaction between volume and momentum). Interactions between different past-return variables are rare in the literature, with Grinblatt and Moskowitz (2004) who consider the consistency of return patterns and Han and Zhou (2013) who construct a trend-factor from past returns of different frequencies being two exceptions.

The literature that investigates interactions has typically used portfolio sorts. This

approach sorts stocks into portfolios based on the characteristics in question, and the returns for each portfolio are evaluated. It is, however, only feasible for a small set of, typically two, characteristics. Three-way or four-way sorts are rarely executed at all because the individual portfolios contain few firms.[12] Correlations between firm fundamentals make it difficult to isolate their individual marginal contribution to expected return prediction.[13]

One might think that a fully interacted version of equation (1.2) can overcome this challenge, but in fact becomes infeasible quickly, too. Consider a model that allows for arbitrary three-way interactions

$$\begin{aligned} r_{i,t+1} = a + \sum_{g=0}^{G} \beta_g^t R_{i,t}(g,1) + \sum_{g=0}^{G} \sum_{j>g} \gamma_{gj}^t R_{i,t}(g,1) R_{i,t}(j,1) \\ + \sum_{g=0}^{G} \sum_{j>g} \sum_{k>j} \delta_{gjk}^t R_{i,t}(g,1) R_{i,t}(j,1) R_{i,t}(k,1) + \epsilon_{i,t}. \end{aligned} \tag{1.6}$$

Even if we consider a small set of $G = 20$ firm characteristics, 190 two-way interactions and 1140 three-way interactions would need to be considered. In the application of Green *et al.* (2014) with $G = 100$, these numbers amount to 4950 and 161700 which is prohibitively large for statistical analysis.[14]

Given the difficulties that stem from comprehensively investigating the interactions between characteristics using these standard methodologies, the existing evidence is restricted to the low-dimensional cases that have been and can be considered, while we may not learn the full extent to which interactions are relevant. Our approach below provides one way to address this question.

1.3.1 Conditional Portfolio Sorts

Our goal is to estimate the conditional expectation in equation (1.1) more flexibly than can be achieved by a globally linear model like a Fama-MacBeth regression or by portfolio

[12] For examples, see Daniel and Titman (1997), Fama and French (2008) or Fama and French (2013).

[13] An early contribution that criticizes portfolio sorts for their inability to deal with correlated signals can be found in Jacobs and Levy (1989).

[14] In general, all k-way interactions are given by $\binom{G}{k}$.

sorts that allow for non-linearities but tha6t, in their usual form, are restricted to one- or two-dimensional cases.

Our estimation is based on the well-known concept of *conditional* portfolio sorts which are illustrated schematically in figure 1.3. Consider sorting stocks into two portfolios based on sorting variable $R(g^{(1)}, 1)$ and threshold $\tau^{(1)}$, such that all stocks with $R(g^{(1)}, 1) \leq \tau^{(1)}$ are pooled together into one portfolio, and stocks with $R(g^{(1)}, 1) > \tau^{(1)}$ are pooled together into another portfolio. For instance, if $\tau^{(1)} = 5$ and $g^{(1)} = 0$, we would sort all stocks with returns below the cross-sectional median in the previous month into one portfolio and all stocks with returns above cross-sectional median in the previous month into another portfolio.[15] The expected stock returns in each portfolio are now $E[r_{i,t+1}|R(g^{(1)}, 1) \leq \tau^{(1)}]$ and $E[r_{i,t+1}|R(g^{(1)}, 1) > \tau^{(1)}]$, respectively, and, if the expected return is modeled as a constant within each portfolio, the prediction is just the average of realizations of next months' returns within each group. Sorting stocks within each portfolio again by another (or the same) characteristic with associated thresholds $\tau^{(2a)}$ and $\tau^{(2b)}$ results in four different portfolios S_1 to S_4, e.g. the stocks in portfolio S_1 in the figure have expected return $E[r_{i,t+1}|R(g^{(1)}, 1) \leq \tau^{(1)}, R(g^{(2a)}, 1) \leq \tau^{(2a)}]$.

A simple way to test whether $R(g^{(2a)}, 1)$ provides additional information over $R(g^{(1)}, 1)$ would compare the sorts on $R(g^{(2a)}, 1)$ within each portfolio sorted on $R(g^{(1)}, 1)$.[16] On the other hand, one could test whether $R(g^{(2a)}, 1)$ creates a return spread only in the portfolio of, e.g. low $R(g^{(1)}, 1)$ firms, therefore, testing for a potential interaction between characteristics $R(g^{(2a)}, 1)$ and $R(g^{(1)}, 1)$.

In appendix A.1, we illustrate a basic conditional portfolio sort with a few standard firm characteristics. Our results complement Fama and French (2008) who sort stocks into three size portfolios first and then sort each portfolio subsequently on a further firm characteristic. In our illustration, we consider conditional portfolio sorts that are each based on two of the

[15]The literature usually considers one-variable sorts of stocks into ten different portfolios. However, our sort into two portfolios is not restrictive because a one-variable sort into multiple portfolios can always be achieved by a repeated sort into two portfolios.

[16]This kind of test is, for example, applied in Bandarchuk and Hilscher (2012).

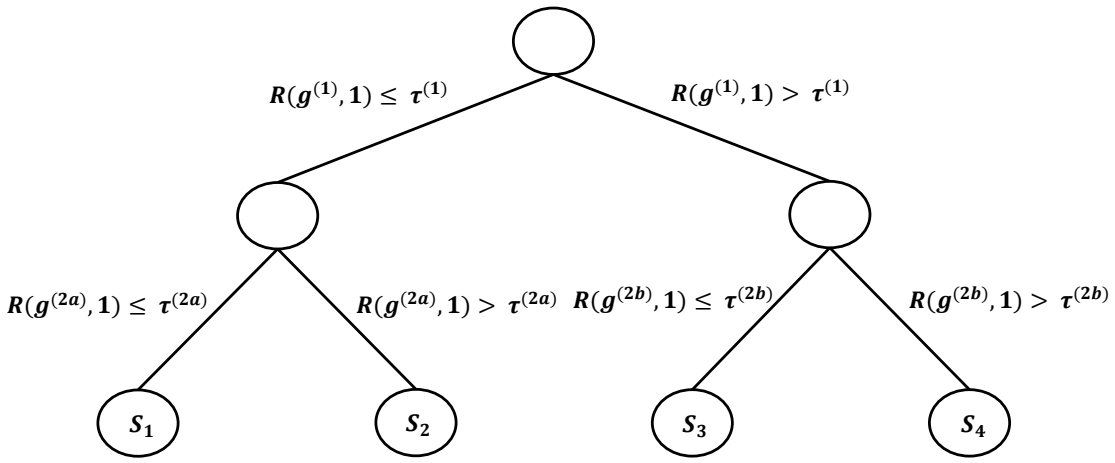


Figure 1.3: *Schematic representation of a conditional portfolio sort*

Notes: First, observations are sorted into two portfolios based on past return $R(g^{(1)},1)$ and threshold $\tau^{(1)}$. The resulting portfolios are then sorted again on variables $R(g^{(2a)},1)$ and $R(g^{(2b)},1)$ with thresholds $\tau^{(2a)}$ and $\tau^{(2b)}$ for a total of four portfolios S_1, S_2, S_3 and S_4.

following variables: short-term reversal, momentum, intermediate momentum, size, gross profitability, and book-to-market.

We refer the interested reader to the appendix for detailed results but highlight a few notable results here. The overall picture that emerges is that of return sorts being relatively stable while accounting-based sorts are less robust to initial sorts on some other return- or accounting-based variable. For instance, size sorts do not work uniformly when stocks are sorted on short-term reversal or momentum first. Interestingly, momentum sorts continue to work well when firms are sorted on intermediate momentum first but the reverse is not true.

Of course, there is the question of how to choose the sorting variables and the sorting thresholds in the first place. The literature typically chooses the sorting variables based on a specific hypothesis and uses thresholds that evenly sort stocks into three, five or ten portfolios. The same sorting variable is used in all branches after the first sort. But, if viewed as a way to approximate a conditional expectation of returns, this restriction might not deliver the best approximation. We relax these constraints in the next section.

1.3.2 Deep Conditional Portfolio Sorts

We suggest to extend the method of conditional portfolio sorts along the following dimensions. First, unlike in our example above that had thresholds and sorting variables chosen ex-ante, we will choose thresholds and sorting variables optimally (where "optimally" will be defined below) within each portfolio in a data-driven way. Second, we apply the procedure to levels deeper than the two levels that are usually considered which gives rise to, what we call, a *deep conditional portfolio sort*. Third, since conditional sorts involve hard thresholds that are sensitive to small changes in the data, their predictions do not work very well out-of-sample. Following Kleinberg (1990, 1996), Ho (1998) and Breiman (2001), we average over many deep conditional portfolio sorts to smooth out the decision boundary which improves predictions significantly, as explained below in more detail.

Our approach draws on parallel concepts from the machine learning literature. The techniques that we use to estimate deep conditional portfolio sorts mirror those that are used to estimate a so-called decision tree in computer science.[17] Model averaging or ensemble methods are also developed in that literature and they are successfully applied to areas as diverse as biology (DNA sequencing), psychology or motion sensing. Applications in economics are rare[18] and our paper can also be read as an attempt to investigate whether these techniques have something to add to academic research in finance and economics. This is the first paper that interprets conditional portfolio sorts from a machine learning perspective, tailors the methodology to similar approaches well-known in finance, and applies it to a comprehensive financial dataset.

[17] For further reading on decision-trees, see Hastie *et al.* (2009), Zhang and Ma (2012), Murphy (2012) or Criminisi and Shotten (2013).

[18] A few examples in a macroeconomic context use decision trees to analyze currency crises (Kaminsky (2006)), sovereign debt crises (Manasse and Roubini (2009)), banking crises (Duttagupta and Cashin (2011)) or to develop early warning indicators for e.g. excessive credit growth (Alessi and Detken (2014)).

Estimation

We start by describing how variables are selected and how thresholds are estimated. The goal is still to estimate the expectation of the return of firm i in period $t+1$ conditional on information in period t as in equation (1.1).

To illustrate estimation start out with the conditional portfolio sort in figure 1.3. Consider the portfolio S_1 in that figure which is defined by variable $R(g^{(1)}, 1)$ being less than threshold $\tau^{(1)}$ and variable $R(g^{(2a)}, 1)$ being smaller than threshold $\tau^{(2a)}$. Other portfolios can be defined similarly by their relations between sorting variables and associated thresholds. Within each portfolio S_l, the predicted expected return is modeled as the average return, μ_l, of all firms in the portfolio, that is,

$$\hat{\mu}_l = \text{Mean}(r_{i,t+1} | \text{Firm i } \in S_l \text{ in period t}) \tag{1.7}$$

In other words, analogous to linear regression, we are interested in approximating the conditional mean of the outcome variable at a value of the regressor by the average of the outcome variable over observations with close values of the regressors. The conditional portfolio sort therefore generates subsets of firm observations that are more homogenous. Suppose for a moment that we have found such a homogenous allocation of firms into portfolios. The prediction function could then be written as

$$\hat{r}_{i,t+1} = \sum_{l=1}^{L} \hat{\mu}_l \mathbb{1}(\text{Firm i } \in S_l \text{ in period t}), \tag{1.8}$$

giving a portfolio-specific expected return prediction for each observation. What we have described so far is nothing more than a formal definition of the common conditional sorting methodology that we carried out in the previous section.

Of course, the conditional sort does not need to end after two levels but can be computed at greater depth. We consider the case in which the depth of the conditional sort, the sorting variables and associated thresholds are not pre-selected but need to be identified from the data. To start with a negative result, it can be shown that finding the optimal solution to this problem requires solving an optimization problem for which a computationally fast

solution does not exist (see (Hyafil and Rivest (1976))).

Instead, we adopt a *greedy algorithm* from the machine learning literature that proceeds in a step-wise fashion. We describe the details in appendix A.2 and give a high-level summary here. The algorithm starts out with all observations and splits them into two subsets. From a given set of variables, it finds the variable and the associated threshold value that minimize the mean squared error over all observations if predictions are computed as in equation (1.7). The algorithm is called greedy because it solves the minimization problem in a brute-force fashion by trying every combination of variable and threshold value. The same procedure is then repeated in each subset until the number of observations in a subset becomes small or if no further split can meaningfully improve upon the mean squared error. The result is a deep conditional portfolio sort, that is, a conditional portfolio sort with many levels.

Figure 1.4 illustrates the results of this procedure using the data and variables described in section 1.2.1 below. Rather than showing the entire iterative sort, the figure only shows the first few nodes. The first selected split variable is R(0,1), the return over the previous month. The associated threshold is 6, that is, all firms with a return over the previous month in the lowest six deciles are sorted into one portfolio, and the remaining ones are sorted into the other. Conditional on this split, R(0,1) is selected again in the left branch at the next level and R(2,1), the one-month return two months ago, is selected in the right branch. The actual iterative sort goes deeper but, for illustration, we have computed the one-month ahead returns in each of the four subsets. Differences are already pretty stark: The subset S_1 which is the set of companies that were in the lower of the two R(0,1) groups, display the highest return, indicating short-term reversal. The right branch illustrates a momentum effect: Stocks with higher values of R(2,1) have a higher subsequent return on average.

Before we move on, we want to point out a few links to other estimation methods in the literature. The greedy algorithm introduced in this section bears some resemblance to forward-selection methods in regression models. Forward-selection starts out with the smallest possible linear model, estimates bivariate regressions of the outcome variable on each candidate regressor separately, and keeps the one with the highest t-statistic (or

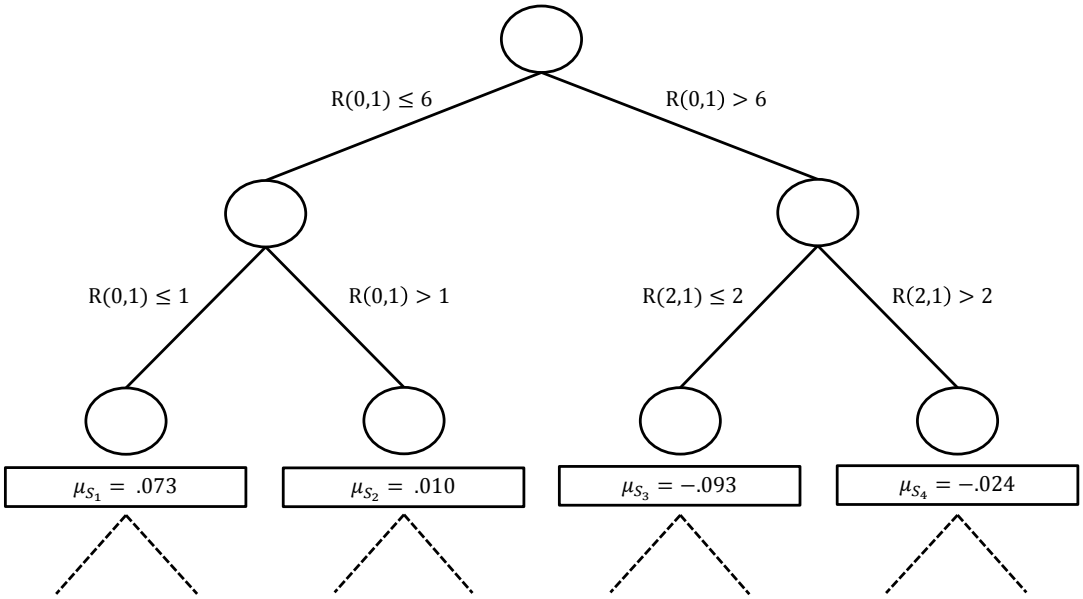


Figure 1.4: *Deep conditional portfolio sort using the entire data set: First nodes*

some other selected performance criterion). The procedure is then repeated for all of the remaining variables with the best-performing variable joining the regression each round until no further variables are significant. As deep conditional sorts, forward-selection works when there are more regressors than observations. On the other hand, forward selection is global in nature in the sense that one regression function is fitted for the entire sample and variable selection is based on performance over the entire sample. In addition, interaction terms would need to be added one-by-one as well leading to a large set of candidate variables whereas the set of candidate variables is always less than the number of main signals in iterative conditional portfolio sorts.

Kernel regression is based on approximating an outcome variable by a (kernel-) weighted average of the outcome at each value of the regressor. Deep conditional portfolio sorts approximate the outcome by the average value of the outcome for a regressor region defined by split points and threshold values. Kernel regressions are very flexible but do not extend easily beyond the bivariate case. A small practical issue is the difficulty to display results in higher dimensions. More importantly, since kernel regression is based on using local

averages, there are few observations in each subspace over which an average is taken as the number of regressors becomes large. This is known as the curse of dimensionality and one can show that the convergence rate for kernel regressions deteriorates sharply with the dimensionality of the regressors. Local linear regressions run into analogous problems in high dimensions.

Model averaging

Constructing deep conditional portfolio sorts in the way described above results in a few challenges. First, as described above, because of the complexity of the optimization, we have to use a greedy algorithm to estimate the model. This algorithm, however, does not guarantee that thresholds and split variables are selected optimally at each node. Second, the threshold rule is discrete, and any error in the estimation of the threshold could greatly distort the correct path for any expected return that is supposed to be predicted from the estimated model. Third, our initial results showed that a single estimated deep conditional portfolio sort summarizes the estimation data well, but the model does not extend well to new observations. In other words, the deep conditional portfolio sort can often overfit the estimation sample. The related machine learning literature acknowledges these issues under the label of *weak learners*, characterized by the fact that their predictions for new observations are often only weakly (albeit positively) correlated with the actual values.

We adopt a solution based on averaging over many deep conditional portfolio sorts that combines elements of Kleinberg (1990, 1996), Breiman (1996), Ho (1998) and Breiman (2001). Kleinberg introduces the idea of "stochastic discrimination" to solve estimation problems without overfitting too much in sample. The idea is to estimate a model a number of times using only random subsets of regressors each time. The resulting models are less prone to overfitting since they are arguably less complex. Kleinberg shows that by combining predictions from such models the accuracy of out-of-sample estimates can be improved upon.[19] Breiman (1996) suggests a related approach, "bootstrap aggregating" (or bagging),

[19]Kleinberg (1990) provides the following intuition: "If one were presented again and again with the same

that leaves the set of regressors intact but estimates a model several times on different random parts of the estimation data. The final prediction is then again constructed as an average over the different models' predictions. Breiman (2001) combines both elements, stochastic discrimination and bagging, in the context of decision trees (which are, what we call, deep conditional portfolio sorts). He finds that this approach that he labels "random forests"greatly improves upon out-of-sample accuracy.[20]

The idea of combining many predictions to construct a more accurate one can be illustrated in a simple voting setup in which people use majority voting to make a decision or to determine the (objective) value of an object. If everyone has the same information set, then nothing can be learned from aggregating individual votes, instead every single vote is a sufficient statistic for the outcome. Only if voters differ in their information, aggregation can lead to a more precise estimate. Stochastic discrimination and bagging induce just such different information sets.

We apply these concepts to deep conditional portfolio sorts. New predicted expected returns are generated by first computing an estimated expected return from each deep conditional sort and then averaging over the individual predictions. More formally, let B be the number of deep conditional sorts that are computed, and let $\hat{f}_b(\Theta_{it})$ be the predicted expected return for stock i at time t that is based on model b. The final expected return

---

poor solution to a problem, he would have little chance of ever creating anything better than that poor solution - on the other hand, if he were presented again and again with equally poor but different solutions to the problem, he would at least be getting diverse information; and in this case, stochastic discrimination will enable him to create from this diverse information an essentially perfect solution."

[20]While applications of these methods are plentiful in computer science, their theoretical properties are not all well-understood. Breiman (2001) shows that bagging decision-trees implies an upper bound on the out-of-sample mean squared error that depends on the strength of the individual models and on the correlation between them. In that sense, bagging shields against overfitting if one can sufficiently de-correlate the individual greedy conditional portfolio sorts. Büchlmann and Yu (2002) analyze the bias-variance trade-off of bagging and they show that bagging reduces mean squared error by substantially reducing variance with only a small effect on bias. They argue that bagging works well for the case of unstable models that are characterized by hard decision rules like splits based on thresholds. Bagging softens these hard decision rules because thresholds vary across models with positive probability. The argument carries over to deep conditional portfolio sorts such that one would expect an ensemble of DCPS to make fewer mistakes than each individual one. Biau *et al.* (2008) provide consistency results of using stochastic discrimination jointly with bagging for decision-trees for the case in which the outcome variable is ordinal (a classification problem). To our knowledge, analog results are not available yet for the case in which the outcome variable is continuous.

estimate is given by

$$\hat{r}_{i,t+1} = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(\Theta_{i,t}). \tag{1.9}$$

In all results that follow, we construct two hundred deep conditional portfolio sorts (that is, $B = 200$) and we use eight out of twenty-five regressors (that is, roughly 30% of the number of regressors) in each of them. We have tried other values for the share of sampled regressors (between 20% and 40%) and also larger values for the number of estimated deep conditional portfolio sorts but have found that results do not vary much with these choices. We settled on the share of 30% of regressors because it is a standard recommendation in the random forest literature, and we chose $B = 200$ because higher values did not have any apparent benefit for the estimation but are more costly in terms of computation.

Discussion and strategies for evaluating the estimations

Our ultimate goal is to provide a new method that is capable of tracing out which firm characteristics predict the cross-section of stock returns well. (Deep) conditional portfolio sorts are potentially interesting because they can account for both the correlation and the interactions of candidate characteristics. Model averaging as described above protects against the risk of in-sample overfitting, and deals with the hard thresholds that sorting induces.

Our suggested approach differs from previous work in a number of ways. First, we do not need to handpick variables in advance; instead, our methodology works well with large sets of many potentially irrelevant variables. Many firm or return characteristics are highly correlated which makes it difficult to judge their contribution when they are considered in isolation. We aim to include many variables and let our algorithm control for the correlation structure between all of them. Second, we can allow for arbitrary interactions between the variables that we include. This is important because, as we have shown in section 1.3.1, these interactions tend to be important. However, the universe of potential interactions is large and can generally not be considered with standard methods.

The flexibility of our approach does not come without costs: Model averaging loses the simple interpretation from a single deep conditional portfolio sort. Moreover, we cannot summarize our model as a simple linear equation in the space of firm characteristics and factors. One reason for the popularity of linear regression methods certainly lies in their apparent transparency. Our approach draws on methods from computer science that are sometimes criticized for producing black box predictions that cannot easily be interpreted. One contribution of this paper is to introduce measures with which the relation between model predictions and regressors can nevertheless be evaluated transparently.

Variable importance Since the relevance of a variable is determined by both its level and its potential interactions with other variables, summarizing statistical significance via a simple t-test is not appropriate. Instead, we rely on a relative variable importance measure that was suggested in Breiman (2001) and that can be interpreted similarly to t-statistics in simple regressions.

For each predictor variable and each deep conditional portfolio sort, we compute the mean squared error (MSE) of the prediction when the values of that variable are randomly permuted, and we express its MSE relative to the model's MSE when all variables are at their original values. This fraction is then averaged over all iterative conditional sorts and predictor variables are ranked by this measure, where higher values imply that random permutations of a predictor variable cause higher increases in mean squared error, and the predictor variable is therefore considered more relevant.

Results are typically displayed relative to the predictor variable that causes the highest increase in mean squared error when it is permuted, a convention that we follow. For example, a value of .8 for a predictor variable means that this variable is associated with an MSE increase equal to 80% of the variable with the highest MSE increase.

Interactions and partial derivatives Another question that one might ask is whether interactions are important in the resulting trees or whether a linear model in the predictor variables would have yielded a similar return forecast. We address this by projecting return

forecasts on the space of predictor variables, that is, we estimate

$$\hat{r}_{i,t+1} = \psi_{cons} + \sum_{g=0}^{24} \psi_g R_{i,t}(g,1) + \epsilon_{it}, i = 1,\ldots,N; t = 1,\ldots,T, \tag{1.10}$$

and we compute the R^2 from this regression. This gives us an answer to the question how much of the variation in forecasts is explained by a simple linear combination of the predictors. In our application below, we find that R^2 is generally low throughout all specifications, illustrating the importance of interaction effects. Then we run the same specification including all two-way interactions of variables to measure the increase in (the adjusted) R^2 which gives us a sense of how important variable interactions and non-linearities are for the return predictions.

To assess directional effects of particular predictor variables on the prediction, we define a measure of partial derivatives that can be applied to deep conditional portfolio sorts. Define $R_{it}(g^-,1)$ as the vector of past return variables that does not include past return g. We approximate a partial derivative of the prediction with respect to past return ranking $R_{it}(g,1)$ as follows. Recall that we construct past return rankings as the cross-sectional decile ranks, that is, $R_{it}(g,1) \in \{1,\ldots,10\}$. For each of the ten values, counterfactually set $R_{it}(g,1) = d, \forall d = 1,\ldots,10$ for all observations and compute the average prediction over firms, time and bootstrap samples,

$$\hat{r}^{g,d}_{i,t+1} = \frac{1}{N}\frac{1}{T}\frac{1}{B}\sum_{i,t,b} \hat{f}_b(R_{it}(g,1); R_{it}(g^-,1)).$$

Repeat this for all values of d, and graph the results for each past return g and each value of d. Our method can easily be extended to varying two (or more) variables at the same time. Below, we also report partial derivatives for two-way interactions of return variables.

Return predictions Finally, we address the question of whether deep conditional portfolio sorts really work in the sense that they make superior return predictions. Based on our model estimates, we predict stock returns for each firm in each month and we sort stocks into deciles each month based on those predictions. We then compute the mean return

spread that is generated across deciles. In addition, we employ a simple trading strategy: Each month, we go long the highest decile of predicted returns and we go short the lowest decile of predicted returns, therefore earning an equal-weighted hedge return.

It is, of course, essential to test the model out-of-sample. While an actual out-of-sample test is difficult to implement, we suggest a standard pseudo-out-of-sample procedure that works as summarized in figure 1.5 and that we also used in section 1.2.3. Deep conditional portfolio sorts are re-estimated each year with data over the past five years. Predicted returns are then calculated for the next twelve months. In each of these months, we trade on our predicted returns as described in the previous paragraph. This approach takes into account the potential time-varying importance of different regressors, and answers whether averaged deep conditional portfolio sorts could, in principle, be used for trading purposes.[21]

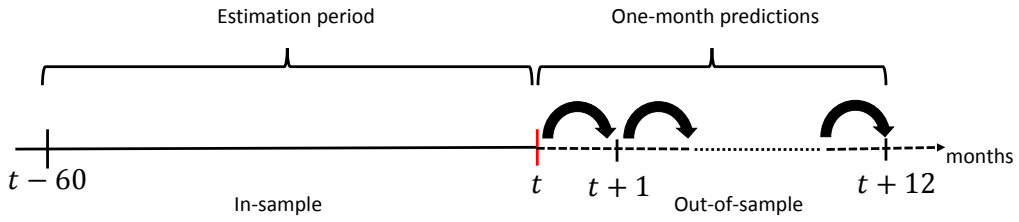


Figure 1.5: *Out-of-sample testing*

Notes: Deep conditional portfolio sorts are re-estimated every year with data over the past sixty months. Predicted returns are then calculated for the next twelve months. The strategy is go long (short) the highest (lowest) decile of those predictions each month.

1.4 Empirics

We apply our method to the prediction of future returns based on past returns. Our main results are as follows. First, deep conditional portfolio sort works well in this setting in the sense that expected return predictions are ordinally accurate. Strategy returns and

[21] An alternative strategy for pseudo-out-of-sample testing is often employed when the data can be assumed to be independently and identically distributed. The model would be estimated once over the entire period with 70% of the data. Predictions would then be computed for the remaining 30% of the data. Even if data were stratified by month, this procedure would not provide proper out-of-sample evidence because returns are cross-sectionally correlated within each period due to common factors.

information ratios based on the model's predictions are much higher than those from alternative models. Second, among return-functions the most important ones refer to the more recent past. Third, superior predictive ability can be traced to flexibly dealing with non-linear relations between past and future returns, and interaction effects between past return functions. The relation between past and future returns is more complex (and more predictable) than can be captured by any one summary return.

1.4.1 Strategy returns

We first show that a strategy that buys high predicted expected returns and that sells low predicted expected returns makes robust and strong risk-adjusted excess returns.We proceed as described in section 1.3.2, that is, we estimate the model with five years of data up to period t, and use the estimated model to predict returns for $t+1, \ldots, t+12$. This procedure is repeated for every year between 1968 and 2012. In both cases, we sort returns into ten deciles from the lowest to the highest predictions each month.

Figure 1.6 shows that the annual strategy return would have been positive for each of the past forty-five years. Returns tend to be somewhat lower after the year 2000 which is consistent with the observation that momentum strategies have not performed well recently (see Lewellen (2013)). Figure 1.7 shows the return to investing \$1 in the long portfolio and the short portfolio and illustrates that the deep conditional portfolio sort works well in both portfolios.

More generally, figure 1.8 illustrates that the deep conditional portfolio sort manages to spread returns more accurately across the entire distribution of firm-months than common past return sorting strategies. It plots the average decile performance for predictions based on the rolling model estimation. The deep sort does consistently better than a simple sorting on a single past return. Although this is not surprising, it is not self-evident that a larger set of explanatory variables will do better in these dimensions. Recall that we evaluate all performances out-of-sample for twelve months by fixing the prediction function based on past estimates. Deep conditional portfolio sorts appear to excel by producing a much more

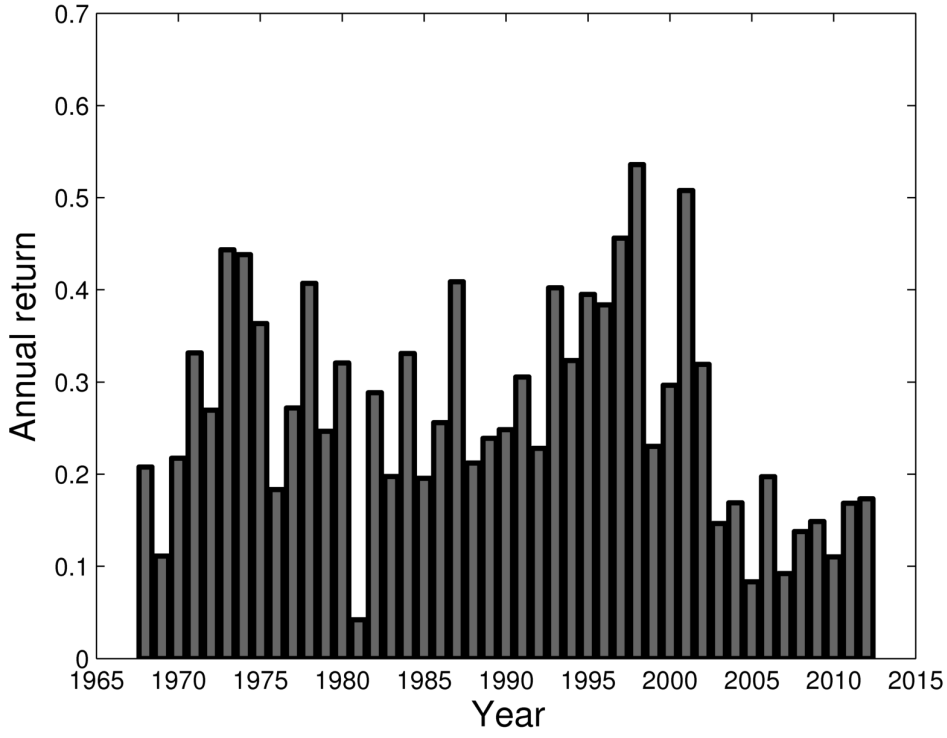


Figure 1.6: *Annual strategy return*

Notes: The strategy is based on the predictions of deep conditional portfolio sorts that relate future returns to past decile sorts of returns. Past return sorts include decile rankings R(g,l) with length l equal to 1 and gap g between 0 and 24 months (i.e. all one-month returns over the two years before portfolio formation). The strategy goes long the highest decile of predictions and goes short the lowest decile of predictions each month. The figure shows the annual return for each of forty-five out-of-sample predictions.

pronounced return spread than simple strategies.

Table 1.4 regresses the return to the long-short strategy on the CAPM, the three-factor model and the four-factor model. The raw average monthly return in column (1) is 2.3 percent. The strategy is significantly positively correlated with the market return with a very low factor loading; however, projecting the strategy return on the market return does not have a strong effect on the average abnormal return. The strategy does not load highly on the size or value factors.

Overall, results for the CAPM and the three-factor model are very similar, with almost no increase in R^2. As is not surprising, time-variation in the strategy return can partially be explained by the momentum factor, but the intercept is still strongly significant and

Table 1.4: *Strategy factor loadings: Deep conditional portfolio sort*

	(1)	(2)	(3)	(4)
Intercept	2.30	2.23	2.25	2.05
	(16.75)	(16.04)	(16.51)	(14.54)
MKT		0.07	0.05	0.09
		(2.14)	(1.53)	(2.78)
SMB			0.08	0.09
			(1.40)	(1.69)
HML			-0.03	0.04
			(-0.39)	(0.61)
UMD				0.20
				(5.57)
R^2		0.02	0.03	0.13
IR		2.90	2.93	2.82
SR	2.96			
N	540	540	540	540

This table shows time-series regressions of strategy returns on factors. Returns are specified in percent per month. Strategies are based on the predictions of a deep conditional portfolio sort that relates future returns to past decile sorts of returns. Past return sorts include decile rankings R(g,l) with length l equal to 1 and gap g between 0 and 24 months (i.e. all one-month returns over the two years before portfolio formation). Predictions are based on the model in section 1.3.2. Strategies go long the highest predicted return decile and go short the lowest predicted return decile. The sample period covers 1968 to 2012, and all results are based on rolling out-of-sample estimates of the models. MKT is the market return, SMB and HML are the Fama-French factors for size and value, and UMD is the momentum factor. SR is the Sharpe ratio and IR is the information ratio. T-statistics are in parentheses, and standard errors were clustered using Newey-West's adjustment for serial correlation.

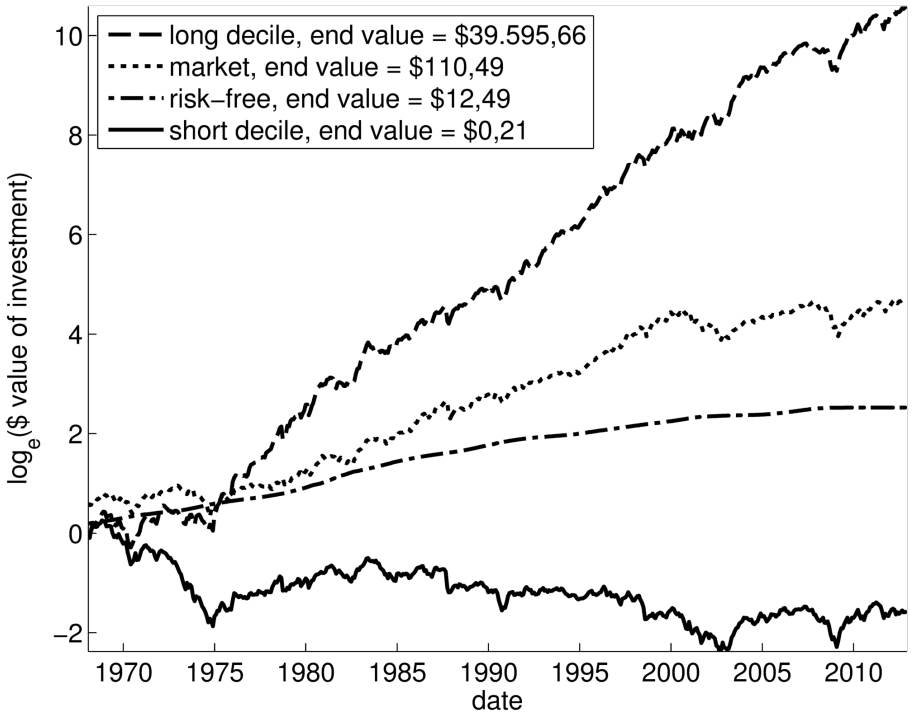


Figure 1.7: *Earned profit from investing $1 in the strategy in 1968*

Notes: The strategy is based on the predictions of a deep conditional portfolio sort that relates future returns to past decile sorts of returns. Past return sorts include decile rankings R(g,l) with length l equal to 1 and gap g between 0 and 24 months (i.e. all one-month returns over the two years before portfolio formation). The strategy goes long the highest decile of predictions and goes short the lowest decile of predictions each month. The figure shows the earned profit from investing $1 in the long and the short portfolio, respectively. For reference, the figure also includes the returns to investing $1 at the riskfree rate and for investing at the rate of the market return over the same horizon.

large with a value of 2 percent per month. The R^2 goes up to .13 which still leaves a large part of the strategy variation unexplained by the equilibrium model. We observe very high information ratios at a value of around 2.9 throughout all specifications. While averaged deep conditional portfolio sorts produce mean excess returns that are somewhat, if not greatly, above those of the standard methods in section 1.2.3, the method seems to do so with a large reduction in variance.

Table 1.5 sheds more light on the decile portfolios that are formed based on the models' predictions. They show the factor loadings of each decile portfolio return for one of four risk models. The returns of all decile portfolios appear to correlate one-to-one with the market

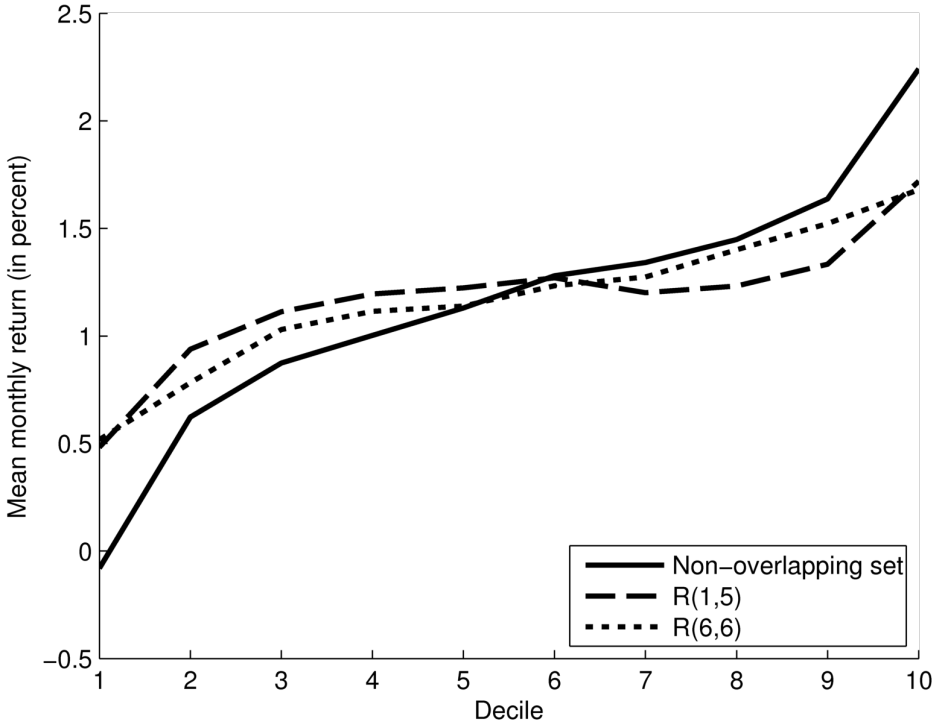


Figure 1.8: *Average monthly decile return for strategy return and simple return strategies*

Notes: The strategy is based on the predictions of a deep conditional portfolio sort that relates future returns to past decile sorts of returns. Past return sorts include decile rankings R(g,l) with length equal to 1 and gaps between 0 and 24 months. The strategy goes long the highest decile of predictions and goes short the lowest decile of predictions each month. Simple return strategies are plotted for comparison. R(1,5) is the strategy that goes long (short) the highest (lowest) decile of returns over the past six months, leaving out the most recent one. R(6,6) is Novy-Marx (2012)'s intermediate return strategy that goes long (short) the highest (lowest) decile of returns that are computed over the six months that skip the most recent six months.

return, with the extreme portfolios experiencing a slightly higher covariance. Second, there is no apparent spread in factor loadings for the size and the value factor. The extreme portfolios load slightly higher on the size factor (an issue that we come back to in appendix A.3), and slightly lower on the value factor. Third, there is a monotone relationship of decile returns with respect to loadings on the momentum factor. Quantitatively, however, these differences are small. Fourth, even though none of these portfolios differ much in their loadings on risk factors, there is a strong monotone relation between the portfolios and their (risk-adjusted) average returns. This stands in stark contrast to the seemingly very similar portfolios in terms of risk loadings. What is more, this relation is not only driven by the

extreme portfolios (although it is particularly strong in those portfolios), but it exists across all ten portfolios. In unreported[22] monotonicity tests based on Patton and Timmermann (2010), we confirm that raw and risk-adjusted returns are monotonically increasing in deciles at all levels of significance.

[22]available on request

Table 1.5: *Factor loadings of decile portfolios: Deep conditional portfolio sort*

	Low	2	3	4	5	6	7	8	9	High	High-Low
Average return	-0.53	0.21	0.44	0.58	0.68	0.80	0.94	1.01	1.22	1.76	2.30
	(-1.74)	(0.77)	(1.66)	(2.29)	(2.61)	(3.10)	(3.52)	(3.78)	(4.27)	(5.54)	(16.75)
						CAPM					
Intercept	-1.52	-0.73	-0.48	-0.32	-0.23	-0.13	0.01	0.06	0.23	0.72	2.23
	(-8.59)	(-4.97)	(-3.60)	(-2.47)	(-1.81)	(-1.03)	(0.08)	(0.42)	(1.55)	(3.97)	(16.04)
MKT	1.12	1.08	1.06	1.04	1.04	1.06	1.06	1.09	1.14	1.20	0.07
	(27.40)	(30.01)	(31.94)	(32.94)	(30.65)	(31.83)	(30.62)	(28.90)	(29.10)	(25.30)	(2.14)
						Three-factor model					
Intercept	-1.64	-0.87	-0.63	-0.48	-0.38	-0.28	-0.13	-0.07	0.10	0.61	2.25
	(-13.12)	(-7.71)	(-6.80)	(-5.58)	(-4.66)	(-3.45)	(-1.42)	(-0.71)	(0.98)	(4.51)	(16.51)
MKT	0.99	0.97	0.97	0.96	0.97	0.98	0.98	0.99	1.02	1.04	0.05
	(27.70)	(29.22)	(34.92)	(36.04)	(38.43)	(38.76)	(32.98)	(29.14)	(32.43)	(27.55)	(1.53)

Continued on next page

Table 1.5: *(continued)*

	Low	2	3	4	5	6	7	8	9	High	High-Low
SMB	0.87	0.74	0.69	0.67	0.65	0.66	0.69	0.71	0.80	0.95	0.08
	(8.64)	(8.49)	(8.76)	(8.51)	(8.80)	(8.69)	(8.51)	(8.84)	(10.21)	(12.08)	(1.40)
HML	0.23	0.25	0.27	0.30	0.29	0.28	0.27	0.25	0.25	0.21	-0.03
	(2.84)	(3.58)	(4.44)	(4.73)	(4.64)	(4.44)	(4.24)	(3.50)	(3.68)	(2.52)	(-0.39)
						Four-factor model					
Intercept	-1.37	-0.67	-0.48	-0.36	-0.29	-0.21	-0.07	-0.02	0.13	0.69	2.05
	(-12.77)	(-7.10)	(-6.32)	(-4.96)	(-4.11)	(-3.10)	(-0.90)	(-0.25)	(1.40)	(5.13)	(14.54)
MKT	0.94	0.94	0.94	0.93	0.95	0.97	0.96	0.98	1.02	1.03	0.09
	(30.12)	(31.11)	(37.89)	(39.00)	(41.63)	(44.66)	(36.30)	(31.58)	(35.46)	(27.00)	(2.78)
SMB	0.86	0.73	0.69	0.67	0.65	0.65	0.69	0.71	0.80	0.95	0.09
	(11.69)	(11.04)	(10.88)	(10.14)	(10.15)	(9.58)	(9.15)	(9.38)	(10.52)	(13.16)	(1.69)
HML	0.15	0.19	0.23	0.26	0.25	0.26	0.25	0.23	0.24	0.18	0.04

Continued on next page

Table 1.5: *(continued)*

	Low	2	3	4	5	6	7	8	9	High	High-Low
	(2.47)	(3.39)	(4.40)	(4.91)	(4.96)	(4.65)	(4.53)	(3.62)	(3.86)	(2.34)	(0.61)
UMD	-0.27	-0.20	-0.15	-0.12	-0.10	-0.07	-0.06	-0.05	-0.03	-0.08	0.20
	(-7.71)	(-6.29)	(-4.87)	(-3.82)	(-2.90)	(-2.26)	(-1.86)	(-1.49)	(-0.93)	(-2.34)	(5.57)

This table shows time-series regressions of decile portfolio returns on factors. Returns are specified in percent per month. Each decile is formed on the predicted returns of a deep conditional portfolio sort that relates future returns to past decile sorts of returns. Past return sorts include decile rankings R(g,l) with length l equal to 1 and gap g between 0 and 24 months. Predictions are based on the model in section 1.3.2. *Low* denotes the lowest decile of predicted returns and *High* denotes the highest decile of predicted returns. The first panel reports the average return, the second panel reports CAPM estimates, the third reports the three-factor model estimates and the fourth panel adds momentum. MKT is the market return, SMB and HML are the Fama-French factors for size and value, and UMD is the momentum factor. The sample period covers 1968 to 2012. T-statistics are in parentheses, and standard errors were clustered using Newey-West's adjustment for serial correlation.

Deep conditional portfolio sorts appear to work well in our application in the sense that they produce high and stable excess returns out of sample that are not explained by standard factor models. This begs the question what the method finds that researchers have not paid attention to. We discuss the discovered structure of predictor variables next.

1.4.2 Exploring the mechanism

Predictor variable importance

Recall that we re-estimate the model each year for a total of 45 different estimated models over time. When we can compute our measure of predictor variable importance for each year, this gives us a ranking of the importance of each variable in each year. As a first summary, we rank past returns by their median rank in these 45 models. Table 1.6 shows the median rank as well as the upper and lower quartile of ranks for each of the top ten past returns.

The top four return functions are related to the most recent six months of returns; all return functions over the most recent six months enter the top ten. In addition, some returns that show up provide information about the intermediate return between six and twelve months before the formation date. In particular, it is interesting and reassuring to see past return functions considered in the preceding literature to rank highly in the list. R(0,1), the return over the most previous month is the return function of Jegadeesh (1990) and many other papers, while R(11,1), the one-month return exactly twelve months ago, is the seasonal effect documented by Heston and Sadka (2008).

There is also considerable time variation in the exact ranks as illustrated by the interquartile range of ranks for each past return. All of them were in the top half for more than fifty percent of the time, and seven out of the ten return functions are in the top ten for at least half of the years. On the other hand, each variable also had periods during which it appears less relevant to the prediction as expressed in the last column of the table. Overall (unreported) we find that the rank correlation (Spearman) of past returns' importance between subsequent years is around .7, which points to the fact that the structure is relatively stable.

Table 1.6: *Most important past return variables: Rank statistics*

	Median	75th percentile	25th percentile
R(0,1)	1	1	1
R(1,1)	4	2	15
R(2,1)	4	3	8
R(3,1)	6	4	12
R(11,1)	7	3	9
R(4,1)	8	6	11
R(5,1)	8	5	14
R(8,1)	11	7	18
R(10,1)	11	7	16
R(9,1)	12	8	15

This table shows the ten most important past returns (by median rank) in the deep conditional portfolio sort that relates future returns to past decile sorts of returns. Past return sorts include decile rankings R(g,l) with length l equal to 1 and gap g between 0 and 24 months (i.e. all one-month returns over the two years before portfolio formation). The model is estimated each year between 1968 and 2012 for a total of 45 different rankings. The table reports the median, and the upper and the lower quartile for the top ten past returns (by median rank) over the 45 estimations.

The fact that the pattern of more recent returns being more relevant than more distant past returns comes out of an agnostic search procedure is intriguing. We find that it is quite a robust fact in the data throughout various specifications. For instance, we find very similar results for past-return based variables when we include other firm characteristics in the estimation (appendix A.3.1). In appendix A.3.2 we consider an expanded set of predictor variables that uses 126 past return functions of different gaps and different lengths such that standard past return functions like R(0,6) (the return over the most recent six months) are also part of the set of regressors. In that exercise, all ten predictor variables are related to the most recent six months of returns and, what is more, the top six return functions are returns of length one that, taken together, summarize the most recent six month return. The fact that a standard return like R(0,6) is not chosen but its components are, illustrates that using the return over the previous year alone (and not the one-month returns that it is based on) leads to a loss of relevant information. One-month returns contain important information that is neglected when summary returns such as R(0,6) or R(1,11) are considered. For both sets of past-return functions we repeat the estimations by firm size in appendix A.3.3 and again find similar results.

Our first intermediate result is, thus, that deep conditional portfolio sorts work because they effectively exploit variation in relatively recent one-month returns. The next sections look at how these variables are combined.

Average partial derivatives

Next, we consider the measure of an approximated partial derivative that we introduced in section 1.3.2. For each one-month return ranking over the previous half year, for all observations, we vary its value from the lowest one (1) to the highest one (10), and compute the counterfactual predictions. This allows us to trace out whether a variable is monotonically related to returns and to evaluate the sign of the average derivative going from the lowest to the highest value of the predictor variable. We focus on the most recent half year before portfolio formation because our results so far suggest that these returns are most important

for return prediction.

Figure 1.9 shows results. Focus on the first row for now (we will get to the second row in section A.3.1) which correspond to the deep conditional portfolio sort that we have considered so far. Each column shows results for one of the most recent past one-month returns. Each panel varies the respective predictor from low to high and averages the prediction for each of ten values. We observe that short-term reversal, the most recent one month return, is negatively related to the return predictions, that is, higher values of the most recent one-month return predict lower returns. For the next return function, R(1,1), the one-month return over the second-to-last month, both high and low values are associated with lower returns. The next return functions are monotonically related to predictions, but in a non-linear way: Low realizations have a large negative effect on the prediction, but high realizations do not have as much of a positive effect. These returns, thus, help to identify stocks with low expected returns but they do not necessarily help much to identify stocks with high expected returns. It is only when we consider one-month returns that are in the more distant past (more than four months out) that we find a standard momentum effect, that is, a monotonically positive and close to linear relation between past and predicted returns.

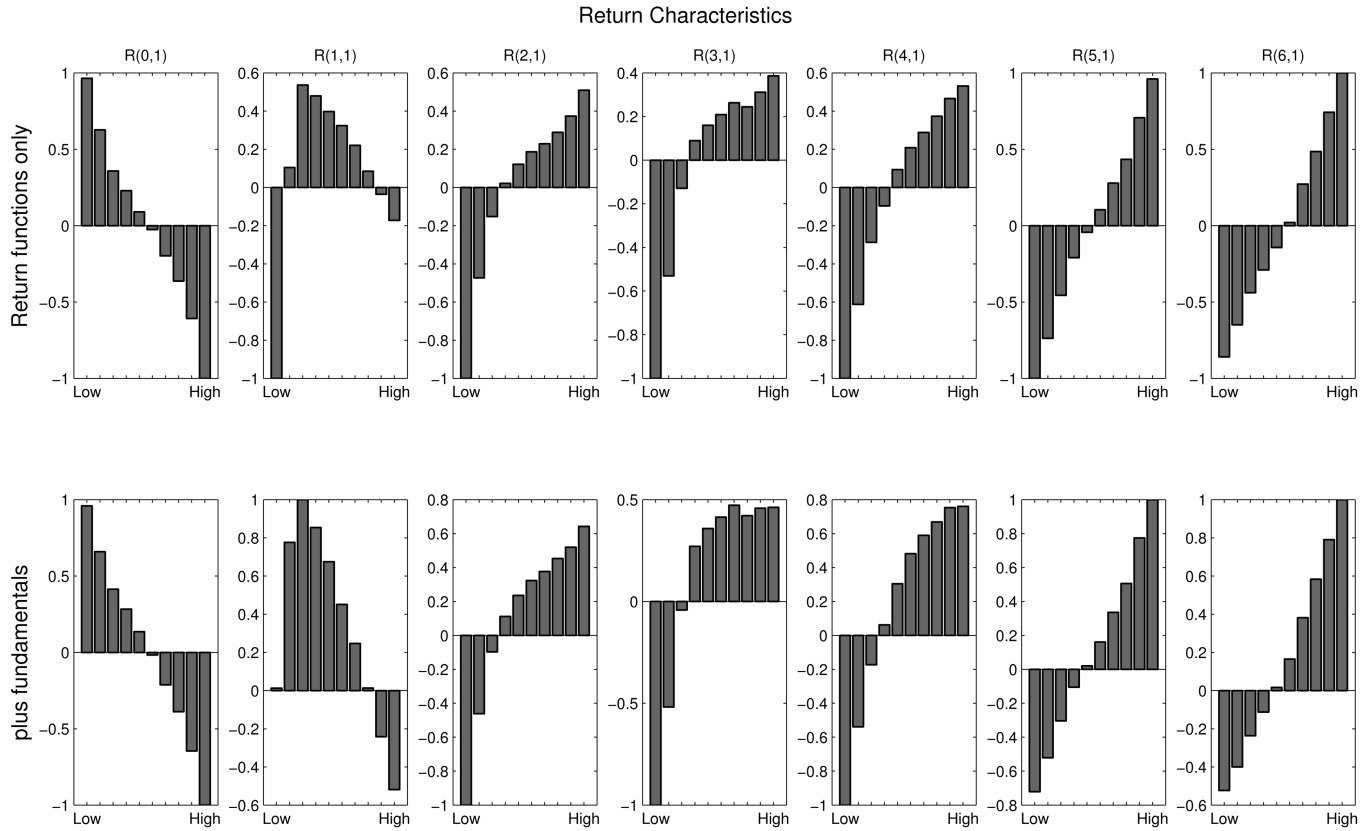


Figure 1.9: *Average partial derivatives for return characteristics*

Notes: The figure shows the average prediction when a characteristic is counterfactually varied from low to high values. Details are in section 1.3.2. The first row shows results when we use only twenty-five past one-month returns as predictors. The second row shows results for the same one-month return functions when other firm characteristics (defined in appendix A.3.1) are included in the estimations as additional variables.

The literature has not paid much attention to non-linear relations between past and future returns. Given that a. predictions from our deep conditional portfolio sorts make high risk-adjusted excess returns, b. short-term return functions have high values in our predictor variable importance calculations and c. the partial effects of these variables cannot all be linearly related to returns, it appears, however, that non-linearities should be investigated further in future research.

Figure 1.10 shows contour plots for all two-way interactions of the most recent one-month return functions. In each panel, darker areas represent lower return predictions and brighter areas represent higher return predictions. A couple of interesting results stand out: First, many return variables interact in non-linear ways. For example, the upper left panel shows the interaction of R(0,1), the most recent one-month return, and R(1,1), the return over the preceding month. Return predictions generally decrease in the value of R(0,1), reflecting short-term reversal. However, within high values of R(0,1), return predictions *increase* in R(1,1), while they *decrease* in R(1,1) within low values of R(0,1). This type of non-linearity holds, with some varying extent, in many panels involving R(0,1). Second, for some return variables, we find monotonically increasing predictions within both return variables, mostly for those that involve returns from four or more months ago. Third, some return predictions neither decrease nor increase monotonically in the predictor variable range, but are non-linearly related to return predictions, once one variable is fixed. For instance, from figure 1.9 we know that R(1,1) is non-linearly related to returns. In figure 1.10, we see that this non-linearity is more pronounced when R(1,1) is interacted with intermediate returns like R(3,1) or R(4,1).

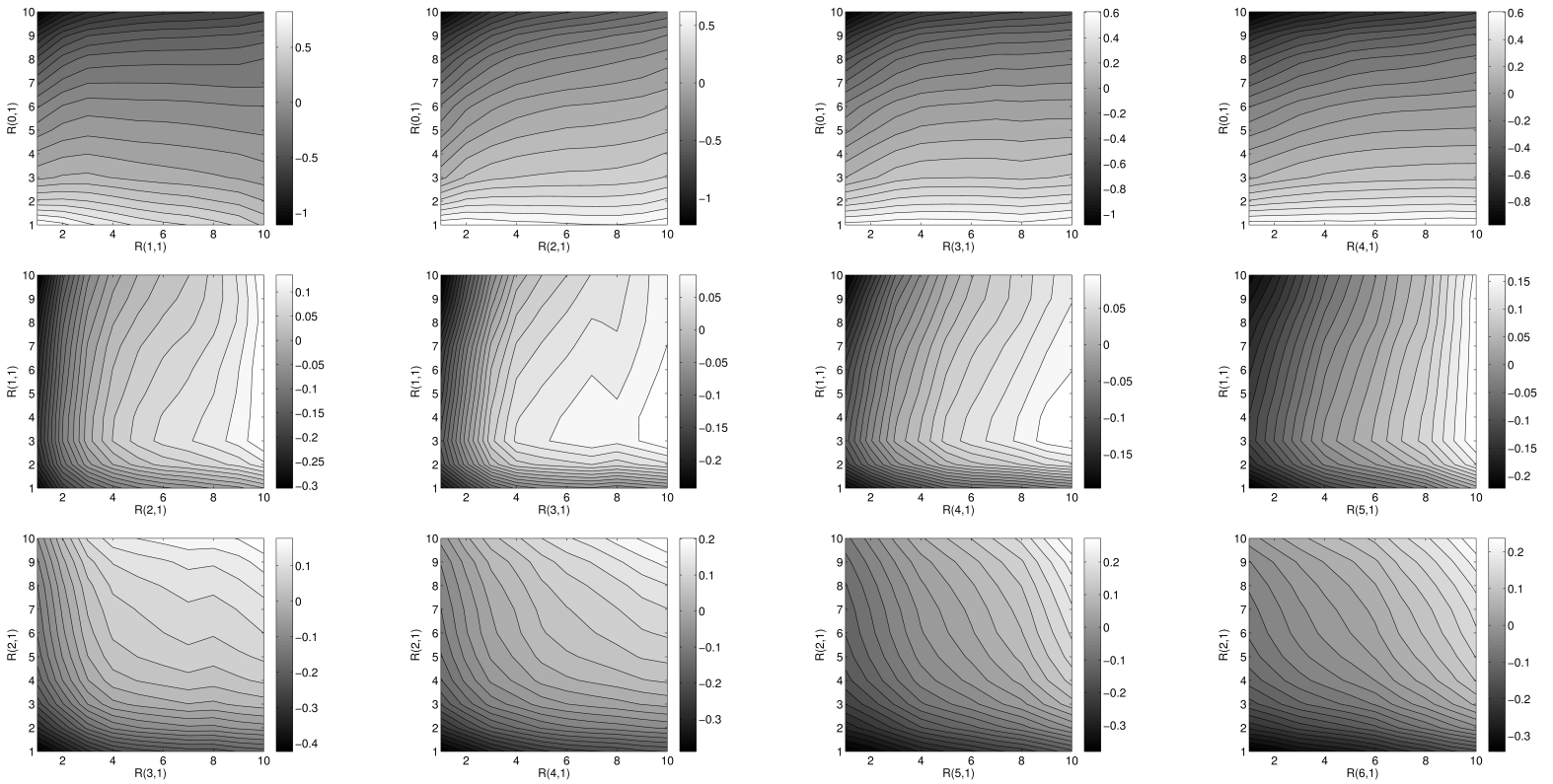


Figure 1.10: *Average double partial derivatives*

Notes: The figure shows the average prediction when two characteristics are counterfactually varied from low to high values. Results are based on rolling optimization of the model and predictions are averaged over the sample period. Details are in section 1.3.2.

Finally, we find evidence that the estimate average partial derivates are time-varying. Figure 1.11 and 1.12 illustrate this for two different variables. Figure 1.11 shows average partial derivatives in eight different years, evenly spaced over the sample period, for R(0,1), the return over the previous month. Short-term reversal is detectable across all years, but its strength varies over time. While our model estimates indicate relatively monotone (or regular) short-term reversal across all ten deciles for the first half of the sample, short-term reversal is more apparent in the extreme deciles in the second half of the sample. Similar conclusions can be drawn from figure 1.12 which shows the same calculations for R(5,1), the one-month return six months before portfolio formation. In the first half of the sample, momentum is apparent and robust across all deciles. In the second half, however, differences in average partial derivatives are more pronounced between extreme deciles than between intermediate deciles. Interestingly, recently (in 2012, the lower right panel), the average partial derivative of R(5,1) has reversed such that lower values of R(5,1) are associated with higher returns in the model estimates. Recall that the estimation period for this panel is 2006-2011 which coincides with an episode of a momentum crash as documented by Daniel and Moskowitz (2014). As we have shown in the previous section, a trading strategy based on our model estimates has not suffered the strong crash that a standard momentum strategy has experienced in this period. The average partial derivative at that time indicates that the model has picked up the weakness of standard momentum and that the estimated relationship was adjusted (in that case: reversed) accordingly.

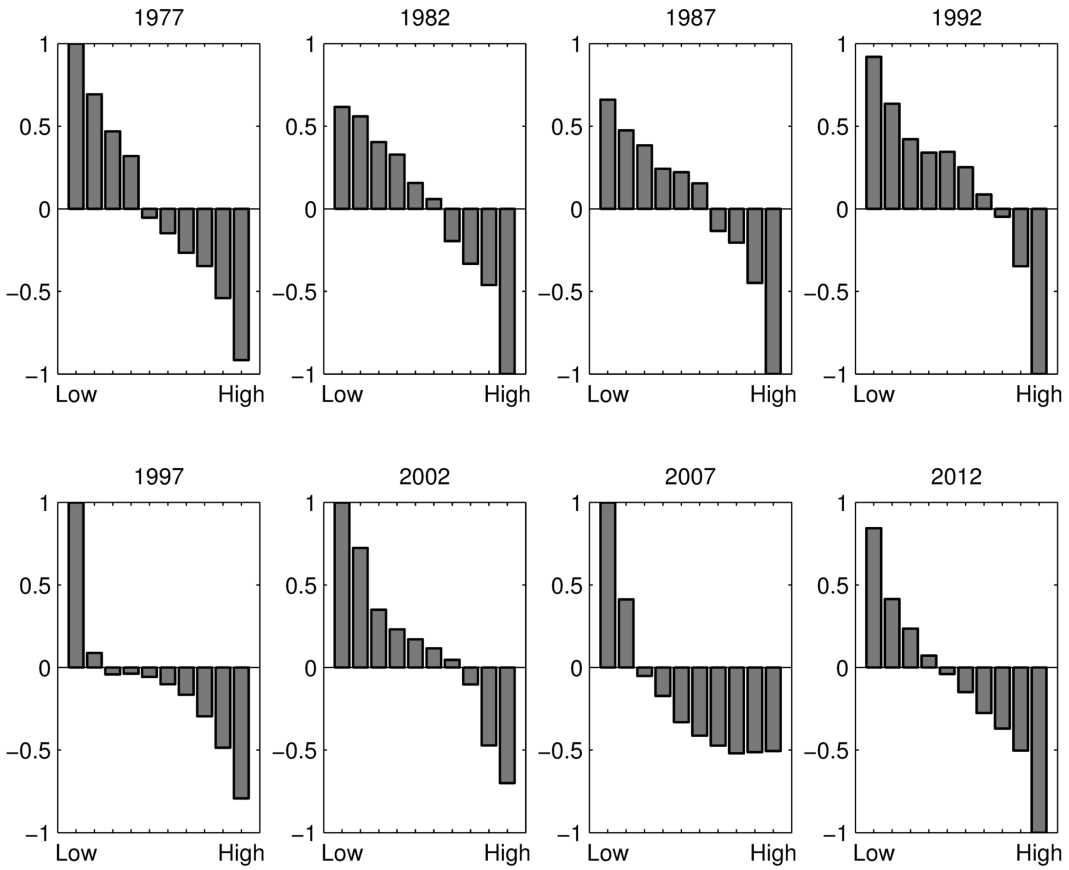


Figure 1.11: *Average partial derivatives in different years*

Notes: The figure shows the average prediction when R(0,1), the return over the previous month, is counterfactually varied from low to high values, and results are displayed for different years to illustrate time-variation. Results are based on rolling optimization of the model, details can be found in section 1.3.2.

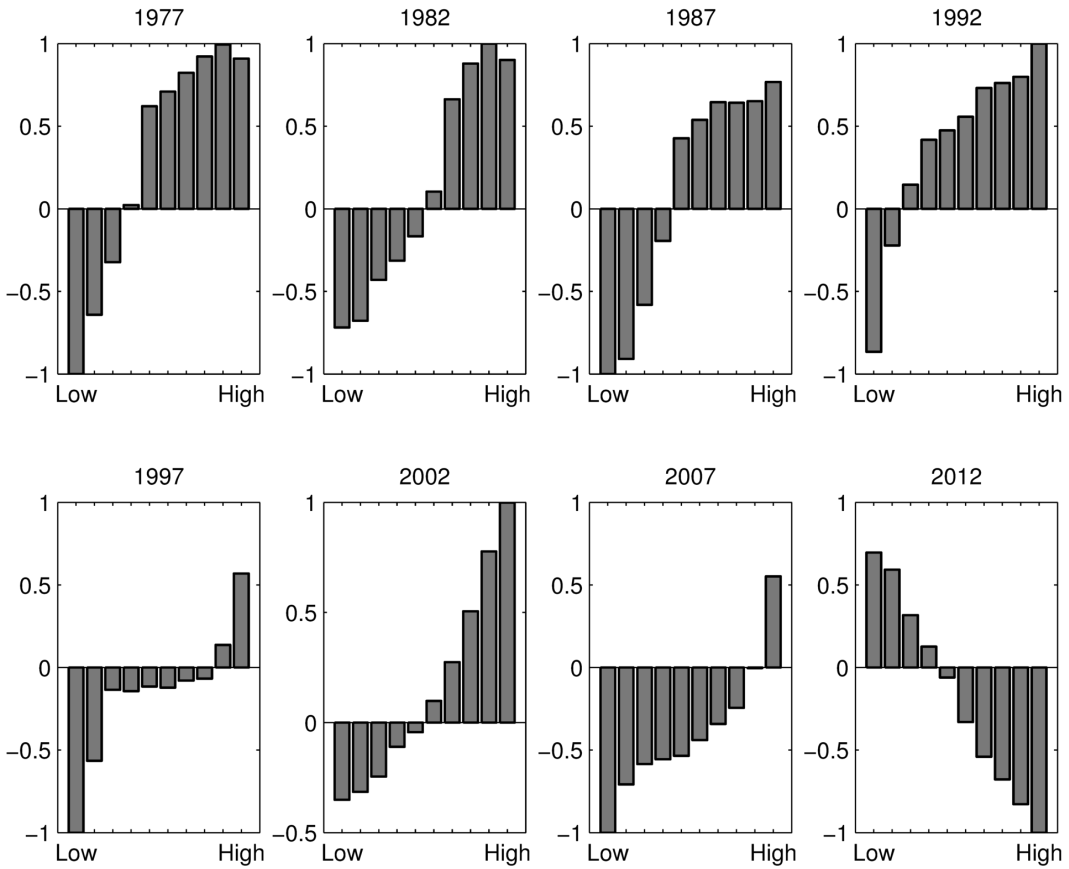


Figure 1.12: *Average partial derivatives in different years*

Notes: The figure shows the average prediction when R(5,1), the one-month return six months before portfolio formation, is counterfactually varied from low to high values, and results are displayed for different years to illustrate time-variation. Results are based on rolling optimization of the model, details can be found in section 1.3.2.

An extensive discussion of the time-variation of all predictor variables is beyond the scope of the paper. The examples above, however, serve to illustrate its importance. The question of whether time-variation in past-return signals can be related to, e.g., the macroeconomic or financial cycle is left for future research.

Do interactions drive the predictions?

We answer this question with the simple exercise outlined in equation (1.10) in section 1.3.2. We regress the deep conditional portfolio sort's predicted expected returns on a linear combination of the regressors that enter the model. The first row of table 1.7 shows that this regression has an R^2 of 1%, that is, only a small portion of the variation in predictions can be explained by a model that is linear in the regressors. The second row adds all two-way interactions between regressors, resulting in a striking ten-fold increase in R^2 (and the adjusted R^2). Regressor interactions can therefore explain a much larger portion of the deep conditional sort's variation in predicted expected returns. While two-way interactions help to explain the predictions, there is still a large part of the variations in predictions that remains unexplained and that should be attributed to higher-order terms.

1.5 Further results

In this section, we investigate a couple of related questions. Section 1.5.1 re-estimates the model when we use only recent past returns and compares the results to a model that uses only intermediate past returns (seven to twelve months in the past) and therefore contributes to the debate about the relative merits of standard momentum and intermediate momentum for cross-sectional return variation started by Novy-Marx (2012). In section 1.5.2, we compare our results to those of a Fama-MacBeth regression that uses recent past returns and two-way interactions. Section 1.5.3 makes sure that the strategy's estimated excess return does not disappear after taking transaction costs into account. Section 1.5.4 addresses the issue of whether the discovered structure should be given a characteristics or risk factor interpretation.

Table 1.7: *Measures of fit: Regressing the predictions onto linear combinations of the predictor variables*

Other characteristics?	Interactions?	R^2	R^2_{adj}
No	No	0.01	.009
No	Yes	0.11	.097
Yes	No	0.017	.015
Yes	Yes	0.102	.055

This table show measures of fit for regressing the predicted returns from the deep conditional portfolio sort on the predictor variables linearly with and without interaction terms. The set of predictor variables contains twenty-five one-month returns over the previous two years, firm fundamentals are a set of 86 firm characteristics as described in appendix A.3.1. Results in the first row are based on the regression

$$\hat{r}_{i,t+1} = \psi_{cons} + \sum_{g=0}^{24} \psi_g R_{i,t}(g,1) + \epsilon_{it},$$

and results in other rows are based on likewise regressions that include other firm characteristics and/or two-way interactions between the regressors. Regressions are fitted for each year of predictions separately and the mean measure of fit over time is reported.

1.5.1 Medium-term momentum

Our results suggest that the most important predictor variables are related to the most recent six months before portfolio formation. One could therefore suspect that short-horizon returns are generally better predictors of future returns than intermediate horizon returns.

We address this question by defining two more sets of return-based functions that split the regressors into those that provide information about the most recent six months and into those that provide information about returns seven to twelve months before portfolio formation. Formally, our split is based on the sum of the gap and length parameters. One set includes all return-based functions for which the sum of gap and length is smaller than 7 months (we call this the *short-term set*), and our second set includes all return-based functions for which the sum of gap and length is between 7 and 12 months (the *intermediate-term set*). The latter set of functions includes the function suggested by Novy-Marx (2012) and other functions that are correlated with it.

Table 1.8 provides the factor loadings of the equal-weighted hedge return strategy that goes long the highest predicted decile and that goes short the lowest predicted decile based on the predictions derived from each set of predictor variables. The first five columns report loadings for the strategy based on the short-term set and the remaining columns report loadings for the strategy based on the intermediate-term set. There are a couple of intriguing results. First, we see that both strategies make high and robust excess returns relative to the CAPM, and the three- and four-factor models. Second, as is immediately apparent, alpha is lower for the medium-term strategy than for the short-term strategy throughout all specifications. In columns (5), we add the strategy return of the intermediate-term set to the factors. As indicated by the t-statistic on the coefficient and the increase in R^2, the two strategies are correlated, and the excess return of the short-term strategy decreases to 1 percent per month.

In column (10), we do the same, and add the short-term strategy return to the factor regression for the intermediate-strategy return. Interestingly, alpha disappears almost entirely once the short-term strategy return is accounted for.

We interpret this as evidence that the most important variation for return prediction purposes stems from short-term variation in returns rather than intermediate-term variation once interactions and confounding returns are included in the estimation. This reconciles the result in Novy-Marx (2012) with Goyal and Wahal (2013) who cannot find the intermediate-term momentum effect in 37 out of 38 markets.

1.5.2 Fama-MacBeth with recent returns only

In this section, we briefly contrast the results from the deep conditional portfolio sorts to the Fama-MacBeth results in section 1.2.3. Deep conditional portfolio sorts can be viewed as either a kitchen-sink regression or as a variable selection method (since a variable is selected for each split). An initial interesting comparison can thus be conducted between the performance of the deep conditional sort in table 1.5 and the Fama-MacBeth regressions in table 1.2. The raw and factor adjusted returns are about .5 percentage points higher than in the Fama-MacBeth regressions and, more interestingly, the information ratios are generally roughly three times as high. Even if we include all two-way interactions in a Fama-MacBeth regression as in table 1.3, average excess returns and information ratios are generally much lower than in our results for the deep conditional sort.

These results let deep conditional sorts shine in two dimensions. If regarded as a kitchen sink method, the greedy conditional sort leads to better performance than the Fama-MacBeth analogue, although both perform well. If regarded as a variable selection device, the Fama-MacBeth method mostly recovers momentum as an important determinant of expected returns whereas the structure discovered by the greedy conditional sort is more stable and cannot be explained by (simple) factor models.

In table 1.9, we contrast this to the case in which only variables that are considered important based on our deep conditional sorts are included in the Fama-MacBeth estimations. In particular, as a consistent set, we focus on the six most recent months of past returns, since our results above indicate that the most recent returns are most important for estimating expected returns. We abstract from variable selection and therefore act as if variable selection

Table 1.8: *Strategy factor loadings: Short-term and intermediate-term return functions*

	Dependent variable									
	Return of short-term strategy					Return of intermediate-term strategy				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Intercept	2.15	2.12	2.17	2.04	1.05	1.78	1.74	1.78	1.44	0.16
	(19.09)	(18.47)	(19.35)	(17.37)	(9.11)	(15.01)	(14.90)	(15.52)	(13.33)	(1.65)
MKT		0.03	0.01	0.03	-0.04		0.05	0.04	0.10	0.08
		(1.38)	(0.46)	(1.64)	(-2.37)		(1.40)	(1.32)	(4.25)	(4.39)
SMB			0.04	0.04	0.06			-0.05	-0.04	-0.06
			(0.65)	(0.85)	(1.95)			(-0.74)	(-0.93)	(-2.27)
HML			-0.08	-0.04	-0.06			-0.08	0.03	0.05
			(-1.52)	(-0.85)	(-1.95)			(-1.15)	(0.78)	(1.85)
UMD				0.13	-0.11				0.35	0.27
				(3.80)	(-4.28)				(11.11)	(10.71)
MT/ST strategy					0.69					0.63
					(15.63)					(15.24)
R^2		0.00	0.02	0.08	0.48		0.01	0.02	0.37	0.64
IR		3.38	3.46	3.37	2.29		2.40	2.47	2.49	0.37
SR	3.42					2.45				
N	540	540	540	540	540	540	540	540	540	540

This table shows time-series regressions of strategy returns on factors. Returns are specified in percent per month. Strategies are based on the predictions of a deep conditional portfolio sort that relates future returns to past decile sorts of returns. Strategies go long the highest predicted return decile and go short the lowest predicted return decile. The short-term strategy is based on predictions from a deep conditional portfolio sort that only uses the most recent six months of past return rankings, while the intermediate-term strategy is based on predictions that use past return rankings from seven to twelve months before portfolio formation. Predictions are based on the model in section 1.3.2. The row "MT/ST strategy" adds the intermediate-term strategy return to the factor regressions for the short-term strategy, and adds the short-term strategy return when the intermediate-term strategy is the dependent variable. The sample period covers 1968 to 2012, and all results are based on rolling out-of-sample estimates of the models. MKT is the market return, SMB and HML are the Fama-French factors for size and value, and UMD is the momentum factor. SR is the Sharpe ratio and IR is the information ratio. T-statistics are in parentheses, and standard errors were clustered using Newey-West's adjustment for serial correlation.

had already been conducted based on the deep conditional sort's results.

Table 1.9: *Strategy factor loadings: Fama-MacBeth predictions using the six most recent one-month returns*

	Levels only				plus relevant two-way interactions				plus all two-way interactions			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Intercept	1.27	1.19	1.23	0.81	1.61	1.47	1.60	1.38	1.58	1.45	1.57	1.35
	(6.57)	(6.04)	(7.21)	(4.22)	(9.39)	(8.73)	(9.67)	(7.28)	(9.29)	(8.60)	(9.44)	(7.08)
MKT		0.09	0.09	0.17		0.16	0.09	0.13		0.14	0.08	0.13
		(1.36)	(1.30)	(3.10)		(3.23)	(1.91)	(3.14)		(3.11)	(1.92)	(3.18)
SMB			-0.03	-0.03			0.10	0.10			0.09	0.09
			(-0.35)	(-0.38)			(1.10)	(1.38)			(1.01)	(1.27)
HML			-0.07	0.06			-0.23	-0.16			-0.21	-0.14
			(-0.48)	(0.61)			(-2.33)	(-2.12)			(-2.21)	(-1.91)
UMD				0.41				0.22				0.22
				(4.13)				(2.49)				(2.59)
R^2		0.01	0.01	0.22		0.04	0.09	0.17		0.04	0.08	0.16
IR		1.02	1.06	0.78		1.48	1.65	1.49		1.49	1.65	1.48
SR	1.09				1.59				1.60			

Continued on next page

Table 1.9: *(continued)*

	Levels only				plus relevant two-way interactions				plus all two-way interactions			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
N	540	540	540	540	540	540	540	540	540	540	540	540

This table shows time-series regressions of strategy returns on factors. Returns are specified in percent per month. Strategies are based on the predictions of a Fama-MacBeth regressions of future returns on past decile sorts of returns. Past return sorts include decile rankings R(g,l) with length equal to 1 and gaps between 0 and 6 months, that is, predictions are based on the equation

$$r_{i,t+1} = \beta^t_{cons} + \sum_{g=0}^{6} \beta^t_g R_{it}(g,1) + \epsilon_{it},$$

or, when two-way interactions are included,

$$r_{i,t+1} = \beta^t_{cons} + \sum_{g=0}^{24} \beta^t_g R_{i,t}(g,1) + \sum_{g=0}^{24} \sum_{j>g} \gamma^t_{gj} R_{i,t}(g,1) R_{i,t}(j,1) + \epsilon_{i,t}.$$

The first four columns include only the levels of past returns, the next four columns include relevant two-way interactions as identified from the deep conditional portfolio sort and the last four columns include all two-way interactions between those returns. Strategies go long the highest predicted return decile and go short the lowest predicted return decile. The sample period covers 1968 to 2012, and all results are based on rolling out-of-sample estimates of the models. MKT is the market return, SMB and HML are the Fama-French factors for size and value, and UMD is the momentum factor. SR is the Sharpe ratio and IR is the information ratio. T-statistics are in parentheses, and standard errors were clustered using Newey-West's adjustment for serial correlation.

The first four columns of table 1.9 use only the return functions themselves and no interactions. The long-short strategy has an average return of 1.27 percent per month, and a four-factor alpha of .81 percent per month, with an information ratio of .78. Results based on using all past return functions are slightly higher, indicating that there is some information to be gained by including more distant past returns as well.

The next four columns of table 1.9 additionally include the most relevant two-way interactions between the six most recent return functions based on our previous results. The average strategy return is 1.61 percent per month and the four-factor alpha is 1.38 percent per month, both higher than in the kitchen sink regression above and also than in the estimations without interactions. Most remarkably, we observe an about 50% increase in the information ratio relative to the kitchen sink regression, from 1 to 1.49, indicating that the strategy return is earned at a much better risk-return trade-off.

The last four columns of the table use all (and not only relevant) two-way interactions between the six most recent return functions. Results are almost identical to including the relevant interactions only. Table 1.10 shows that this can be attributed to the fact that the non-relevant interactions have small and insignificant coefficients and therefore do not have a big impact on the predictions.

Note that the returns in table 1.9 are still lower than the strategy returns in the original deep conditional portfolio sort. The Fama-MacBeth regressions only include two-way interactions, and recall from section 1.4.2 that two-way interactions explain only around 10% of the variance of the estimated expected returns of the deep conditional sort. While the Fama-MacBeth regression with two-way interactions goes some way to achieve similar-sized returns, the remaining differences can be attributed to the actual return structure being more involved than can be captured by including levels and two-way interactions of past returns alone.

Table 1.10: *Fama-MacBeth regression coefficients and t-statistics: Using the six most recent one-month returns*

	Coefficients			t-stats		
	Levels only	Relevant Int	All Int	Levels only	Relevant Int	All Int
R(0, 1)	-1.63	-6.29	-6.32	-9.20	-16.30	-16.52
R(1, 1)	0.06	-2.87	-2.82	0.43	-8.83	-8.69
R(2, 1)	0.64	-0.93	-1.48	4.76	-4.65	-4.80
R(3, 1)	0.42	-0.87	-1.06	3.44	-4.25	-3.67
R(4, 1)	0.43	-0.42	-0.47	3.26	-1.99	-1.66
R(5, 1)	0.62	-0.11	-0.20	5.19	-0.60	-0.72
R(6, 1)	0.39	-0.33	-0.70	3.51	-1.80	-2.50
R(0, 1) X R(1, 1)		2.33	2.33		12.45	12.42
R(0, 1) X R(2, 1)		2.19	2.16		11.84	11.81
R(0, 1) X R(3, 1)		1.55	1.54		8.83	8.88
R(0, 1) X R(4, 1)		0.88	0.87		5.10	5.11
R(0, 1) X R(5, 1)		0.82	0.86		5.20	5.48
R(0, 1) X R(6, 1)		0.77	0.79		4.85	5.00
R(1, 1) X R(2, 1)		0.60	0.53		3.64	3.21
R(1, 1) X R(3, 1)		0.73	0.69		4.40	4.18
R(1, 1) X R(4, 1)		0.61	0.60		3.73	3.66
R(1, 1) X R(5, 1)		0.48	0.49		2.95	3.04
R(1, 1) X R(6, 1)		0.54	0.55		3.25	3.27

Continued on next page

Table 1.10: *(continued)*

	Coefficients			t-stats		
	Levels only	Relevant Int	All Int	Levels only	Relevant Int	All Int
R(2, 1) X R(3, 1)			0.27			1.59
R(2, 1) X R(4, 1)			0.29			1.82
R(2, 1) X R(5, 1)			0.36			2.24
R(2, 1) X R(6, 1)			0.20			1.20
R(3, 1) X R(4, 1)			-0.01			-0.08
R(3, 1) X R(5, 1)			-0.10			-0.66
R(3, 1) X R(6, 1)			0.22			1.41
R(4, 1) X R(5, 1)			-0.26			-1.61
R(4, 1) X R(6, 1)			0.10			0.65
R(5, 1) X R(6, 1)			0.08			0.49

This table shows coefficient estimates and t-statistics for the three regression models in table 1.9. Past returns include return-based functions R(g,l) with length equal to 1 and gaps between 0 and 6 months. The sample period covers 1968 to 2012. "Levels only" only includes the levels of past return functions, "Relevant Int" includes relevant two-way interaction terms and "All Int" includes all two-way interaction terms between the six most recent past returns. The first three columns show the coefficient estimates times 100, and the last three columns show t-statistics.

Table 1.10 shows the Fama-MacBeth coefficient estimates averaged over the entire sample and corresponding t-statistics for the three regression models in table 1.9.[23] The second column shows coefficients in the levels-only regression. We observe the short-term reversal effect while all other past return variables enter with a positive sign. This is in line with the

[23]Note that for the prediction exercise we based predictions on rolling estimates of past coefficients as described above, while table 1.10 gives an average over the entire sample period.

standard reversal and momentum effects in the literature.

Column three illustrates how these results completely flip when interaction terms are introduced in the regression. All level effects are on average negatively associated with expected returns while interaction terms are positive. This result is robust to including further (less relevant) interaction terms in column four. A possible interpretation of this finding is that momentum is more likely to exist when returns are more consistent. For instance, we find that the effect of high returns in either the last month or in the second-to-last month indicate low returns. When both returns are high, however, the interaction effect of this consistently high return works against the reversal effect of the two individual returns. Return consistency effects in momentum have been documented before by, among others, Watkins (2003) and Grinblatt and Moskowitz (2004).

How do the estimated Fama-MacBeth interactions compare to the average double partial derivatives in figure 1.10?[24] We find both similarities and differences. When we calculate the same average derivatives for the Fama-MacBeth model, we find that interactions of returns display the aforementioned consistency effect, that is, consistently high past returns predict high returns. These patterns coincide with the ones in figure 1.10. We also see that in two-way interactions that involve R(0,1), returns are less sensitive to the more distant returns, as in the top row of figure 1.10. On the other hand, in the Fama-MacBeth results, the interactions sometimes overturn the reversal effect, unlike in the deep conditional portfolio sort. Owing to their simplicity, the Fama-MacBeth regressions do not capture the more involved interaction patterns between R(1,1) and more distant returns that are apparent in the second row figure 1.10.

To summarize, we have emphasized the flexibility to control for variable interactions as one of the strengths of deep conditional portfolio sorts before. Now we see that the (two-way) interactions could have been discovered in a Fama-MacBeth regression framework, too. The deep conditional portfolio sort, however, is an efficient way to screen out the irrelevant

[24]Note that since we do not include higher-order polynomials of the past decile ranks, the average partial derivatives with respect to each variable will be linear and therefore cannot capture non-linear effects.

interactions when the set of candidates is potentially large. At the same time, it also allows to control for more involved interactions.

1.5.3 Transaction costs

While the strategy returns in our deep conditional portfolio sort appear high, they could still disappear after taking transaction costs into account. Strategies that are based on past returns generally have been found to have relatively high turnover (see de Groot *et al.* (2012) or Frazzini *et al.* (2013)), especially so, when they are based on recent past returns. As the deep conditional portfolio sort mainly exploits variation in the most recent past returns, we expect turnover to be high as well.

The first row of table 1.11 shows that this expectation is correct: An equal-weighted hedge strategy that goes long \$1 and short \$1 in the extreme portfolios has an average monthly turnover of 318%. Turnover is also high using the less extreme hedge returns that go long the ninth or eight decile and that go short the second or third decile, respectively.[25]

[25]These numbers are similar to those reported in de Groot *et al.* (2012) or Frazzini *et al.* (2013) for strategies based on short-term returns.

Table 1.11: *Turnover and trading costs*

	Low	2	3	4	5	6	7	8	9	High	High-Low	9-2	8-3
Turnover (monthly)	1.56	1.74	1.76	1.78	1.78	1.8	1.78	1.76	1.76	1.62	3.18	3.5	3.52
Trading cost (annual)	3.71	4.09	4.13	4.17	4.17	4.21	4.17	4.13	4.13	3.83	7.13	7.81	7.85
Gross return (annual)	-6.18	2.55	5.41	7.19	8.47	10.03	11.88	12.82	15.66	23.29	31.37	12.82	7.06
Net return (annual)	-9.88	-1.54	1.28	3.02	4.30	5.82	7.71	8.69	11.53	19.46	24.24	5.01	-.79

This table shows turnover and trading costs for the decile portfolios formed on predictions of the deep conditional portfolio sort in section 1.4.1. In all rows, the unit of the estimates is percent per month. The first row (turnover) shows monthly turnover for the ten decile portfolios and for the equal-weighted hedge strategies. Turnover is computed for a strategy that goes $1 long and $1 short. Trading costs are extrapolated using the results in Frazzini *et al.* (2013). The gross return is taken from table 1.5 and the last row computes net return as the difference between gross return and the estimated trading costs.

Recent research has noted the large heterogeneity of trading costs across different types of investors. Keim and Madhavan (1997) suggest a simple model to estimate transaction costs for a sample of institutional traders. However, as de Groot *et al.* (2012) note, this model can give rise to negative transaction costs in recent years. We, therefore, went with a rough calculation that extrapolates transaction costs from the turnover estimates in Frazzini *et al.* (2013). Even though the numbers might not apply to our sample exactly, they should be of similar magnitude, given the similarities of the data sample.

Using this approximation, we find that trading costs are around 7-8 percent per year (second row of table 1.11), close to the trading costs of the standard short-term reversal strategy investigated in the aforementioned papers. The last row of the table subtracts the approximate trading costs from the gross annual returns that we reported in table 1.4. After adjusting for trading costs, the hedge strategy that trades the extreme portfolios has an excess return of 24% per year. Trading the ninth minus the second decile (recall that these companies are larger and therefore probably more suited to the extrapolation from Frazzini *et al.* (2013)) yields an excess return of 5% per year. The excess return of trading the eight versus the third decile is insignificant and slightly negative. In other words, the iterative conditional portfolio sort manages to profitably spread 40% of the companies, even after adjusting for transaction costs.

While our strategy implementation is standard in the stock market anomalies literature, more sophisticated variants could be designed for trading purposes when transaction costs are taken into account. de Groot *et al.* (2012) suggest to reduce turnover of the short-term reversal strategy by holding on to the position in stocks even when they are not ranked in the extreme portfolios. We do not pursue their implementation here, but, given the return spread in the less extreme portfolios, it is plausible that such an implementation could be constructed here as well in order to reduce turnover and trading costs further.[26]

[26]For instance, Novy-Marx and Velikov (2014) finds that many anomalies can be exploited by following an (s,S)-type strategy that, e.g. buys stocks when they are in the highest decile but only sells them if they drop out of the highest quintile.

1.5.4 Risk factors or return characteristics?

Our results in section 1.4.1 indicate that portfolios that are based on forecasted returns from the estimated model have similar loadings on the Fama-French factors and the momentum factor, and yet consistently have different expected returns. While, at the surface, this seems to be a challenge to the four-factor model, we investigate the issue further in this section.

Table 1.12 shows the bivariate correlation between the strategy return, formed from the extreme portfolios, and the four standard risk factors. The strategy return displays low correlation with the market return and the value factor. It is somewhat higher correlated with the size factor and, as one would expect from a past return-based strategy, correlated with the momentum factor. In general, however, these correlations are low relative to the correlations between the other factors, which makes the strategy return a potentially suitable candidate factor.

Table 1.12: *Strategy return correlations with four factors*

	MKT	SMB	HML	UMD	DCPS
MKT	1.000	0.306	-0.320	-0.140	0.125
SMB		1.000	-0.241	-0.032	0.129
HML			1.000	-0.146	-0.081
UMD				1.000	0.291
DCPS					1.000

Bivariate correlations between the market return (MKT), the size (SMB) and value (HML) factors, the momentum factor (UMD) and the strategy return from an estimated deep conditional portfolio sort (DCPS).

This motivates table 1.13, which mirrors the analysis in Haugen and Baker (1996). It shows the average values of various firm characteristics in each decile of expected returns. The first panel of the table shows measures of risk across the ten deciles with no clear

(monotone) pattern. Average market beta is higher in the extreme deciles. The same holds for the profitability measures in the second panel. Interestingly, gross profitability is very similar in each decile, but expected returns are very different. This illustrates that our sorting is not driven by Novy-Marx (2013)' measure of gross profitability. Panel three shows that book-to-market is balanced across deciles, as one would expect from the balanced factor loadings in table 1.5. The last panel shows that the firms in the extreme deciles are on average smaller.

More intriguingly, since the strategy is based on the extreme deciles, it is worthwhile to compare the average values within these two deciles. Note that the values of most firm characteristics are very similar in these two deciles. The strategy appears to be based on, on average, riskier, less profitable and smaller companies. Yet, within the set of these firms, there are stark differences in returns that can be systematically predicted.

Table 1.13: *Firm characteristics: Portfolios based on deep conditional portfolio sort*

	Dec. 1	Dec. 2	Dec. 3	Dec. 4	Dec. 5	Dec. 6	Dec. 7	Dec. 8	Dec. 9	Dec. 10
Risk										
Debt to Equity	2.61	2.33	2.75	2.47	3.29	2.52	2.70	3.62	2.60	3.55
Long-term debt to Equity	1.43	0.78	1.15	0.81	1.61	0.74	0.92	1.98	0.90	2.09
Debt Ratio	0.51	0.52	0.53	0.53	0.54	0.54	0.54	0.54	0.54	0.54
Beta	1.18	1.09	1.07	1.05	1.05	1.05	1.06	1.08	1.12	1.19
Profitability										
Gross Profitability	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34
Return on Assets	-0.02	0.01	0.02	0.02	0.02	0.03	0.02	0.02	0.01	-0.02
Return on Equity	-0.15	0.03	-0.07	0.05	-0.15	0.04	0.04	-0.25	-0.18	-0.33
Profit Margin	-2.79	-1.31	-1.20	-1.04	-1.11	-1.60	-1.11	-0.75	-1.20	-2.41
Gross Margin	-1.31	-0.63	-0.26	-0.31	-0.19	-0.22	-0.27	-0.16	-0.37	-1.04
Earnings per Share	0.87	1.31	1.45	1.58	1.63	1.64	1.59	1.55	1.34	0.93

Continued on next page

Table 1.13: *(continued)*

	Dec. 1	Dec. 2	Dec. 3	Dec. 4	Dec. 5	Dec. 6	Dec. 7	Dec. 8	Dec. 9	Dec. 10
Basic Earnings Power Ratio	0.04	0.06	0.07	0.07	0.08	0.08	0.08	0.07	0.07	0.03
Price level										
Price Earnings Ratio	4.40	5.18	4.68	6.68	6.15	5.46	3.87	7.04	3.99	4.52
Book to Market	0.79	0.81	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.87
Price Sales Ratio	2.00	1.35	0.91	0.88	0.84	0.88	0.88	0.71	0.93	1.42
Dividend Yield	0.04	0.03	0.03	0.04	0.04	0.03	0.04	0.03	0.03	0.03
Activity										
Current Ratio	3.17	2.94	2.83	2.78	2.78	2.74	2.78	2.77	2.82	2.93
Quick Ratio	2.05	1.88	1.81	1.77	1.76	1.74	1.76	1.77	1.79	1.86
Net Working capital Ratio	0.30	0.29	0.28	0.27	0.27	0.27	0.27	0.28	0.28	0.29
Cash Ratio	1.49	1.23	1.13	1.07	1.07	1.04	1.07	1.06	1.11	1.22
Assets - Turnover Ratio	1.17	1.17	1.16	1.15	1.15	1.15	1.15	1.16	1.18	1.21

Continued on next page

Table 1.13: *(continued)*

	Dec. 1	Dec. 2	Dec. 3	Dec. 4	Dec. 5	Dec. 6	Dec. 7	Dec. 8	Dec. 9	Dec. 10
Inventory-Turnover Ratio	20.60	19.24	20.12	23.77	22.12	23.97	23.85	21.34	22.33	20.25
RandD	0.09	0.08	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.09
Others										
Size	681.95	976.24	1113.54	1177.19	1180.10	1177.22	1156.71	1085.86	967.02	633.43

Each month, all stocks are ranked by their estimated expected return based on a deep conditional portfolio sort that is based on all one-month return functions over the two years before portfolio formation. The table reports the average value of each firm characteristic in each decile over time.

Recall that alternative strategies that are based on buying the second (third) highest decile and selling the second (third) lowest decile rather than the extreme portfolios also make robust excess returns. Comparing deciles two and nine, or deciles three and eight, illustrates that the two corresponding portfolios are again very balanced throughout characteristics. As a sole characteristic, return-on-equity is lower in decile nine (eight) than in decile two (three), but other measures of profitability indicate that the portfolios are comparable along this dimension. While the non-extreme decile portfolios display similar characteristics, their excess returns vary and (see table 1.5) can be predicted from past returns. Since the portfolios based on the deep conditional portfolio sort appear not to be discernible based on many characteristics, we would not expect the strategy return that is based on it to help explain other anomalies.

As the strategy return itself appears to be unpriced by equilibrium models and unrelated to standard characteristics, the return could, in principle, be added as an additional risk factor to standard equilibrium models. However, in unreported results, we found that the strategy return, as expected, only weakly helps to explain other asset pricing anomalies, which is why we prefer the interpretation from a characteristics' rather than a risk factor perspective.

1.6 Conclusion

Some fifty years after the Capital Asset Pricing Model of Sharpe (1964), Lintner (1965) and Mossin (1966), and some twenty years after the three-factor model of Fama and French (1992), there is still a remarkable lack of consensus about which variables can be related to expected stock returns. To date, the literature has found more than 300 variables that spread returns in a way that is unaccounted for by the standard equilibrium models. This has led Green *et al.* (2013) to conclude that "either US stock markets are pervasively inefficient, or there exist a much larger number of rationally priced sources of risk in equity returns than previously thought." Surely, many of these variables contain correlated information, and some will not hold up out-of-sample, but, so far, the literature has not rigorously

identified which ones are fundamentally important. Furthermore, we have illustrated that some variables interact in non-trivial ways, making it more challenging to single-out the important ones with standard methodologies.

We provide a framework, deep conditional portfolio sorts, that is designed to deal with a large number of variables and their potential non-linearities and interactions. It also puts emphasis on systematic out-of-sample testing of all results. It connects model evaluation in finance to the machine-learning literature in computer science and can serve to bridge the two fields.

We apply our framework to find information in past returns that can be related to future returns. A simple, linear Fama-MacBeth framework finds moderate excess returns relative to the four-factor model. Using the same variables in the deep conditional portfolio sort framework, on the other hand, yields high and stable excess returns, indicating that the linear framework does not exploit all relevant information in the data.

Finance has criticized machine learning for producing black box predictions without any possibility to "get insights into the underlying structure of the data" (Breiman, 2002). We show that, even though the structure does not come in the form of simple equations, one can still extract interpretable information from the resulting deep conditional sorts. First, we find that, among the prior two years of one-month return functions before portfolio formation, the more recent ones are the most important for accurate return predictions. Second, some of these return-functions are non-linearly related to future returns, mostly returns between two and four months before portfolio formation. For instance, both high and low values of the return over the second-to-last month forecast lower returns. Third, many of the return functions display non-trivial interactions. For instance, the one-month return over the second-to-last month, is positively related to returns for stocks with low returns last month, but is positively related to returns with high returns last month. At a minimum, our results indicate that the relation between past and future returns is more complex than can be captured by any one summary return, like momentum or intermediate momentum.

Our results are robust to including a larger set of correlated return functions, and to the inclusion of other firm characteristics. Similar structures are also discovered within different size-sorted portfolios.

The finance literature that tries to understand the drivers of cross-sectional variation in expected returns, and the literature in machine learning that tries to predict stock returns have largely developed unnoticed of each other. The machine learning literature has focused on predicting stock returns from a few return-based and accounting-based variables jointly, but then has largely ignored the structure of the prediction equation, and has analyzed the quality of the prediction itself instead.[27,28] This article can also be viewed as an attempt to connect the two and to provide a synthesized framework that can be used in either field.

Lastly, deep conditional portfolio sorts can accommodate the inclusion of new predictor variables quite easily. Starting from the observation that if a predictor variable is relevant, it should show up among the most important variables that the method finds, one could just add the variable in question to the existing set of variables. Running the estimation on this extended set effectively controls for correlations with other variables and takes potential interactions and non-linearities into account. Our hope is that a framework around deep conditional portfolio sorts can significantly speed up the process of scientific discovery in this literature.

---

[27]See, e.g. Tsai *et al.* (2011) or Huerta *et al.* (2013).

[28]The variables that this literature uses for prediction are typically not motivated by results from the finance literature, but they appear to be chosen based on their availability in different datasets (convenience samples). In analyzing the predictions itself, the joint hypothesis problem (Fama (1965, 1970)) is usually ignored and the evaluation is conducted for raw return estimates.

Chapter 2

Inductive Testing: Theory and an Application to the Disposition Effect[1]

2.1 Introduction

In the last two decades, both the scale and breadth of data have erupted, so much that it has spawned a new term: "big data". These changes have affected many other fields and will surely affect empirical social science research (as they already have to some degree[2]). There are some fairly natural ways this will happen. First, as the scale of data — the number of observations — increases, standard errors fall and estimates become more precise. This allows researchers to test hypothesis at much greater granularity: empirical relationships that would otherwise be swamped by noise can be seen more clearly. Second, as the breadth of data–the number of variables–increases, new phenomena such as social networks or well-being, can now be brought under the empirical lens.[3] In both cases, big data boosts the

[1]Co-authored with Jon Kleinberg, Sendhil Mullainathan and Chenhao Tan

[2]See Einav and Levin (2014) for an overview of applications in economics.

[3]For instance, Dahl *et al.* (2014) use administrative data to uncover peer effects in maternity leave, Burke and Kraut (2013) use Facebook data to investigate how helpful strong and weak ties are in finding a new job after unemployment, and Dodds *et al.* (2011) extract happiness patterns from messages of over 60 million Twitter users.

power of existing empirical tools: the larger the dataset the more we can do with the tools we already have.

Big data, though, has also brought with it new empirical tools. Specifically, machine learning techniques appear to represent a different way to analyze data sets and are particularly suited to large data sets. In this paper we argue that machine learning can be profitably applied to social science research and in particular provide a new way to test theories.[4]

A fictional example illustrates our approach. Imagine a medical researcher in the 19th century who has a novel theory: that mental health and physical health are connected. Pessimism, she conjectures, impedes the healing process and thereby mortality is a self-fulfilling prophecy. She recognizes that simply correlating reported pessimism with mortality is an inadequate test of her theory since pessimism may be correlated with other factors that affect health. So instead she isolates another empirical variable for a more precise test. She notes that a patient should be more pessimistic if his roommate dies, and roommates can be randomly assigned. Amongst the theories she knows, there is still one potential confound–doctor quality–since roommates could share doctors. So she collects data on doctor assignments and controls for this in her experiment.

To her delight, she finds support for her theory: roommate mortality predicts own mortality. Of course, today we know that her test was confounded. The more likely interpretation for the researcher's finding is that bacterial infections spread easily between roommates. Our 19th century researcher, however, isn't aware of the germ theory of disease and did not control for it. Two points are worth noting about this example. First, this is not simply a case of "omitted variable bias". The causal effect of roommates is correctly estimated. The empirical relationship is right; the theory that predicted it is wrong. We might call it "omitted theory bias" because she did not control for the germ theory. Second, this faulty interpretation is clear only in hindsight. Because we now know the powerful

[4]Other interesting work that uses or conceptualizes machine learning or high-dimensional data techniques in social science include Belloni *et al.* (2011), Belloni *et al.* (2012), Belloni *et al.* (2013), Farrell (2013), Gelman and Imbens (2013) or Fan *et al.* (2011). See also the recent survey by Belloni *et al.* (2014) and the references therein.

effect of bacterial infections on health, we see immediately this alternative interpretation. How could a researcher blind to this future discovery anachronistically have been expected to anticipate it and control for it? Unlike omitted variable bias, omitted theory bias seems impossible to address. We will argue though that machine learning techniques can at least partly address it.

To understand how machine learning can do so, it first is useful to see how it is used by computer scientists for the problems it was initially created to address. Take sentiment analysis: a classic artificial intelligence problem where the task is to classify some text–such as a movie review–as either favorable or unfavorable. Early artificial intelligence approaches began by noting that we, as humans, solve this problem trivially. So they would introspect about what words (or phrases) constituted a favorable or unfavorable review, such as "gripping", "cliched" or "moving". An algorithm could then scan reviews for those words and classify reviews into favorable and unfavorable ones based on the frequency with which those words occurred. This method for choosing words does turn out to work significantly better than randomly guessing the sentiment of a review. However, it is also limited by the candidate lists of words that researchers could create from introspection.

But there is an alternative strategy, reflected in current approaches to the problem. Instead of *deducing* from priors what words might matter, why not *induce* the words that work? Specifically, put together a large set of training data–examples of text where sentiment is known–and see which words are diagnostic of sentiment in this data set. Recasting problems as prediction problems is a key element of the machine learning framework–in this case, given a movie review, to predict its sentiment using a set of features derived from the review, such as the words in it. A second key element in the machine learning framework is the effective construction of the variables to be used in this prediction. If we applied traditional regression techniques to this problem, we would need to decide upon a modest set of variables to include as regressors where the left hand side variable is sentiment. This is better than ex ante deciding which words are diagnostic but it limits the set of words we can search over.

Instead, machine learning approaches to this problem begin with a very large set of features as potential inputs. For example a standard bag-of-words technique ("unigrams") would form a vector of dummy variables indicating the presence of every single word in the database. So if there are 5,000 words used in any review then there are 5,000 dummies. One could also include a vector of "bigrams": a variable for each pair of consecutive words appearing in every review. This most often results in a dataset with more variables than observations. Machine learning, or statistical learning, techniques effectively (and provably) can search such large spaces and find prediction functions that work well. In the case of sentiment analysis, for example, prediction rules extracted in this way trump carefully crafted rules based on words deduced a priori.[5] This story replays itself in many areas, machine learning making tractable problems that were once thought to be intractable, from handwriting recognition to word sense disambiguation.

Sentiment analysis highlights several features of machine learning algorithms. First, note that machine learning techniques are powerful exactly because they can handle large numbers of independent variables (k) relative to sample size (n). As in the example above they can even handle situations where $k > n$. Second, note that they are designed exactly to address over-fitting: they do not simply do well in the data provided but on new data (drawn from the same distribution). Finally, and most importantly note that machine learning techniques focus on prediction, while econometrics focuses on estimation. Put simply, if we think about problems involving the analysis of a relationship $y \approx \beta \cdot x$, machine learning focuses only on $\hat{y}$ and traditional econometrics focuses on accurate estimates of $\hat{\beta}$. Machine learning techniques therefore often provide a weaker guarantee: that $\hat{y}$ will be close to y. These three ingredients are what facilitate an inductive approach to problems such as sentiment analysis.

Empirical social science research is more interested in understanding theories — in other words, more interested in $\hat{\beta}$. We would like to argue, however, that machine learning techniques, initially designed for prediction, can also be used for theory testing. Let's

[5] See Pang *et al.* (2002)

return to our 19th century researcher and compare her problem to the sentiment analysis problem. Her goal is different–she is not trying to come up with a black box mortality predictor but instead is trying to test a theory. But in one crucial respect her problem was similar. In deciding what theories to control for she took a deductive approach. Much like early attempts at sentiment analysis she used her prior knowledge of alternative theories to decide on what variables to include–in this case measures of doctor quality. The steps she followed in our story are basic to the scientific method: formulate a theory, deduce from it a hypothesis that a specific target variable should affect some outcome, and then test this hypothesis by controlling for known theories. We refer to this procedure as *deductive testing*.

But with machine learning she could have done better. Let's imagine a 19th-century researcher who doesn't know the germ theory but does have access both to modern machine learning techniques *and* a large data set including many variables — not just roommate random assignment and other variables she would hand-collect. If she were now to come up with the best predictor of mortality, she'd notice two facts. First, variables that she could not understand–such as sharing of scalpel or test instrument with a sick person–would predict mortality. That by itself does not invalidate her theory; it merely suggests there is more for her to understand. Second, and more importantly, once these variables are accounted for she might find that roommate assignment no longer matters. To be more precise, she could ask the question: "How well does a machine learning algorithm with access to roommate health do relative to one that does not have access to roommate health?" If pessimism is right, having access to roommate health ought to improve prediction. We call this approach *inductive testing*.

The steps for inductive testing are slightly different from deductive testing. As before, a theory must be formulated and, given a data set, a variable relevant to the theory must be found in it. Greater care must now be taken though to ensure that other variables in the data strongly related to the theory are also excluded. For example, if some other variable proxies for roommate health that too must be noted. But unlike deductive testing, inductive testing does not curate the control variables. Instead as many of the control variables as

possible are included. This is the sense in which inductive testing makes full use of machine learning. The inductive test then involves comparing predictive performance (on y) between an algorithm given access to the full variable set and the variable set that excludes those related to the theory to be tested. Inductive testing is appealing because it addresses omitted theory bias to a degree. It allows us not just to control for known theories but some of the unknown ones.

Three important points are worth noting about the inductive method. First, the quality of an inductive test — its power so to speak — depends on the scope of the data. A data set with very few other variables will inevitably produce a weak inductive test. Second, theories that exclude a large set of variables as auxiliary variables are poor candidates for inductive tests. In the extreme, if every variable is in principle explainable by the theory, then an inductive test is not feasible. Of course, one could ask whether theories of this kind are well-formed. Finally, inductive tests are only tests. When they reject a theory, they do not necessarily give guidance as to why. Looking at the prediction function, one might gain some insight into what is happening. But it need not be the case. In the example above, induction simply says that the presence of bacteria predicts mortality: it does not "discover" the germ theory of disease, although it may provide a roadmap to guide future scientific research.

In this paper, we first develop these ideas by creating a simple model of science. This model helps to clarify, among other things, the distinction between a theory test and an empirical hypothesis test. For example, it clarifies how even if there is no omitted variable bias a theory test can fail, as in the fictional medical example above. In this model, we define a deductive test and then show how a machine learning algorithm can be used to create an inductive test. We show how, under performance guarantees of the machine learning algorithm, an inductive test can outperform a deductive one.

We apply these ideas by inductively revisiting a classic deductive test. One of the key observations in behavioral finance is the disposition effect: the idea that investors are reluctant to realize their losses. Theories for the disposition effect are based in part on the

notion of loss aversion, i.e. losses loom larger than gains. While their is no consensus in the literature, a popular theory–"realization utility"–makes the further assumption that this disutility is only realized when the gains or losses are realized (Barberis and Xiong, 2012).[6] Realization utility is an interesting theory because of its broad applicability, affecting everything from stock markets to housing markets (Genesove and Mayer, 2001). It is also important because evidence for realization utility forms one of the most interesting pieces of evidence in behavioral finance that is not directly related to predictability of asset prices.

We begin by replicating the classic Odean (1998) study on this topic. Using his data, we find–as he did– strong deductive evidence of realization utility: stocks are much more likely to be sold when they are in the gain domain than in the loss domain. For the inductive test we create a large set of variables unrelated to realization utility, such as: How does the price today compare to the price in the last t days? Importantly, these variables do not include any information on the purchase price, but only on recent price movements. We find, of course, that our full variable set does much better at predicting selling behavior than the gain information alone. But, importantly, we find that once these variables are available to a machine learning algorithm, incorporating information about gain relative to purchase price does not improve predictive performance at all. Whatever was generating the disposition effect does not appear to be related to purchase price.

As noted above, an inductive test does not directly give guidance as to why a theory is rejected. But indirectly one can examine the variables that are included in the model and gain some informal understanding. In this case, we note that a variable which indicates whether the purchase price is in the top quartile of prices seen recently matters significantly for selling. Being in the bottom quartile also matters somewhat, though significantly less. Of course, if prices for this stock are high relative to what the investor has seen, the stock will also be (on average) more likely to be in the gain domain. But once this fact is accounted for being in the gain domain adds no value at all.

[6]It is important to note that by realization utility we mean here the combination of realization utility and some loss-averse utility function based on *purchase* price.

Put simply, realization utility (based on purchase price) passes a deductive test but not the inductive one. The caveat regarding purchase price is important here, of course, since other realization utilities might be defined differently. At the least, though, the original disposition effect does not seem robust to an inductive test.

In summary, the contributions of this paper are three-fold. First, we introduce a formal framework for theory testing that illustrates what scientists are assuming in hypothesis testing and how hypothesis testing is related to theory testing. Second, we propose the idea of inductive testing, and combine existing findings in the machine learning literature to propose a procedure for inductive testing. Third, we apply our framework and conduct inductive testing to the disposition effect.

The paper is organized as follows. Section 2.2 develops our theory more concretely and provides conditions under which deductive and inductive tests work. Section 2.3 describes how we apply inductive testing to the disposition effect. Section 2.4 concludes.

2.2 Model

Setup and assumptions. This section outlines a theoretical framework to think about the relation between a (binary) outcome of interest, an underlying theory and its empirical proxies. We define deductive and inductive tests and derive conditions under which the two lead to the same or different answers, respectively. The reader should note that this section is tentative but highlights some of the main ideas and concepts. A more complete model is work in progress.

Our goal is to examine an outcome variable Y as a function of unobserved concepts or theories M_0, M_1 and M_2, where $Y = g(M_0, M_1, M_2)$. Here, M_0 is the theory that the researcher wants to test, M_1 are other theories that are known to the researcher and M_2 are unknown theories. In our example in the introduction, Y would be patient mortality, M_0 would be the pessimism theory and M_1 would be doctor quality. We would like to know how well these theories explain patient mortality.

Assumption 1 (Data-generating process)**.** *Let binary Y be a function of underlying models with $Y = g(M_0, M_1, M_2)$ and let $M_j = h(X_j), \forall j$ such that by composition $Y = f(X)$. X is drawn from distribution $\mathcal{D}$.*

In any empirical application, the data generating process can therefore be written in terms of Y as a function of X, where X have a distribution that we leave unspecified and the number of measurable variables X can potentially be infinite.

In what follows, we will work with an abstract performance function $\Pi(Y; Z)$ to measure how well Y is explained by Z. This performance measure can be a statistical loss function or some other function; for know, we only assume that it exists and that it has nice properties.

Definition 1 (Performance measure)**.** *A performance measure $\Pi(Y; Z)$ that reports how well Y is explained by Z is available for any set of Zs. Normalize $\Pi(Y; \emptyset) = 0$.*

Models M_j only exist in scientists' minds. They allow generalization to different contexts and their structure extends past any one data set. In any particular context, however, the scientist has to choose one or more empirical variables that stand in for the model under investigation. The distinction is often glossed over but we aim to model it explicitly. Normally, the test of an empirical relationship is synonymous with testing a theory, but, of course, a deduced hypothesis may be empirically true while the theory is false. It could simply be the case that other theories also imply the same hypothesis. While we know this distinction, it is rarely part of the formal empirical framework. Rather, it is left to the iterative process of science itself.

Let us define what it means for a set of variables X to be an empirical approximation to model M_j as follows.

Definition 2 (Proxy)**.** *A set of variables X_j is a* proxy *for theory M_j if, for all X_{-j}*

$$\Pi(M_j; X_j, X_{-j}) - \Pi(M_j; X_{-j}) > \gamma.$$

In words, X_j is a proxy for theory M_j if there is at least some residual signal in X_j to explain M_j.

We then assume that researchers can find proxies for the theory under investigation and for known theories.

Assumption 2. *Researchers are able to find empirical proxies for known theories M_1 and for the theory under investigation M_0.*

Assumption 2 reduces the role of the researcher to the *curation* of the appropriate X_j variables for any theory M_j.

Assumption 3 implies that having better proxies leads to better performance in explaining the outcome variable.

Assumption 3 (Smooth model relevance)**.** *Let $\hat{M}_j$ be an empirical approximation to M_j and let $Y = g(M_0, M_1, M_2)$ as before. Define $\mathcal{K}$ implictly by*

$$\frac{\Pi(Y; \hat{M}_0, \hat{M}_1, \hat{M}_2)}{\Pi(Y; M_0, M_1, M_2)} = \mathcal{K}(\Pi(M_0, \hat{M}_0), \Pi(M_1, \hat{M}_1), \Pi(M_2, \hat{M}_2)).$$

M_j is smooth-relevant *if*

$$\exists \theta_j > 0 : \mathcal{K}(\Pi(M_j, \hat{M}_j), .) - \mathcal{K}(\Pi(M_j, \hat{M}'_j), .) > \theta_j(\Pi(M_j, \hat{M}'_j) - \Pi(M_j, \hat{M}'_j)).$$

We will use *relevant* and *smooth relevant* interchangeably. The smooth relevance assumption implies that, as we find better proxies for a theory, we observe an increase in the performance of predicting Y.

Finally, definition 3 lays out what it means for a function $f()$ to be learnable with high probability.

Definition 3 (Learnability)**.** *$f(X)$ is learnable if and only if, given data drawn from $\mathcal{D}$ and $Y = f(X)$, we can find a function $\hat{f}(X)$ such that the performance of $\hat{f}(X)$ is close to the performance of $f(X)$ when the sample is sufficiently large. Formally, f is* learnable *if and only if*

$$\forall \frac{1}{2} > \epsilon > 0, \frac{1}{2} > \delta > 0, \exists n' : \text{ If } n > n', \text{ then } Pr(|\hat{\Pi}(Y; X) - \Pi(Y; X)| > \epsilon) \leq \delta.$$

This definition is motivated by results in the field of statistical learning theory, a field that tries to provide statistical guarantees for the generalization error (the error when predicting

new data points) of estimated prediction functions under minimal assumptions.[7] Such guarantees have been derived for a variety of algorithms, including decision-trees that we use below. We, therefore, make the following assumption about learning algorithms that we use.

Assumption 4 (Technology)**.** *The researcher has access to a learning technology such that f is learnable.*

Deductive and inductive testing. We define a deductive test as follows. First, estimate $\Pi(Y; X_0, X_1)$ and $\Pi(Y; X_1)$. The relevance of M_0 is assessed by comparing these two quantities; in particular, M_0 is accepted as a relevant theory if

$$\hat{\Pi}(Y; X_0, X_1) - \hat{\Pi}(Y; X_1) > \kappa.$$

In contrast, an inductive test accepts M_0 as relevant if

$$\hat{\Pi}(Y; X) - \hat{\Pi}(Y; X - \{X_0\}) > \kappa.$$

Note that a deductive test can at most control for all theories that are known. An inductive test, on the other hand, controls for *all* theories, known and unknown, that can possibly be covered by measurable variables X. The inductive test relies on key insights from the machine learning literature. First, the recent literature has developed ways to estimate models, even when there are more explanatory variables than there are observations. In an inductive test, we want to include as many X variables as possible (as long as they provide diverse information) since, by definition, we do not know which variables could proxy for unknown models M_2. This implies a change in the scientific approach to measurement of variables: Scientists should not only measure and use what they think is important; in fact, "useless" variables can turn out to be very useful. At the same time, to set up an inductive test, scientists need to carefully curate the variables X_0 that relate to the theory to be tested, while curation of other variables is unnecessary. Second, in such a variable-rich situation,

[7] See e.g. Valiant (1984).

concerns about overfitting the data naturally arise. But the machine learning literature has developed ways to deal with this issue (cross-validation, boosting) that we explain in more detail below. Statistical performance guarantees bound the generalization error on new data and, they imply that this error becomes very small with high probability when the number of observations in the estimation data increases.

Proofs. We first establish conditions under which deductive and inductive testing both accept a theory M_0.

Theorem 1 (Acceptance)**.** *Suppose M_0 is relevant, X_0 is a proxy for M_0 and Y is learnable, $X = \{X_0, X_1, X_2\}$. Then deduction and induction accept.*

Proof. Let us denote the sample size by n. Because Y is learnable, there exist ϵ, δ and n' so that

$$\forall n > n' : Pr(|\hat{\Pi}(Y; X) - \Pi(Y; X)| > \epsilon) < \delta$$
$$\forall n > n' : Pr(|\hat{\Pi}(Y; X - X_0) - \Pi(Y; X - X_0)| > \epsilon) < \delta,$$

thus

$$\forall n > n' : Pr(\hat{\Pi}(Y; X) > \Pi(Y; X) - \epsilon) \geq 1 - \delta$$
$$\forall n > n' : Pr(\hat{\Pi}(Y; X - X_0) < \Pi(Y; X - X_0) + \epsilon) \geq 1 - \delta.$$

On the other hand, because M_0 is relevant and X_0 is a proxy for M_0,

$$\Pi(Y; X) - \Pi(Y; X - X_0) > \theta_0(\Pi(M_0; X) - \Pi(M_0; X - X_0)) > \theta_0\gamma$$

It follows that

$$\forall n > n' : Pr(\hat{\Pi}(Y; X) - \hat{\Pi}(Y; X - X_0) > \theta_0\gamma - 2\epsilon) \geq (1 - \delta)^2 > 1 - 2\delta.$$

Hence, if $n \to \infty$, $\epsilon \to 0$ and $\delta \to 0$, as long as $\theta_0\gamma > \kappa$, induction accepts the theory. The proof works analogously for deduction if we replace X_{-0} with X_1.

□

Theorem 1 implies that a theory M_0 is accepted if it is sufficiently relevant (θ_0 is large) and it can be proxied well (γ is large).

Hence, if a theory M_0 is true, then both deductive and inductive tests accept it under the same conditions. If a theory is false, however, there are situations in which a deductive test accepts it as true, while an inductive test rejects it, as illustrated in theorem 2.

Theorem 2 (Counterexample)**.** *Suppose $Y = g(M_1, M_2)$ (M_1 and M_2 are relevant), and X_0 is a proxy for M_2 conditioned on X_1 but is no longer a proxy conditioned on X_2. Then deduction accepts but induction rejects.*

Proof. First, deduction accepts, because following the proof of theorem 1, as $n \to \infty$, $\hat{\Pi}(Y; X_1, X_0) - \hat{\Pi}(Y; X_1) > \theta_2\gamma > \kappa$.

On the other hand, for the inductive test, because $\Pi(M_1; X) - \Pi(M_1; X - X_0) \to 0$, $\Pi(M_2; X) - \Pi(M_2; X - X_0) \to 0$ and $\theta_0 = 0$, $\hat{\Pi}(Y; X) - \hat{\Pi}(Y; X - X_0) \to 0$.

□

2.3 An Application to the Disposition Effect

We apply inductive testing to the disposition effect: the idea that investors are reluctant to realize their losses. Theories for the disposition effect are based in part on the notion of loss aversion, i.e. losses loom larger than gains. While there is no consensus in the literature, a popular theory makes the further assumption that this disutility is only realized when the gains or losses are realized (Barberis and Xiong, 2012). We refer to this combination of loss aversion with purchase price as the reference point and that this utility is felt at the time of sale as "realization utility."[8] This is an interesting theory because of its broad applicability. The disposition effect is directly implied by realization utility and could potentially affect everything from stock markets to housing markets (Genesove and Mayer, 2001). It is also

[8]The specific assumption here is that the purchase price is the reference price. Other reference points are obviously possible and could also be called realization utility.

important because evidence for realization utility forms one of the most interesting pieces of evidence in behavioral finance that is not directly related to predictability of asset prices.

The best-known piece of evidence for the disposition effect is provided by Odean (1998) who focuses on data from a large brokerage house. His transaction level data allows him to know, for each trader, which stocks were sold and which stocks were kept day-by-day for several years. Using this data, he performs a simple deductive test. He compares whether stocks in the gain domain — whose current prices are above purchase price — are more likely to be sold than those in the loss domain — whose current price are below purchase price. In fact, he finds that stocks in the gain domain are almost 50% more likely to be sold, an effect that reverses in December, presumably due to tax reasons (Ivkovic *et al.*, 2005).

This example has all the key ingredients for an inductive test. First, the data set is very large. We use data from 40,000 traders. The scope — containing at least the entire time series of price data for every stock — is also large. Second, the theory to be tested is precise. It does not suggest many auxiliary variables that might matter. Because the disposition effect is so clearly focused on the experienced gain or loss, using the initial prices as the reference point, the auxiliary set of variables the theory could explain is not large.

To implement inductive testing, we look at a prediction problem. We view each observation as a time series of data that begins with a purchase at time 0. Each point in the time series is labeled "sell" if the account sold the stock on that day, otherwise "hold". Some observations involve a "sell" at the end and others involve a "hold" at the end. Notice that if someone holds a stock for T time periods, they create in effect T such time series. This creates a prediction task: classify whether an observation was "sell" or "hold" (corresponding to the outcome Y), given features of the time series. For each time series, we encode a large set of variables, including a variable that encapsulates the disposition effect. We call this variable *Gain* (X_0 in our framework).

First, as a simple check, we replicate Odean (1998)'s original finding that *Gain* is in fact predictive on its own. Indeed, we find that traders are significantly more likely to sell a stock in gain than in loss. Second, for the inductive test, we then run two prediction algorithms

on this large data set. In all the cases, we find that — though Gain is individually valuable — it does not help the predictive performance much when all the variables are considered. But as an inductive test, this might be too restrictive. After all, our variables also include other variables that are close to Gain. We, therefore, repeat the test excluding not only Gain but also a number of auxiliary variables. Again, we find no reduction in performance. When we subsequently exclude more and more auxiliary variables that could be related to Gain, performance never changes. All this suggests that while Gain appears to be strong support for the disposition effect in a deductive test, it is in fact likely capturing some other phenomenon.

To diagnose more carefully exactly what Gain is proxying for, we move to a simpler version of this classification problem in section 2.3.5. We use a brute-force *table lookup* procedure to enumerate all possible combinations of variables. This is an extreme form of a very wide dataset that allows us to zoom in and understand actually what is happening for every configuration of values. We analyze a *scatterplot* of performance values as a function of every possible subset of a reduced variable set to show that a stock's price quartile relative to the history of prices matters instead of Gain. We also find that recent price changes are very predictive of selling. These findings suggest that, in fact, Gain may really be the tip of the iceberg for some other phenomenon, one unrelated to the disposition effect. As noted earlier, an inductive test cannot give firm guidance to new theories but, as in this case, it can give some clues. The fact that quartile matters so strongly suggests that individuals might be trying to market-time their sales based on the prices they have observed. There is some evidence that their attempts at timing depend both on the long run of prices they have seen — they sell when the price is in the top quartile — as well as the short run price dynamics — recent consistent upward or downward trends are both very predictive of selling. These results hint at directions for future theoretical and empirical research.

2.3.1 Data description

Data for our study come from two datasets. Terrance Odean provided us with individuals' trading record data that are also used in Barber and Odean (2000). The dataset originates from a large discount brokerage house and contains all trading activities for a sample of 78,000 households in 158,034 accounts from 1991 to 1996. A typical observation in that dataset is made up of an account identifier, a security identifier, a date, an action (buy or sell), the quantity that is traded and the trading price. Using CUSIP numbers as a joint identifier, we supplement these data with stock-specific information from the Center for Research in Security Prices' (CRSP) daily stock file. The information from both datasets allows us to construct and evaluate security and portfolio returns for each individual account on a daily basis.

For the purpose of our study, we adjust our sample in line with the previous literature (see, for example, Odean (1998)). We restrict the dataset to trades in common stocks, we drop short-selling activities, we drop observations that include buys before the beginning of our sample, and we adjust stock prices for stock splits. A trade by an investor is sampled into our study if the security is held for at least 25 trading days and if there is only one selling action over the 250 trading days following the buying decision. We call each (account number, security number) - combination that is part of the study a *time series*.[9] Sales can be partial reductions or full liquidations of the position. In order to employ all variables discussed in section 2.3.4, we require that for each time series price information for at least 200 days before the purchase day is available. Table 2.1 shows summary statistics for the time series that are part of our study. Subject to our selection criteria, we sampled 166,894 time series that come from 39,956 different accounts and that contain 6,141 different securities. The average holding time is 103 days.

[9] If an account sells a security entirely and then buys it again, this is counted as a new time series.

Table 2.1: *Summary Statistics*

Total number of time series	166,894
Average holding days	103.08
Number of accounts	39,956
Number of securities	6,141

2.3.2 Deductive testing

Following Odean (1998), we define $Gain(X_0)$ as $I(\text{selling price} > \text{purchase price})$ to capture the disposition effect. Table 2.2 replicates the findings of Odean (1998) for our two datasets (the full sample and the balanced sample in the classification task). For every (account number, security number) - combination in the sample and each day t in the sample, we compute whether the security was trading at a gain or at a loss on that day relative to purchase price, and whether it was sold or not. This differs from Odean (1998) because he only evaluates gains and losses on days when the investor was active (traded at least one security). Our results can be regarded as an unconditional version of the statistics that he reports. We follow Odean in computing the proportion of gains that were realized (PGR) as the number of realized gains divided by the sum of realized gains and gains that investors held on to. The proportion of realized losses (PLR) is computed in the same fashion. Table 2.2 shows that the proportion of realized gains is significantly higher than the proportion of realized losses in either dataset, which confirms the validity of an unconditional version of Odean's finding in his original study using trading accounts of 10,000 investors. Thus, a simple deductive test lends support to the hypothesis that investors are more likely to sell gains than losses.[10]

[10]Here, we present the Odean test in its original form. Note that it can also be written as the regression problem

$$Sell_{it} = \beta_0 + \beta_1 I(\text{selling price}_{it} > \text{purchase price}_{it}) + u_{it}$$

which implies $\hat{\beta}_1 = \text{PGR} - \text{PLR}$. Using e.g. R^2 as a performance measure, the simple t-test here and the deductive test as a function of a performance measure in section 2.2 are equivalent.

Table 2.2: *PGR and PLR for two datasets.*

	Classification task	Full data
PGR	0.535	0.0049
PLR	0.452	0.0038
Difference	0.083	0.0011
t-statistic	46.558^{***}	50.748^{***}

This table compares the aggregate Proportion of Realized Gains (PGR) to the aggregate Proportion of Realized Losses (PLR), where PGR is the number of realized gains divided by the number of realized gains and the number of gains investors held on to, and PLR is defined likewise. Gains and losses are defined relative to the security's purchase price. The t-statistic reported is for the hypothesis that PGR and PLR are equal. The test statistic is computed as $(PGR - PLR)/\sqrt{\frac{PGR(1-PGR)}{n_{ug}+n_{rg}} + \frac{PLR(1-PLR)}{n_{ul}+n_{rl}}}$, where n_{ug}, n_{rg}, n_{ul} and n_{rl} are the numbers of unrealized gains, realized gains, unrealized losses and realized losses.
*** $p < .001$, ** $p < .01$, * $p < .05$.

2.3.3 The prediction problem

To map this problem into our framework in Section 2.2, we formulate a prediction problem to explain traders' selling actions. For all the time series that satisfy the requirements in Section 2.3.1, we construct a dataset that includes all days on which a stock is sold and we randomly sample another day from the same time series on which the stock was not sold.[11] The prediction task can now be formulated as follows: given an observation of a time series, can we predict whether it is a selling day or a holding day? We call this the *classification task*. For the classification task, an algorithm can use any combination of the features discussed in section 2.3.4. We randomly sampled 100,000 time series as training data and the remaining 66,894 time series as testing data to evaluate the expectation of the performance of different algorithms. We use cross-validation to select parameters in the

[11]It is computationally too expensive if we include all points in the time series in the prediction task (17M data points). Instead we create a dataset by choosing a selling point and randomly sampling a holding point for all time series where the holding time is longer than 25 days and shorter than 250 days. We require the "hold" day to be at least 20 days after purchase to leave enough space for computing the features defined below in Section 2.3.4.

machine learning method. Cross-validation works as follows: we split the sample into five parts; four parts are then used to estimate the model, and one part is used to evaluate the prediction error. The procedure is repeated using each of the five parts as testing part in turn. We apply this procedure in the training data and choose the parameter that gives the best average cross-validation performance. We then use that parameter and estimate a model based on all the training data, and report the expected performance of the estimated model based on the held-out testing data.

2.3.4 Inductive testing

Now we investigate whether an algorithm will discover the disposition effect when it is presented with a large number of potentially important variables. Given the time series that we see until a given day, we can, of course, extract many different variables other than *Gain*. A large subset of these variables will not be related to the theory of realization utility, but may be able to explain traders' selling behavior. In Section 2.3.4, we first introduce the variables that we construct. We then explain the different machine learning methods that we use in this paper. Finally, we present results based on the inductive testing framework that show that the disposition effect becomes irrelevant once other variables are accounted for.

Variable description

The disposition effect can be represented by a particular function of past and present price movements: if a security is trading at a price higher than the purchase price, an investor should be more inclined to sell the security. The standard notion of the disposition effect would tell us that this feature is helpful to predict investors' selling actions. We, therefore, make sure that our variable set includes variables that can be related to the disposition effect so that they can potentially be discovered in the learning exercise.

Define p_0 as the buying price and p_j as the price on day j. In order to predict a selling action on day t, the algorithm can use functions of $p_0, \ldots, p_t$. In principle, there is an infinite number of functions of prices that might be related to trading decisions. We discipline the

search by permitting the algorithm to use five types of functions that operate on a given price domain, and that naturally cover a variety of theories. Some notation will be helpful to define functions and domains.

The domain on which functions can operate is defined by two types of mappings from integers (i, j) into prices, which we term *A-ranges* and *B-ranges*. *A-ranges* start from both ends of the price series and move towards the middle, $A : (i, j) \rightarrow \{p_i, \ldots, p_{t-j}\}$. For instance, $A(0, 0)$ denotes the set of all prices $p_0, \ldots, p_t$. *B-ranges* start at date t and go back in time, $B : (i, j), j > i \rightarrow \{p_{t-j}, \ldots, p_{t-i}\}$. For instance, $B(0, 1)$ defines the domain that includes only the two most recent prices, p_{t-1} and p_t. These two types of domains allow us to compute information that is based on the entire holding period of traders until the selling day as well as short-term and long-term price changes. In our analysis, we consider the following domains on which a function can operate:

- $A(i, j)$, where $0 \leq i < 10$, $0 \leq j < 10$. These domains define a broad range of prices from the distant past around the buying action to the recent past close to time t. They are also commonly associated with the disposition effect.
- $B(i, j)$, where $0 \leq i < 5$, $i + 1 \leq j < 5$. These define recent price movements.
- $B(i, j)$, where $i = 0$, $j \in \{20, 40, \ldots, 200\}$. These define medium term to long-term price movements relative to t.

We define five different types of functions that operate on these domains. The functions are chosen to reflect a broad array of potential theories that we might or might not have little knowledge about. Each function defines a variable for a given price domain $A(i, j)$ or $B(i, j)$. Some of the variables that can be generated by these functions have the flavor of the disposition effect even though they slightly differ from the original implementation. These are the auxiliary variables in our exercise. Other variables are not related to the disposition effect at all. They mimic the process of scientific discovery when we move to inductive testing. The variables fall into the following five categories:

- *Gain*: This function is equal to 1 if the price at the end of the range is higher than the price at the beginning of the range, -1 if it is lower and 0 if it is identical. Note how the gain function defined on an A-range re-builds variables that are typically related to the disposition effect. To provide an example of how our notation works, $\text{Gain}(A(0,0)) = 1$ if $p_t > p_0$ and $\text{Gain}(A(0,0)) = -1$ if $p_t < p_0$, which is the most basic definition of a variable that is used in studies of the disposition effect. (This serves as our variable X_0.) We vary the domains to generate auxiliary variables that still preserve qualities of the disposition effect.[12]

- *Quartile*: These functions are equal to 1 if p_t is in a certain quartile of a given price range. For instance, if p_t is in the lowest quartile of the prices $\{p_0, p_1, \ldots, p_{t-1}\}$, the function $\text{Q1}(A(0,1))$ is equal to 1, and it is zero otherwise.

- *Max*: This function is equal to 1 if p_t is higher than the maximum of prices in a given price range, -1 if it is lower and 0 if it is identical. For instance, if p_t is the maximum price over $\{p_0, \ldots, p_t\}$, the function $\text{Max}(A(0,0))$ is equal to 1.

- *Min*: This function is equal to 1 if p_t is higher than the minimum of prices in a given price range, -1 if it is lower and 0 if it is identical. For instance, if p_t is the minimum price over $\{p_0, \ldots, p_t\}$, the function $\text{Min}(A(0,0))$ is equal to 1.

- *RefGain*: While *Gain* indicates whether the current price is greater than some fixed price in the past, *RefGain* allows for comparison between the current price and a reference price that is evolving over time. In particular, iteratively define $RefPrice(t, \eta) = \eta * RefPrice(t-1, \eta) + (1-\eta) * p_t$, where η is a parameter that adjusts the weight between the current prince and the past price. *RefGain* is equal to one if the current price exceeds the current *RefPrice*, and it is zero otherwise. We construct four such variables, operating on the A(0,0) domain for parameters $\eta = 0.9, 0.99, 0.999, 0.9999$.

[12]In the original work by Odean (1998) only the purchase price is considered as the reference point, but others have argued that the reference point would be better modeled as expectations (Barberis and Xiong (2009), Meng (2013))

Applying these functions on the 120 domains yields a total of $120 \cdot 7 + 4 = 844$ candidate features to consider.

Learning algorithms

We apply three different learning algorithms to our classification problem to make sure that results do not depend on a particular machine learning method. Here, we describe in more detail how each of them is implemented.

Lasso. Lasso (*l*east *a*bsolute *s*hrinkage and *s*election *o*perator) was first introduced in Tibshirani (1996). The Lasso objective function minimizes the squared deviations from an outcome variable (as in ordinary least squares regression) subject to a penalty term on the regression coefficients:

$$\min_{\beta} \frac{1}{2n} \|Y - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1 .$$

This estimator enforces a *sparse* representation of Y, that is, many coefficients will be exactly equal to zero. The penalty parameter λ is chosen by cross-validation, that is, we pick the λ that results in the lowest value of the cross-validated objective function. As discussed in section 2.2, cross validation guarantees the selection of a λ parameter that performs approximately as well as an ideal λ that recovers the true model. The performance is measured by the reduction in the squared loss from a classifier that guesses randomly, $0.25 - \frac{1}{n} \|Y - \mathbf{X}\beta\|_2^2$. We also report accuracy as a commonly used metric in the machine learning literature, which is the percentage of correctly predicted outcomes. We use the implementation in Pedregosa *et al.* (2011).

Logistic regression. Logistic regression is a model familiar to both computer scientists and economists. Using the two classes (sell/hold) as a binary outcome variable, we aim at maximizing a log-likelihood function on the training sample. Similar to lasso, we use L1-regularization in the logistic regression framework. Thus the objective function is

$$\min_{\beta} C \sum_{i=1}^{n} log(1 + e^{-y_i \beta^T x_i}) + ||\beta||_1.$$

We run cross validation on the training data and select the best C. Performance improvements are measured relative to the average log probability ($log(0.5) - \frac{1}{n} \sum_{i=1}^{n} log(1 + e^{-y_i \beta^T x_i})$). In addition, we report accuracy on the held-out test data. The estimations use the optimization procedure of Fan *et al.* (2008).

The decision tree algorithm. The decision tree algorithm is fairly distinct from the family of generalized linear classifiers. For several reasons, decision-tree learning lends itself very naturally to our problem. It is robust to monotone transformations of variables, it is easily scalable to large datasets and it deals relatively well with irrelevant variables.[13]

Decision tree learning tries to construct a decision tree from a given dataset by working top-down from a root node. The name stems from the fact that the results can be visualized as a decision tree that, at each node, splits the remaining observations into two branches, based on the value of a feature and based on some threshold value. Both the feature that is used for the split and the threshold need to be estimated. In general, multi-class classification is possible, but in our case, the decision tree algorithm needs to classify instances into only two categories, a selling and a holding category.[14] We define the best split to be the one that provides the largest *information gain* at a given node, that is, at each node the algorithm searches for the feature that is most discerning for the data at that node. A split is defined by a variable x^f and a value v such that observations that satisfy $x^f \leq v$ end up in one group and observations that satisfy $x^f > v$ end up in the other group. For example, $Gain(A(0,0)) \leq 0$ is one potential split at a node.

Starting with the root node, we estimate the decision tree with a greedy algorithm, i.e. we compute the information gain for all possible splits resulting from all features x^f and

[13] See (Hastie *et al.*, 2009, ch. 10) for an extensive discussion of the relative merits of different learning algorithms.

[14] For a more detailed description of the decision tree algorithm, see also chapter 1 of this dissertation.

values v, and at each node choose the combination of x^f and v that maximizes information gain.

We repeat the procedure on the next level of the tree, and stop when the number of observations in a node becomes smaller than a minimum threshold. The best minimum threshold is again selected by cross-validation. For the decision tree classifier, we simply use accuracy to measure the performance (noting that randomly guessing would result in an accuracy of 50%).

Results of inductive testing

The idea of inductive testing is that we give an algorithm access to many variables such that we can check whether performance changes significantly when we add variables that represent the theory to be tested (X_0). If significant improvement is made, it suggests that the theory is valid. In this particular example, we look at whether $Gain(A(0,0))$ affects the performance of the algorithm. We also formulate a more relaxed version of the disposition effect hypothesis, by comparing the performance of all the variables except the full set of $\{Gain(A(i,j))\}$ to the performance of all the variables including all of $\{Gain(A(i,j))\}$. Thus for ease of comparison, we show the performance in a reverse order, i.e., we show the performance using all variables and then the performance when we remove certain variables related to the theory.

Table 2.3 shows the performance of different variable sets using the Lasso algorithm. The first two rows show the performance using either the standard disposition variable only or all variables jointly. $Gain(A(0,0))$ by itself is useful in predicting traders' selling behavior: The accuracy is above 50% and the improvement in MSE is larger than 0. Not surprisingly, it is significantly worse than the performance using all features in the second row.

Following the inductive testing procedure, the third row computes the performance using all variables except for $Gain(A(0,0))$. Note that there is neither a significant difference in mean squared error nor in accuracy relative to the second row which includes all variables

Table 2.3: *Inductive testing using Lasso regression*

Variable set	MSE improvement	Accuracy
Gain(A(0, 0)) only	0.002***	0.541***
All variables	0.018	0.606
Remove Gain(A(0, 0))	0.018	0.606
Remove Gain(A(0, 0)) and all RefGain	0.018	0.606
Remove Gain(A(i, j)), $0 \leq$ i,j < 3 and all RefGain	0.018	0.606
Remove all Gain(A(i, j)) and all RefGain	0.018	0.606
Remove all Quartile	0.015***	0.591***

The table shows the performance on the held out testing data using regularized linear regression with an L^1 norm. The performance in column (1) is measured by mean squared error and in column (2) is measured by accuracy defined as the proportion of observations that the algorithm predicts correctly. In column (1), mean squared error is normalized by the mean squared error of a model that has no explanatory variables.
For each cell in the table, we test whether performance in that cell is significantly different from the cell that uses **all variables**. The p-value is computed by a permutation test.
*** $p < .001$, ** $p < .01$, * $p < .05$.

(and $Gain(A(0,0))$. A natural concern is whether this result might arise because other $Gain(A(i,j))$ variables that are close to $Gain(A(0,0)$ can stand in for that particular variable or because we might be considering too narrow a definition of the disposition effect.

However, when we remove all gain features on A-ranges within 3 days from the purchase day and the last day in the time series, that is, we remove all *Gain* variables with $A(i,j), 0 \leq i < 3, 0 \leq j < 3$, there are no significant performance difference either (row 4). Row 5 removes all *Gain* variables on the A-ranges, again having no effect on the prediction performance.

Thus, an inductive test provides evidence that the disposition effect does not matter for stock trading decisions. Again, we note that an inductive test cannot directly suggest new theories; rather, it can only suggest directions for future iterations of theory-development.

With such future iterations in mind, the last row of table 2.3 shows the performance after we remove all quartile variables. Here, the drop in MSE performance compared to the set of all variables is roughly 18% and it is significant.

Table 2.4 conducts the equivalent analysis for logistic regressions and decision trees. The exact same findings are also seen with these two methods, indicating that the conclusion is fairly robust to the particular learning method that is used.

Table 2.4: *Inductive testing using logistic regression and decision tree.*

	Logistic regression		Decision tree
Variable set	Improvement	Accuracy	Accuracy
Gain(A(0, 0))	0.003^{***}	0.541^{***}	0.541^{***}
All variables	0.038	0.606	0.626
Remove Gain(A(0, 0))	0.038	0.605	0.626
Remove Gain(A(0, 0)) and all RefGain	0.038	0.606	0.627
Remove Gain(A(i, j)), $0 \leq$ i,j < 3 and all RefGain	0.038	0.606	0.628
Remove all Gain(A(i, j)) and all RefGain	0.037	0.605	0.624
Remove all Quartile	0.031^{***}	0.592^{***}	0.601^{***}

The table shows the performance on the held out testing data using logistic regression regularized with an L^1 norm and with decision trees as described in the text. *Improvement* in column (1) is measured by improvement in the mean log probability as discussed in the text and accuracy in columns (2) and (3) is measured as the proportion of observations that the algorithm predicts correctly. P-values are computed by a bootstrap procedure analogous to the one in table 2.3.

*** $p < .001$, ** $p < .01$, * $p < .05$.

To summarize, although gain functions look individually valuable, they turn out not to be relevant when they are considered jointly with many other variables. The inductive tests reject the theory that disposition-related variables are relevant for the selling decision, and the results suggest that the gain functions are likely capturing other phenomena. While our results unambiguously support this conclusion, they do not give guidance as to *why* gain

functions do not pass the inductive test. In the next section, we move to a simpler version of the classification task to get more insight into the underlying forces that drive the results.

2.3.5 A simpler problem

The inductive tests in the previous section indicate that the disposition effect is not relevant to predicting selling actions. In this section, we use a smaller set of variables to better understand how the disposition effect's influence diminishes in the presence of selected other variables. We also try to shed light on the question which variables other than disposition effect - related ones can account for the prediction performance.

Table lookup

Many complex machine learning algorithms are designed to optimize predictions. When large sets of variables are used, the outputs of such algorithms are often difficult to interpret. By using a smaller set of variables, we can move to a more interpretable method that allows us to dig deeper into the underlying interplay between the different variables. When the variable set is small, we can compute the joint distribution of all variables' values and all selling decisions. For instance, suppose we only consider two variables, Gain(A(0, 0)) and Gain(A(0, 1)). We would then group instances by each possible combination of values of the two variables, i.e. $(\text{Gain}(A(0,0)) = i, \text{Gain}(A(0,1))) = j, i \in \{-1,0,1\}, j \in \{-1,0,1\}$, for a total of 9 different combinations. For each combination, we compute whether it is more likely to sell or to hold a security in the training data. This is then used to classify observations of the test data. We call this method *table lookup* since we literally have a table from which we learn the prediction for each possible pattern. Of course, this method is really only feasible when the set of variables is small, although the number of combinations can be relatively large.

The variable set that we consider is a small subset of the variables that we use in Section 2.3.4 and that we think is representative for each group of variables. We include variables and domains that convey both long- and short-run information about prices. In particular,

we define the following variables:

- *Gain*: This is Gain(A(0, 0)), the gain function operating on the entire domain of past prices. It is the original notion of gain and loss as in Odean (1998).
- $Trend_t$: This is Gain(B(0,1)). It describes how the price of today compares to yesterday's price.
- $Trend_{t-1}$: This is Gain(B(1,2)). It describes how the price yesterday compares to the price two days ago.
- $Trend_{t-2}$: This is Gain(B(2,3)). It describes how the price two days ago compares to the price three days ago.
- *Max*: This is Max(A(0, 0)). It describes whether today's price is greater than the maximum price during the holding period.
- *Min*: This is Min(A(0, 0)). It describes whether today's price is greater than the minimum price during the holding period.
- $Quartile_1$: This is Q1(A(0, 0)). It describes whether today's price is in the lowest quartile of the holding period.
- $Quartile_2$: This is Q2(A(0, 0)). It describes whether today's price is in the second-to-lowest quartile of the holding period.
- $Quartile_3$: This is Q3(A(0, 0)). It describes whether today's price is in the second-to-highest quartile of the holding period.
- $Quartile_4$: This is Q4(A(0, 0)). It describes whether today's price is in the highest quartile of the holding period.

Table 2.5 shows summary statistics for the subset of features that we use.

Table 2.5: *Summary statistics for variables in both tasks (mean and standard deviation).*

Variable	Classification	Game
Gain	0.211	0.171
	(0.970)	(0.980)
Quartile_1	0.281	0.291
	(0.450)	(0.454)
Quartile_2	0.135	0.154
	(0.342)	(0.361)
Quartile_3	0.154	0.169
	(0.361)	(0.375)
Quartile_4	0.430	0.386
	(0.495)	(0.487)
Trend_t	-0.050	-0.016
	(0.929)	(0.927)
Trend_{t-1}	-0.002	-0.016
	(0.931)	(0.927)
Trend_{t-2}	0.016	-0.016
	(0.929)	(0.927)
Max	-0.793	-0.879
	(0.589)	(0.461)
Min	0.884	0.927
	(0.443)	(0.352)

Variables are defined in the text. Gain, Trend_t, Trend_{t-1}, Trend_{t-2}, Max and Min can take on values in $\{-1, 0, 1\}$; Quartile_j can take on values in $\{0, 1\}$ for $j \in \{1, \ldots, 4\}$. Standard deviations are in parentheses.

A new prediction problem

In order to better understand the effects of different variables, we return to the full sample and we switch from the classification task to a new prediction game that we introduce now.

The payoff of the guessing *game* is defined in Table 2.6. Suppose you are presented with the price path of a security, starting at the buying price and then evolving for 250 trading days. If you were to bet on which day the investor sold his position (full or partial), on which day would you bet? The entries in table 2.6 display the payoff structure for that gamble. The payments are designed in such a way that a naive prediction that predicts holding on all days or selling on all days receives a payoff of zero. The parameter α determines the relative rewards for predicting selling versus predicting holding on to a security.[15]

In practice, the selling days are not evenly distributed over the time interval between 25 and 250. For instance, many more sales happen on day 25 than on day 249. In order to avoid biases towards long or short holding periods, we adjust the payoffs for betting on different days by correcting for their baseline probability $prob_t$. In other words, the potential reward for betting on day 25 being the selling day is lower than the potential reward for betting on day 249 being the selling day. A nice property of this game and table lookup is that the reward is additive for different feature combinations. For instance, the reward of betting on selling for $Gain = 1$ is independent of that for $Gain = -1$.

In the game, we randomly split all the time series into training data and held-out testing data. The entries for our lookup table are learned on the training data and the variables are then evaluated on the test data. The training data contain 145,705 time series, while the held out testing data contain 29,141 time series.

Results

We replicate the inductive tests with the smaller set of variables using table lookup. A 5-fold cross validation on the training data is used to find the set of variables that provides the best performance.[16]

As in section 2.3.4, our inductive tests compare the best performance using all variables

[15]Throughout this section, we set $\alpha = 1$, that is, one can only get a reward by predicting selling correctly. We confirmed in unreported computations that the choice of α does not affect the main results.

[16]Note that because of overfitting (fitting training data too well to generalize on testing data), the best set may not be the full set that uses all features.

Table 2.6: *Payoff matrix.*

	Realized sell	Realized hold
Predicted sell	$\alpha(\frac{1}{prob_t} - 1)$	$-\alpha$
Predicted hold	$-(1-\alpha)$	$(1-\alpha)(\frac{1}{1-prob_t} - 1)$

This table shows the payoff matrix for the prediction game introduced in section 2.3.5. Each cell shows the payoff for a predicting sell/hold-realized sell/hold combination. $prob_t$ is the probability to sell on day t. The parameter α determines the relative rewards for predicting selling versus predicting holding on to a security.

to the best performance when a particular variable is not used. Note that we use all available data in this section and not only the balanced subsample of section 2.3.4.

Table 2.7 shows the results for the *Game* task. Column 2 reports, in each row, the reward that could have been achieved by using only the row variable(s) to make predictions. Column 3 reports the reward that could have been achieved by using all variables *except* for the row variable(s). All variables are individually important as is apparent from column 2. But when we sequentially exclude variables from the set of all variables, a different picture emerges: In some cases, the performance deteriorates significantly, while in other cases performance is completely unaffected. Excluding the Gain variable does not have any significant effect on the performance. It seems that, even within this small set of variables, Gain is not very helpful for making predictions. The last two rows show results when we exclude groups of features that relate to short-term movements. In both cases, performance is significantly worse, indicating the importance of recent price movements for the selling decision. We also see a complementary effect, meaning that the reward of using Trend_t and Trend_{t-1} jointly is better than the combined reward of using Trend_t or Trend_{t-1} separately.

Discussion

Table 2.7 shows that Gain is not a key determinant of selling, even when only a few other variables are considered. With this smaller variable set, we are able to dig deeper into the

Table 2.7: *Performance using table lookup in the Game task*

Variable(s)	Using only row	Using all but row
All	31.973	
Trend_t	8.651***	26.638***
Trend_{t-1}	4.307***	29.618***
Trend_{t-2}	5.480***	32.178
Gain	11.950***	31.973
Max	13.191***	31.286*
Min	4.840***	30.900**
Quartile_4	21.575***	28.943***
Trend_t, Trend_{t-1}	15.975***	26.660***
Trend_t, Trend_{t-1}, Trend_{t-2}	18.810***	26.415***

The first row shows the reward that the algorithm achieves on the testing data using all variables. Column 2 reports, in each row, the reward that could have been achieved by using only the row variable(s) to make predictions. Column 3 reports the reward that could have been achieved by using all variables *except* for the row variable(s). P-values are computed by a bootstrap procedure analogous to the one in table 2.3.

*** $p < .001$, ** $p < .01$, * $p < .05$.

underlying mechanism. Table 2.7 indicates that short-term price movements and quartile variables are very relevant to predict the selling decision.

In this section, we investigate in more detail the relative relevance and the interplay of short-term price movements and quartile features.

Focusing on Trend_t and Trend_{t-1} and all quartile variables, there are 16 different ways in which these variables can be combined (ignoring Trend_t or Trend_{t-1} being equal to zero). For instance, one combination would be two price increases over the last two days and the price at day t being in the highest quartile. This combination would require $\text{Trend}_t = 1$, $\text{Trend}_{t-1} = 1$ and $\text{Quartile}_4 = 1$. We find all observations in the data that obey this pattern

and compute the average reward that one could earn based on betting on that pattern.

Table 2.8 shows all the possible variable-value combinations of Trend_t, Trend_{t-1}, and Quartile_j. The top pattern is the one that we just discussed. Observing this pattern, one should predict a selling decision for an average reward of 11.986. The remaining columns answer a thought experiment: How would the prediction change if we altered only one of the variables' values? For the top pattern, we see that changing the sign of the short-term price variable Trend_{t-1} changes the prediction from selling to holding. Moreover, $\text{Quartile}_4 = 1$ is essential in this pattern; if we change the price to be in any other quartile of the price distribution, the predicted result flips. Another interesting thing about the first row is that flipping Trend_t does not affect the prediction result, but flipping Trend_{t-1} does. This suggests that if the price goes up yesterday, and is right now in the top quartile of the history, people are more likely to sell no matter whether the price is going up or down today.

More generally, the table allows us to investigate which variables are essential or pivotal to a pattern's prediction. Note that many predictions flip depending on whether or not the price is in the highest quartile of the price distribution. In addition, note that flipping short-term price movements often results in opposite predictions, too. This suggests that these variables play a pivotal role in informing an investor's selling decision.

A failure of exclusion

When we think about a particular variable X_0, we have thus far implicitly acted as if our assumptions about X_0 from Section 2.2 fully held. Often, however, this is hard to tell in practice. For instance, one special case that often occurs in empirical work is what we might call the *clone problem*: Suppose we have an army of clones of a useful feature, and we observe that removing the useful feature does not have a significant effect on some performance measure. According to our inductive tests, we may think that this feature is not useful. People have tried to avoid this problem by proposing features that are almost "orthogonal".

Here we propose the beginnings of an alternate solution to this problem; we highlight

some results that seem intriguing and suggest the need for a more complete formalization. Suppose the variable that we are interested is f, and the overall performance is P. We will introduce a new task in which the goal is to predict f; we will denote the performance on this prediction task by P_f. The idea is to see how the overall performance P changes as a function of P_f. Now let us see how this helps solve the clone problem. If there is a clone of variable f, causing us to find that removing f does not degrade P, then we will also see that P_f is high in the subset of variables with f deleted. Thus if there is a clone problem, we will see that P increases as P_f increases. On the other hand, if P does not increase as P_f increases, it shows an even stronger consequence of the inductive test: a better understanding of f does not help our prediction task at all.

In the particular problem of optimizing performance in the game we are studying here, because we have a small number of variables, we can enumerate every possible variable subset, and plot the performance in the game vs the accuracy of predicting a given fixed variable. Figure 2.1 shows the scatter-plot of the performance in the game vs the accuracy of predicting three specific features: Disposition (Gain), q4 (Quartile$_4$), and t0 (Trend$_t$). Each point in the scatter-plot corresponds to a single subset of the features, representing this subset's accuracy in predicting disposition, q4, or t0 on the x-axis, and its reward in the game on the y-axis. Note, for interpreting these scatter-plots, that whenever the feature on the x-axis is included in a subset of the features, the accuracy for predicting that feature becomes 1 by definition, and so the maximum x-value over the points in the scatter-plot is always 1.

As we can see in the figures, the reward from the game gets saturated as the accuracy of predicting disposition increases, which serves as an indication of the ineffectiveness of disposition as a prediction variable. For t0 and q4, on the other hand, the increase in the performance in the game is almost linear with the increase in accuracy on the x-axis. Note that the structure of the scatter-plots for t0 and q4 are actually quite different from one another. For t0, it is very hard to predict the value of t0 from any set of other variables that does not include it (consistent with the notion that stock price changes are difficult to

forecast), while for q4, we can see that as the accuracy increases, there is a consistent linear trend of increase, which validates the effectiveness of q4.

2.4 Conclusion

The emergence of powerful machine-learning techniques has made it possible to perform robust prediction in settings where the number of variables is very large relative to the number of observations. Here we have shown how these techniques can be adapted to provide methods for testing theories. Instead of traditional *deductive tests*, which can only control for known theories, we develop *inductive tests* that make use of large variable sets and can be interpreted as implicitly controlling for theories that have not yet been formulated, provided only that they are sufficiently covered by the available variables. We apply this inductive testing methodology to analyze the disposition effect, a long-standing issue in behavioral finance; our inductive tests have revealed potential limitations of the disposition effect in ways that escape the power of standard deductive tests.

This work suggests a number of directions for further investigation, of which we highlight two here. First, we have constructed the inductive testing framework so that it can be applied quite broadly, ideally with relatively little adaption from one domain to another. Our understanding of inductive testing would clearly benefit from further theoretical development at a general level beyond the specifics of any one particular machine learning technique, since the underlying principles seem clearly to operate at this more general level. But the formulation we give here already suggests some of the power of the approach for basic questions in the social sciences, and it would be interesting and arguably relatively straightforward to apply this framework to other problems where it is possible to evaluate theories in light of large feature sets.

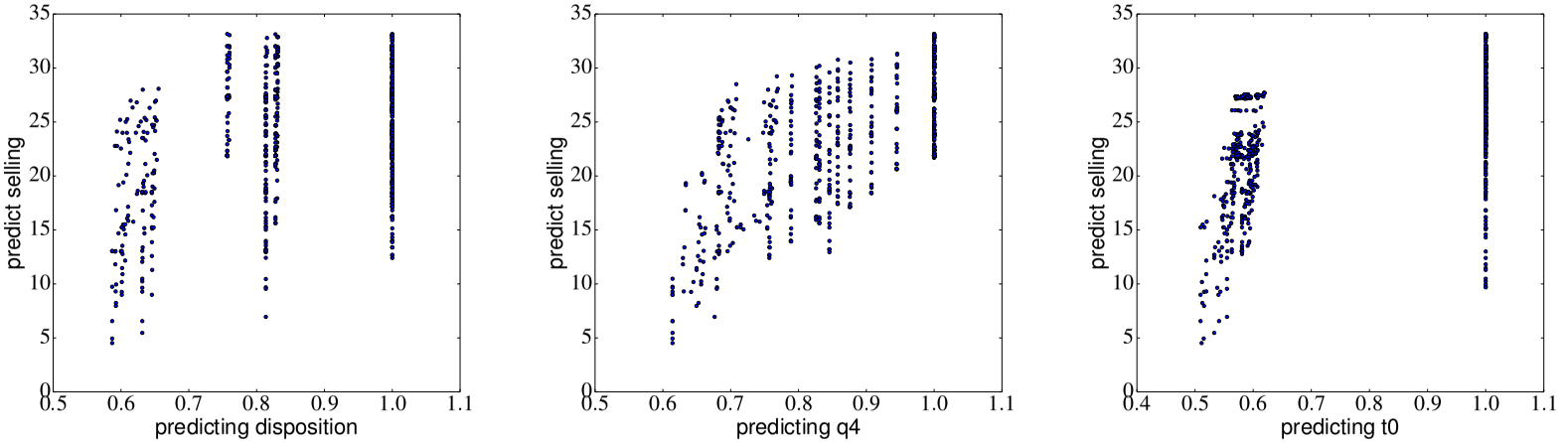
Second, our focus here is on tests to accept or reject a known theory M_0; as we have noted at several points, the inductive testing framework does not directly give guidance in the formulation of new competing theories. But it would be interesting to consider whether techniques from machine learning could be useful for this latter task — the problem of

generating new theories from large variable sets. This leads to a range of further questions that could help deepen our understanding of the trade-offs between deduction and induction as modes of investigation and evaluation in these areas.

Table 2.8: *Top patterns depending on Trend_t, Trend_{t-1} and Quartile functions.*

Pattern	Average reward	Prediction	Flip Trend_t	Flip Trend_{t-1}	Q1	Q2	Q3	Q4
$\text{Trend}_t = 1$, $\text{Trend}_{t-1} = 1$, Q4	11.986	sell	sell	hold	hold	hold	hold	sell
$\text{Trend}_t = -1$, $\text{Trend}_{t-1} = -1$, Q4	4.142	sell	hold	sell	sell	sell	sell	sell
$\text{Trend}_t = -1$, $\text{Trend}_{t-1} = -1$, Q1	4.098	sell	hold	hold	sell	sell	sell	sell
$\text{Trend}_t = -1$, $\text{Trend}_{t-1} = 1$, Q4	3.896	sell	sell	sell	hold	hold	hold	sell
$\text{Trend}_t = -1$, $\text{Trend}_{t-1} = -1$, Q3	1.385	sell	hold	hold	sell	sell	sell	sell
$\text{Trend}_t = -1$, $\text{Trend}_{t-1} = -1$, Q2	0.412	sell	hold	hold	sell	sell	sell	sell
$\text{Trend}_t = -1$, $\text{Trend}_{t-1} = 1$, Q3	-1.208	hold	hold	sell	hold	hold	hold	sell
$\text{Trend}_t = -1$, $\text{Trend}_{t-1} = 1$, Q2	-1.424	hold	hold	sell	hold	hold	hold	sell
$\text{Trend}_t = 1$, $\text{Trend}_{t-1} = 1$, Q3	-1.552	hold	hold	hold	hold	hold	hold	sell
$\text{Trend}_t = -1$, $\text{Trend}_{t-1} = 1$, Q1	-1.915	hold	hold	sell	hold	hold	hold	sell
$\text{Trend}_t = 1$, $\text{Trend}_{t-1} = -1$, Q4	-1.994	hold	sell	sell	hold	hold	hold	hold
$\text{Trend}_t = 1$, $\text{Trend}_{t-1} = 1$, Q2	-2.170	hold	hold	hold	hold	hold	hold	sell
$\text{Trend}_t = 1$, $\text{Trend}_{t-1} = 1$, Q1	-2.547	hold	hold	hold	hold	hold	hold	sell
$\text{Trend}_t = 1$, $\text{Trend}_{t-1} = -1$, Q3	-3.067	hold	sell	hold	hold	hold	hold	hold
$\text{Trend}_t = 1$, $\text{Trend}_{t-1} = -1$, Q2	-3.282	hold	sell	hold	hold	hold	hold	hold
$\text{Trend}_t = 1$, $\text{Trend}_{t-1} = -1$, Q1	-4.027	hold	sell	hold	hold	hold	hold	hold

The table shows all possible patterns involving Trend_t, Trend_{t-1} and the quartile functions. Qj is short for $\text{Quartile}_j = 1$. We exclude patterns with $\text{Trend}_t = 0$ or $\text{Trend}_{t-1} = 0$. The first column show the pattern, the third column shows the predicted result, and the second column shows the reward in the game. Columns 4-9 shows the predicted result if we flip the value of Trend_t or Trend_{t-1} or if we change a price's place in the price distribution over the holding period.



(a) *Reward in the game vs. predicting disposition*

(b) *Reward in the game vs. predicting q4*

(c) *Reward in the game vs. predicting t0*

Figure 2.1: *Scatterplot of reward in the game vs. predicting a given variable.*

Chapter 3

The disposition effect and the size of returns: Reconciling evidence from individual investors[1]

3.1 Introduction

A long-standing puzzle in finance research is investors' tendency to hold losing investments too long and to sell winning investments too soon, a phenomenon that has been coined the "disposition effect" by Shefrin and Statman (1985). In a careful study, Odean (1998) brings the effect to individual trading data, and rules out various competing explanations. The disposition effect is persistent, even when controlling for higher transaction costs of selling losers or tax incentives.[2] Other explanations, such as re-diversification or private information, similarly fail to capture important features of the data.

Several theories have been proposed to explain the disposition effect. The most prominent explanation invokes prospect theory, developed by Kahneman and Tversky (1979). Odean (1998) argues that with reference-dependent preferences, investors are risk-averse over gains

[1] Co-authored with Frank Schilbach

[2] Ivkovic *et al.* (2005) show that the disposition effect interferes with a "lock-in effect" for capital gains.

and risk-seeking over losses, such that they sell winning stocks prematurely and gamble on stocks that lost value in the past.[3]

However, Barberis and Xiong (2009) challenge this informal argument by presenting a stylized asset pricing model involving investors with prospect theory preferences.[4] This model does *not* predict a disposition effect for many reasonable parameter values, since loss aversion pushes investors to only purchase stocks with high enough expected returns. In particular, if an investor is in the gain region, she will in expectation be further away from the reference point than when she is in the loss region, such that a disposition effect may not arise for mild curvature of the value function. In a related paper, Barberis and Xiong (2012) introduce the concept of *realization utility* and apply it to the disposition effect. In their model, investors derive utility from realizing gains and losses, implying that an investor only sells a stock once the return exceeds a certain, potentially investor-specific, threshold.

The main contribution of this paper is to contrast these competing explanations based on their predictions for realizing different *sizes* of gains and losses. We use the Barber and Odean (2000) data to establish two empirical facts. First, for all holding periods longer than one month, and, for both gains and losses, the probability to sell a stock declines monotonically in the size of the absolute return. That is, individual investors are not only more likely to sell gains than to sell losses, but they are also more likely to sell stocks with small absolute returns (i.e. stocks with prices close to the purchase price) than to sell stocks with large absolute returns. Second, the disposition effect is more pronounced for relatively high returns, i.e. the difference in selling probability between gains and losses of similar magnitude is more pronounced for large returns (i.e. comparing large gains to large losses) than for small returns (i.e. comparing small gains to small losses). Theories that attempt

[3]Odean (1998) notes that an irrational belief in mean reversion of stock returns could also explain the disposition effect, but he is not able to separate the two hypotheses. He speculates that investors themselves might not make a clear distinction: "For example, an investor who will not sell a stock for a loss might convince himself that the stock is likely to bounce back rather than admit his unwillingness to accept a loss." Weber and Camerer (1998) control for individuals' beliefs in an experimental setting and find the disposition effect as well.

[4]Barberis and Xiong (2009) do not include probability weighting.

to explain the disposition effect also need to be consistent with these additional facts. In particular, we argue that our first fact is not consistent with the model of realization utility as proposed by Barberis and Xiong (2012), but it is consistent with a version of prospect theory that we outline below.

Our empirical analysis employs two methodologies. First, in our preferred approach, we follow Ivkovic *et al.* (2005) and construct portfolios of investors' stock holdings for each month. This method allows us to measure the survival time of stocks in investors' portfolios, conditional on the stock being in the gain or in the loss region. We extend this approach by splitting up the sample into gain and loss quantiles, and then consider the selling patterns across quantiles, controlling for stock holding periods, which generates the results described above.

Second, following the original estimation approach of Odean (1998), we construct an investor's portfolio for every day at which an investor made at least one trade. Replicating Odean's results, we compute the fraction of stocks valued at a gain (loss) that were actually sold relative to all stocks that could have been sold at a gain (loss). Extending this approach to stock returns of different magnitudes, again controlling for the holding period, we find that the disposition effect persists for all return sizes. However, in contrast to the results from our first approach, the selling probability *increases* in the size of the return for gains and is constant in returns for losses.

We trace the apparent tension between results from the two approaches to different conditioning sets of the estimates. The duration model of Ivkovic *et al.* (2005) computes an *unconditional* probability of selling (for given holding periods), while the Odean (1998) methology estimates a probability of selling a stock *conditional on investor activity*. To reconcile the results from the two estimation approaches, we establish that an investor's propensity to make a trade is largest for small absolute portfolio returns, using several parametric and non-parametric estimators.

Our paper is closely related to a recent contribution by Ben-David and Hirshleifer (2012), who investigate the relationship between past security returns and subsequent sales using

an approach similar to ours. Ben-David and Hirshleifer find that the probability to sell a security is “asymmetrically V-shaped". That is, larger returns are more likely to be sold, this effect is more pronounced for positive returns compared to negative returns, and the selling probability does *not* have a discontinuity at a stock return of 0. Since these results appear to be in contrast to our findings, we reconcile these findings with our results. We document that the probability to sell is asymmetrically V-shaped only for short holding periods. In particular, if results are pooled over different horizons, we do find a pronounced discontinuity at 0 returns. We also find that the selling probability decreases in the absolute value of the return, in line with our previous findings.

The remainder of the paper is organized as follows. Section 3.2 derives simple theoretical predictions for prospect theory and realization utility. Section 3.3 describes the data and our methodological approach. In section 3.4 we present our main empirical results and robustness checks, and in section 3.5 we relate our results to Ben-David and Hirshleifer (2012). Section 3.6 concludes.

3.2 Theoretical background

In this section, we provide a sketch of models of realization utility and of prospect theory that have been suggested as explanations for the disposition effect. We are particularly interested in the predictions of these models for the probability to sell a stock for different returns, and we show that the models we consider generate quite different predictions in this regard. This allows us to disentangle the explanations by comparing their predictions to the actual probability to sell a stock for different returns.

3.2.1 Realization Utility

Barberis and Xiong (2012) argue that investors derive utility from *realizing* gains and losses of assets that they own (as opposed to from consumption of the proceeds). The authors set up a dynamic optimization problem and show that investors with realization utility will sell a stock when the return exceeds some (positive) *liquidation point*. In other words, "if the

investor buys a stock, he voluntarily sells it only if its price rises a sufficient amount above the purchase price". In particular, an investor only sells at a loss if he is forced to do so by a liquidity shock.

For illustrative purposes denote by P_0 the investor's purchase price and by P_t the price in period t. Define $g_t := P_t/P_0$. An investor will then sell the stock if $g_t \geq g^i_*$, where $g^i_* \geq 1$ is the investor's liquidation point which depends on individual and stock characteristics.[5] Figure 3.1 shows the trading behavior of the individual investor depending on g_t. The probability to sell a stock below g_t is drawn to be greater than zero to take liquidity shocks into account. If g_t exceeds the investor's liquidation point, the stock is sold in t. While realization utility predicts a step function for the individual's probability to sell a stock, the function is smoothed out when we consider the aggregated prediction. Since the liquidation point depends on individual characteristics,[6] there will be heterogeneity of liquidation points across investors.

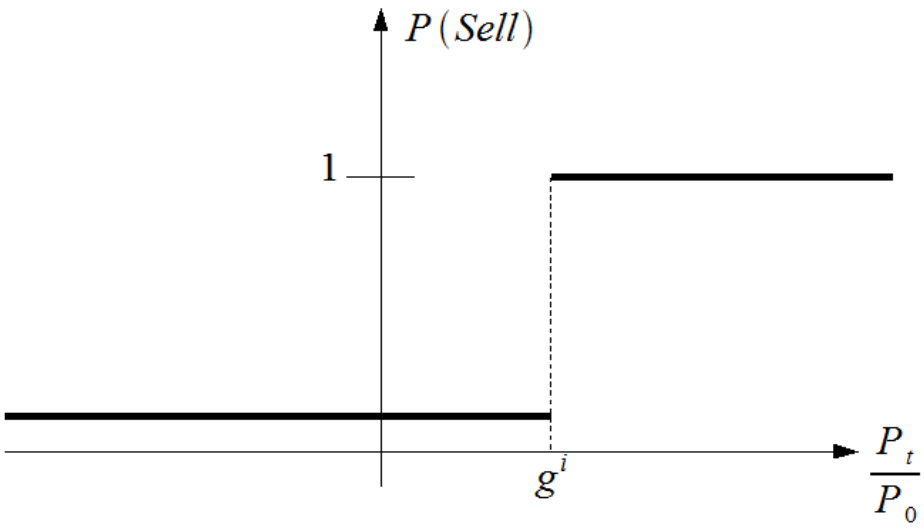


Figure 3.1: *Realization utility prediction for individual trading*

Figure 3.2 illustrates the implications for the aggregate probability that a stock is sold.

[5] g^i_* is strictly greater than 1 if the investor faces some transaction cost.

[6] It depends, for instance, on the time discount rate, the transaction cost, and the likelihood of a liquidity shock, which might all vary across investors.

As investors do not realize losses, the probability to sell is constant in the loss region (liquidity shocks). When the return is slightly positive, some investors start selling the asset because they have a relatively low liquidation point. A higher return implies that additional liquidation points are exceeded. The exact shape of the function obviously depends on the distribution of liquidation points among investors. Of course, there is no reason to expect this relationship to be linear[7], but the probability to sell should certainly be an increasing function of the realized return.

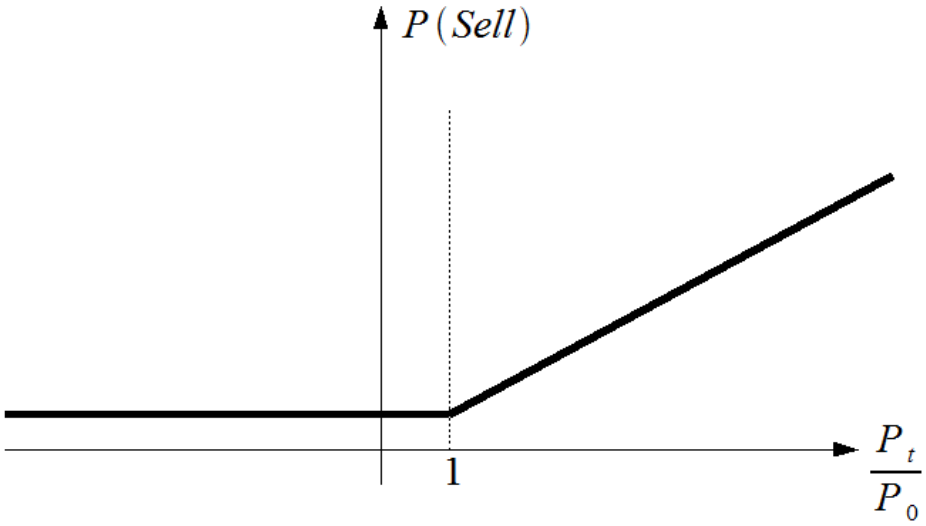


Figure 3.2: *Implied trading pattern in the realization utility model*

3.2.2 Prospect Theory

In their pioneering work, Shefrin and Statman (1985) and Odean (1998) linked prospect theory to the disposition effect using informal arguments. In a formal analysis, however, Barberis and Xiong (2009) find that a model including investors with prospect theory preferences[8] does not generate the disposition effect for many combinations of parameter

[7] e.g. a positively skewed distribution (many observations close to 1) would yield a concave function

[8] The authors abstract from probability weighting.

values.[9] Prior arguments had neglected the effect of the kink on the initial buying decision. Because of the kink, an investor will only buy a stock initially if the expected return is high enough. That implies that, in the next period, the investor is relatively far away from the reference point when the stock trades in the gain region, whereas she is closer to the reference point when the stock is in the loss domain. When the value function has only mild curvature, the investor is almost risk-neutral in the gain region when she is relatively far away from the reference point and she will therefore hold the stock after a gain. On the other hand, after a loss, she is still close to the reference point and sells the stock for many parameter values of the value function. Meng (2013) argues that a modified reference point can help to get around this result. Using expectations as reference point, she proposes a simple model in which prospect theory generates a disposition effect for investors.

We present two views of prospect theory both of which are able to generate the disposition effect. First, we consider Barberis (2012) implementation of prospect theory in a model of casino gambling. This model, very different from most other models in this area of research, takes nonlinear probability weighting into account. It features some interesting predictions that we think are transferable to a model of individual investment decisions.

Consider an asset that in every period with probability 0.5 either increases or decreases by h. Figure 3.3 shows the possible prices after six periods. Barberis' model implies that even though this stock has an expected return of 0, a prospect theory agent might buy it because she can give its return a favorably skewed distribution by overweighting small probabilities and by choosing a suitable ex-ante exit strategy (e.g. sell the stock as soon as you acquire losses). Note that this is in stark contrast to Barberis and Xiong (2012)' s model which implies that prospect theory agents only buy stocks with high expected return. Barberis (2012) then investigates implications for subsequent gambling behavior. He finds that naive prospect theory agents (that is, those who are not aware of their nonlinear probability weighting) almost never exit after making losses and stop gambling too early after making gains; in other words, they do not stick to their original plans. The actual exit

[9]This has also been studied by Hens and Vlcek (2011).

pattern is illustrated by the curved line in figure 3.3. Intuitively, the pattern comes from a trade-off between the nonlinear probability weighting (which pulls towards keeping on gambling, but less so if the gain is already large) and the concavity of the utility function in the gain region.

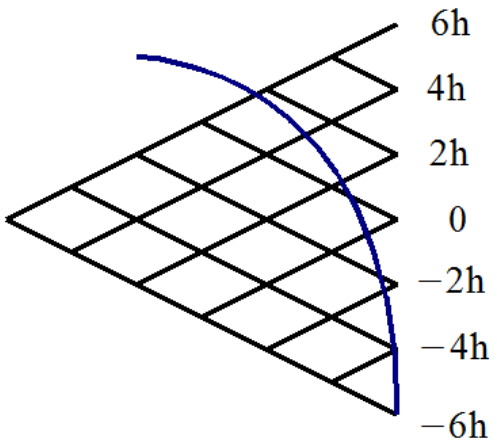


Figure 3.3: *Stock price for six periods and resulting exit behavior*

Carrying this result over to the stock market, it implies that higher positive returns are more likely to be realized. Figure 3.4 shows the trading pattern that can be derived from the model. In particular, investors never realize losses under this specification, but are increasingly willing to realize gains.

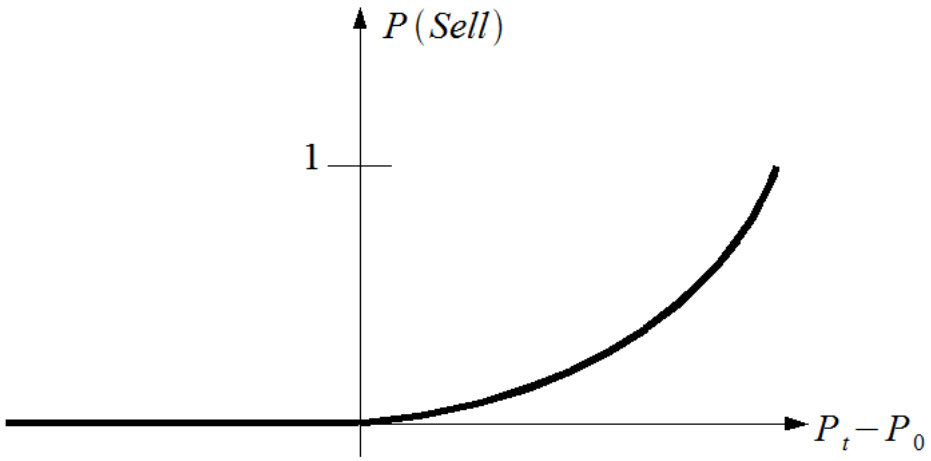


Figure 3.4: *Implied trading pattern in the casino gambling model*

A different, quite natural prediction arising from prospect theory focuses on the kink in the value function. As is well-known (see e.g. Barberis *et al.* (2006)), the kink induces *first-order risk aversion* around the reference point (Figure 3.5 illustrates this). Even though investors are risk-seeking in the loss domain, the kink induces them to reject a fair bet involving small amounts if the price is close to the reference price.[10] Being far away from the reference return, on the other hand, implies only mild curvature of the utility function (i.e. mild concavity/convexity) and, therefore, very similar behavior in these regions.

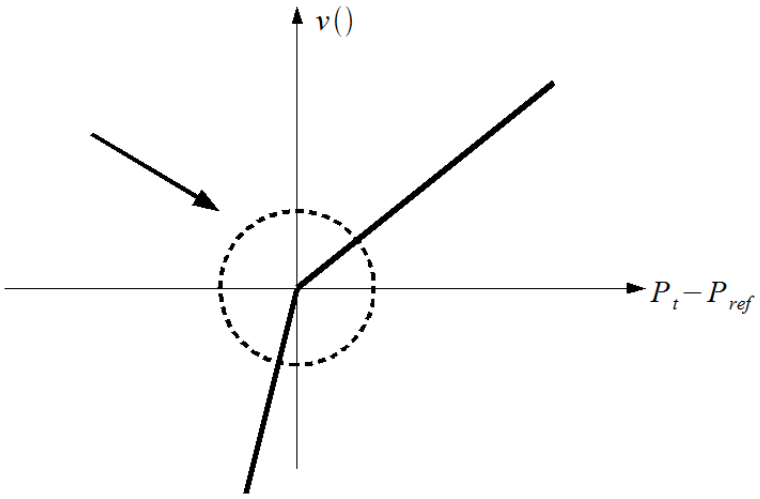


Figure 3.5: *Local risk aversion in the value function*

This gives rise to the prediction that the probability to sell will be higher around the reference point, a phenomenon that we label *bunching*, and lower for larger gains and larger losses. If, in addition, we assume that some degree of concavity/convexity away from the reference point is preserved, we can also conclude that the probability to sell gains is generally higher than the probability to sell losses (risk aversion vs. risk loving). This is summarized in figure 3.6. Note that the plot is qualitatively similar to Kaustia (2010) or Meng (2013) who develop models along these lines.

[10]Note that a fundamental problem of this explanation is the question of why they would buy stocks in the first place.

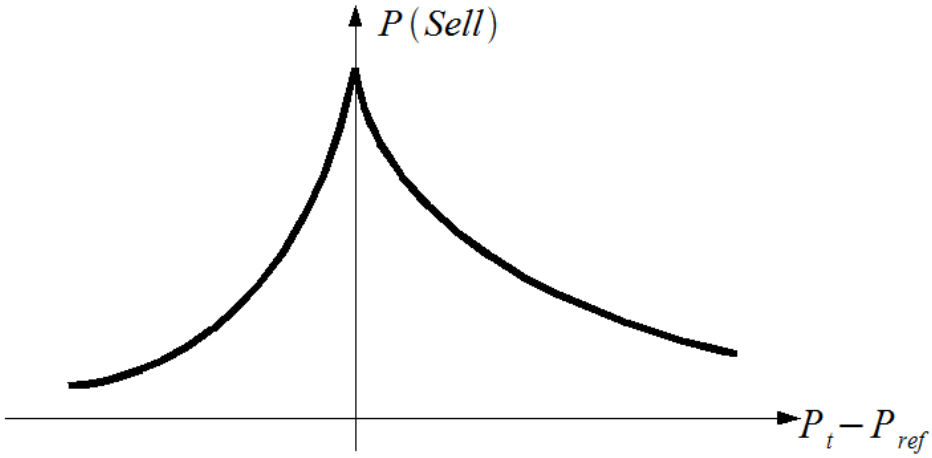


Figure 3.6: *Implied trading pattern in the bunching model*

3.3 Data and Methodology

In this section, we describe the data that we use for our analysis. We then describe two different approaches to estimation of the disposition effect that have been used in the related literature. Section 3.3.2 discusses the duration model approach of Ivkovic *et al.* (2005) and how it can be applied in our context. Section 3.3.3 describes the original approach of Odean (1998). We apply both approaches and reconcile their results later in section 3.4.4.

3.3.1 Data

Terrance Odean kindly provided us with the dataset used in Barber and Odean (2000), which is very similar to the one used in Odean (1998). This dataset from a large discount brokerage house contains all trades as well as end-of-the-month positions for 158,034 US accounts (belonging to 78,000 households) for the time period 1991-1996. Among other variables, the data comprises household and account identifiers, dates, selling and purchase prices, quantities and security identifiers.[11] The data also feature some demographic information that we use to control for investor characteristics in robustness checks of our analysis below. Table 3.1 shows the the main demographic variables used in our analysis. Our data contain

[11] A detailed description of the data can be found in Barber and Odean (2000).

relatively few female individuals, more than 50% of the individuals say they have some knowledge about the stock market (self-reported), and 15% of the households are labeled as frequent traders by the brokerage house (i.e. they trade more than 48 times per year).

Table 3.1: *Demographic characteristics of investors*

Variable	Obs	Mean	Std	Min	Max
Married	37642	0.78	0.41	0	1
Female	47586	0.13	0.34	0	1
Age	41654	50.63	12.76	18	94
Home owner	54914	0.77	0.42	0	1
Knowledge	27179	0.56	0.5	0	1
Equity ('000s)	77981	53	277	-.97	51900
Frequent trader	77984	0.15	0.36	0	1
Taxable account	77984	0.63	0.48	0	1

We match the trades file with specific information for each account and trade (e.g. account type, trading activity, product type). We get monthly and daily price data of securities from the Center for Research in Security Prices (CRSP). Prior to our analysis, we eliminate all trades other than trades of common stocks (e.g. foreign stocks), all trades that involve short-selling and all trades including securities purchased before 1991. Also, we drop all observations for which price data are not available and accounts that own only one stock. This procedure closely follows Odean (1998). To get a sense of the data, and because it is going to be important for our subsequent analysis, figure 3.7 shows the histogram of security holding durations in the sample. As one would expect, a lot of stocks are being held for a short period of time, and the number of observations declines monotonically in

the holding duration.

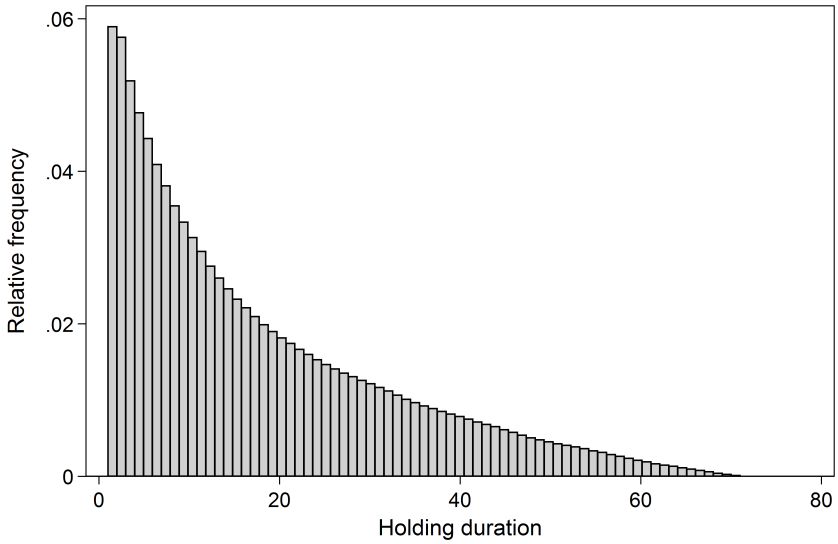


Figure 3.7: *Histogram of stock holding durations*

3.3.2 Duration model

Ivkovic *et al.* (2005) expand the same dataset that we use by including every month between an initial purchase of a stock and the first sale of the stock (or the end of the observation period, if the security had not been sold by then). For every month they match price data to their dataset and determine whether the stock was sold/could have been sold for a gain or a loss by comparing the buying price to the daily price, just like in Odean (1998). For a given duration they calculate the hazard rate for both gains and losses (i.e. the share of gains/losses that are sold). The resulting figure plots the frequency of selling (and leaving the sample) given that an investor has not sold the security yet, conditional on whether it is a gain or a loss (relative to the buying price), an estimator that is commonly referred to as the nonparametric Kaplan-Meier estimator of the hazard rate.[12]

[12]The estimate of the hazard rate is given by

$$\hat{\lambda}(t_k) = \frac{s_k}{n_k}$$

We take this analysis further and, for every holding period, compute the hazard rate as a function of return magnitude. For illustrative purposes, we show a graph with 3 quantiles of returns for all holding periods (small, medium and large) below, but due to the large sample we can focus on a much finer grid in our analysis when we keep the holding period fixed.

As in Ivkovic *et al.* (2005), we then turn to the estimation of a Cox-proportional hazard model of the form

$$\lambda(t, x_i) = \lambda_0(t) exp(x_i'\beta), \tag{3.1}$$

where $\lambda_0(t)$ is the baseline hazard rate and x_i contains individual-specific information which will allow us to control for various stock-holder (and stock) characteristics, such as those presented in table 3.1.

3.3.3 Proportion of realized gains and losses - The Odean (1998) approach

Odean (1998) does not simply compare realized gains and losses[13] but constructs a more sophisticated measure. He calculates portfolios for each account at each trading date by adding up trading records in chronological order. Every time a sale takes place, he compares the average buying price[14] to the selling price. An observation counts as a *realized gain* if the selling price is higher than the average buying price. It counts as a *realized loss* if the average buying price exceeds the selling price. The observation is omitted if the prices are equal. *Paper gains* and *paper losses* are defined similarly. Consider, for example, a security that is in an investor's portfolio at the beginning of the day, and it is not sold. The observation counts as a paper gain if the average buying price is lower than both the high and low price on that day. It is omitted if the average buying price lies between the high and low price for

where n_k is the number of observations in the sample in period t_k and s_k is the number of observations that leaves the sample in t_k. Both numbers can be conditioned on gains and losses.

[13]This would yield spurious results in an upward-trending market.

[14]i.e. the average price of a security for all purchases up to that date

the day. Paper losses are defined equivalently. Odean (1998) then computes the proportion of realized gains of all gains(PGR):

$$PGR = \frac{\text{\# of realized gains}}{\text{\# of realized gains + \# of paper gains}} \tag{3.2}$$

The proportion of realized losses (PLR) can be calculated in the same way. Odean then tests for the presence of a disposition by testing whether PGR>PLR.

The original implementation of the approach suffers from a bias which comes from the differential treatment of paper gains/losses relative to realized gains/losses. Note that paper gains/losses are only counted when the average buying price is less than/exceeds both the high and low price of that day, otherwise they are not counted. Realized gains and losses, on the other hand, are determined relative to the actual selling price regardless of whether the average buying price lies within the daily high and low price of that stock. Since it is more likely that small returns are between the daily high and low prices, this procedure systematically overstates small realized gains/losses relative to small paper gains/losses. Figure 3.8 illustrates this. The lower panel shows an observation that counts as a realized gain although $P_b \in [low, high]$, whereas the observation would not have been considered a paper gain had it not been sold (upper panel).

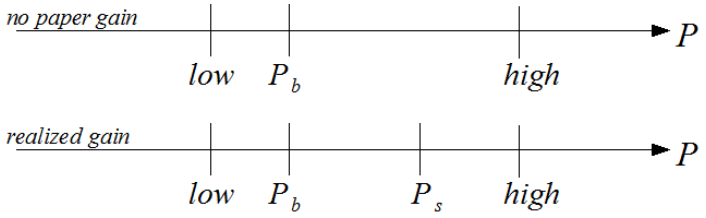


Figure 3.8: *Bias towards realized returns in the Odean approach*

This bias can, of course, be avoided by applying the same rule to both paper gains/losses and realized gains/losses (that is, by dropping realized gains when their average buying price falls in the interval between the low and high price of that day), a convention that we follow throughout our analysis. While we think that our approach is more rigorous than the original treatment of gains and losses in Odean (1998), we have confirmed that it does

not have a substantial effect on any of the results.

3.4 Results

Our main analysis consists of evaluating the duration model Kaplan-Meier estimates for different returns, fixing the holding period. But first, we show how the main results can be seen in simple histograms. In section 3.4.3, we present results using the methodology of Odean (1998).

3.4.1 A first look at the data using histograms

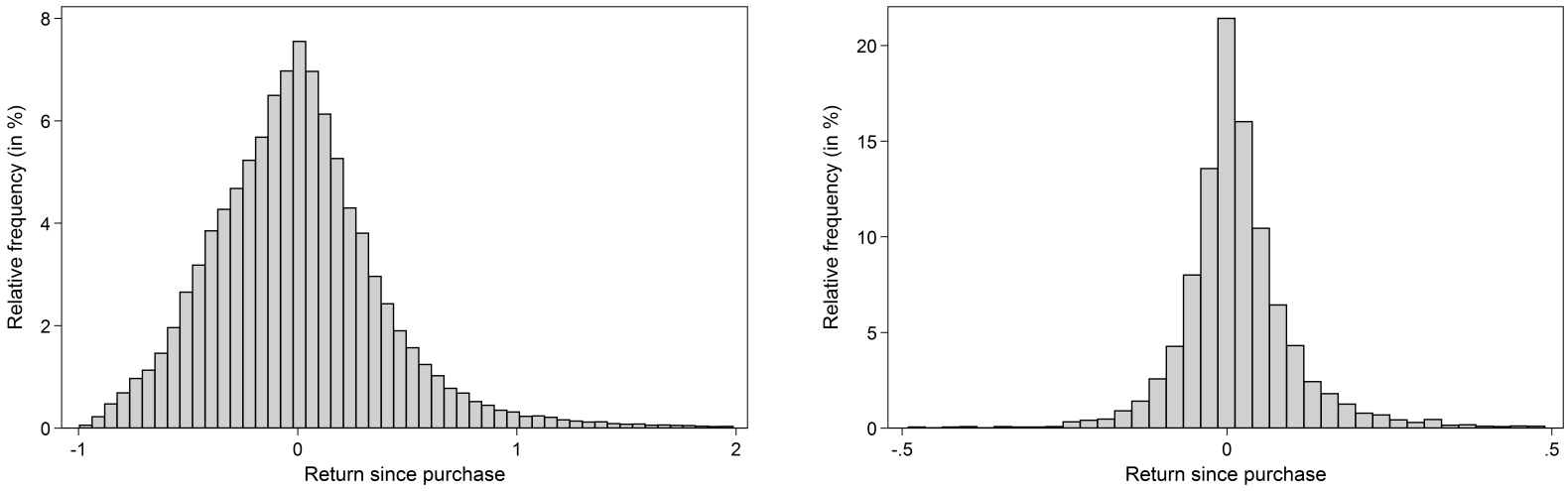
Figure 3.9 shows histograms of the returns for a fixed holding period of 12 months. Panel 3.9a shows the histogram for all returns, whereas panel 3.9b conditions on returns that were realized.[15] The conditional histogram contains relatively fewer observations in the tails and is more concentrated around 0.

Figure 3.10, which shows the corresponding kernel density estimates, illustrates that fact. The dashed line is the kernel density conditional on selling, and the solid line is the unconditional kernel density function. Probability mass is more concentrated around 0 conditional on selling than it is for all returns. This suggests that investors are overly likely to sell stocks that trade close to a zero return. Note that the figure is for a fixed holding period of 12 months for illustration but the implied pattern is robust and holds for all different holding periods. We turn to a more rigorous analysis of those patterns in the next section.

3.4.2 Duration model

Next we consider the Kaplan-Meier estimates. Figure 3.11 provides the starting point of the analysis. This figure plots the estimated hazard rate as a function of the holding period for both gains and losses for the entire sample for holding periods between 1 and 30 months.

[15] We deal with outliers by truncating the histograms at a return of 200%.



(a) *Unconditional*

(b) *Conditional on selling*

Figure 3.9: *Histograms of returns for a holding duration of 12 months*

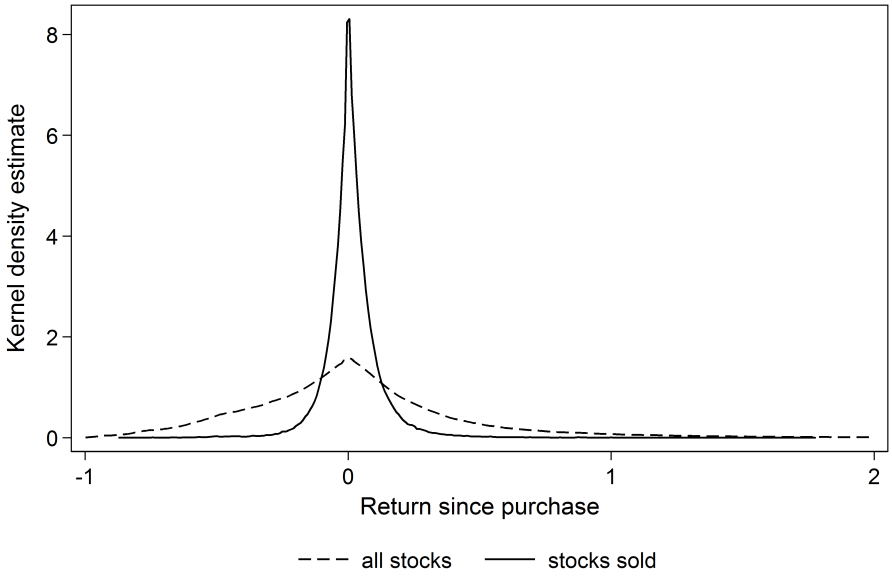


Figure 3.10: *Kernel density estimates*

The results are qualitatively and quantitatively in line with Ivkovic *et al.* (2005). In particular, the probability to sell gains is greater than the probability to sell losses for all holding periods, that is, we observe a disposition effect for all holding periods. Notice also that the probability to sell a stock declines with the holding period.

Figure 3.12 extends the analysis to the case of different return quantiles. We use quantiles instead of return intervals for two reasons. First, since we are interested in comparing differently-seized gains and losses, we want to make sure that the respective classes of returns that we compare contain equal numbers of observations. Second, we cannot match very high returns with same-sized low returns, because losses cannot exceed 100 %.[16]. The quantile procedure has proven valuable when trying to estimate non-linear functions and has been used by others before (e.g. DellaVigna and Pollet (2009)). Analyzing a very rich dataset enables us to split the data into many quantiles in the subsequent analysis. For illustrative purposes, figure 3.12 considers only 3 quantiles, small, medium and large gains

[16]If we restrict ourselves to gains smaller than 100% and use intervals instead of quantiles, we do get qualitatively similar results.

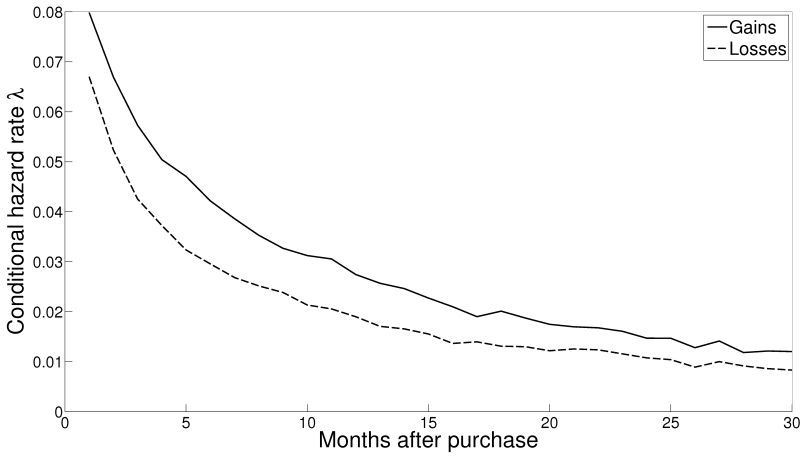


Figure 3.11: *Conditional hazard rate for gains and losses*

and losses respectively. In the figure, solid lines correspond to gains relative to purchase price and dashed lines correspond to losses. Circles denote the smallest gains and losses, triangles denote medium-sized ones and squares denote the largest tercile of gains and losses, respectively. The disposition effect is apparent for all three quantiles of returns (that is, the probability to sell gains exceeds the probability to sell losses for each respective pair of quantiles) , and that the overall probability to sell appears to decrease with higher returns (inward shift of the respective graphs).

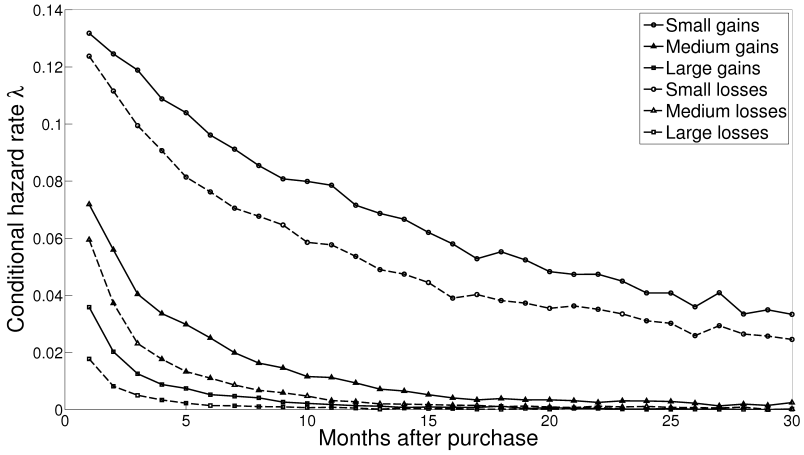


Figure 3.12: *Conditional hazard rate for gains and losses by return tercile*

We will investigate these preliminary findings in more detail. In particular, note that figure 3.12 is suggestive, but not conclusive. It plots the probability to sell as a function of the holding period. However, we are interested in the probability to sell a stock as a function of the return in order to compare the empirical results to the theories that we considered in section 3.2. Therefore, we repeat the calculations behind figure 3.12, split the sample into 25 gain and 25 loss quantiles and plot the probability to sell as a function of those quantiles *keeping the holding period fixed*.

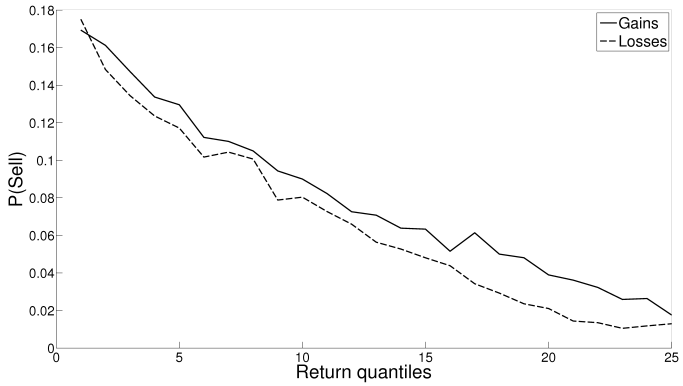
Figure 3.13 presents our main finding. It shows the hazard rate as a function of return quantiles for four different holding periods of 1, 6, 12 and 24 months (the findings are robust for other holding periods).

The disposition effect is apparent in all four panels: The probability to sell gains is larger than the probability to sell losses throughout most holding periods and return magnitudes. Small returns with short holding periods and large returns with long holding periods do not display a disposition effect.[17] Even more remarkable, however, is the pattern of the probabilities as a function of return. For both gains and losses, the probability of selling is largest for small returns and steadily declines for higher (absolute) returns, a consistent pattern throughout all panels.

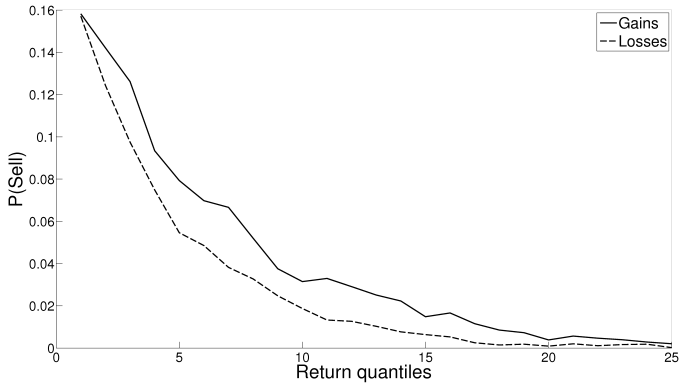
This finding stands in stark contrast to most model predictions that we have derived in section 3.2. The realization utility view and the casino gambling prospect theory view imply an increasing probability to sell in the gain region. Now we see that the opposite is actually the case, smaller returns are more likely to be sold than high returns. In addition, the probability to sell in the loss region is not constant, which also does not speak in favor of the realization utility view. On the other hand, the bunching view of prospect theory at the end of section 3.2 appears consistent with this evidence.

To further explore the relationship between the propensity to sell a stock and its return relative to purchase price, we estimate a Cox proportional hazard model. The parametric

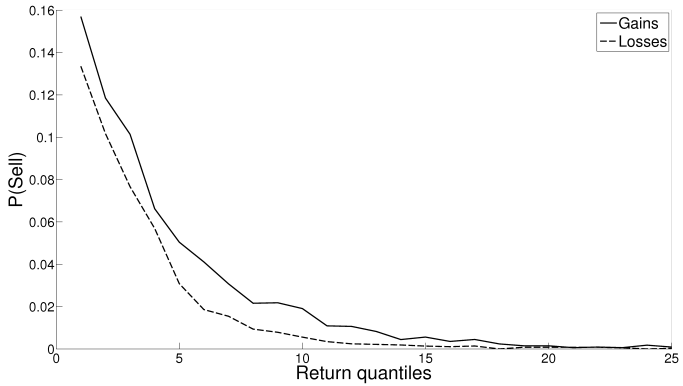
[17] Note that we have relatively few observations for holdings of exactly 24 months which makes the estimates less precise.



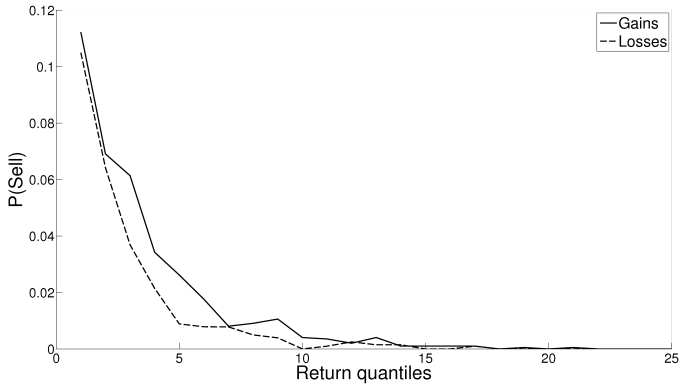
(a) *Holding period: 1 month*



(b) *Holding period: 6 months*



(c) *Holding period: 12 months*



(d) *Holding period: 24 months*

Figure 3.13: *Conditional hazard rate for different holding periods as function of stock returns. Return quantile 1 denotes small returns and return quantile 25 denotes large retunrs.*

approach has the advantage that we can control for other investor characteristics. Every investor-security combination is treated as an observation i, such that the hazard rate is given by:

$$\lambda_i(t,x) = \lambda_0(t) exp\{x'_{it}\delta\}, \text{ where} \tag{3.3}$$

$$x'_{it}\delta = g_{l,it}\beta_l + g_{m,it}\beta_m + g_{h,it}\beta_h + l_{l,it}\gamma_l + l_{m,it}\gamma_m + l_{h,it}\gamma_h + demo_i \tag{3.4}$$

Here, $g_{l,it}$ is a dummy that is equal to 1 if individual-security combination i trades at a low gain in holding period t, the other dummies are interpreted likewise with m being medium gains and losses and h being large gains and losses. $demo_i$ denotes the additional individual-level demographic control variables from table 3.1 such as age, gender and wealth. E.g. β_1 is the effect of a small gain on the probability to sell a security that has not been sold yet.

We divide the stock returns into three groups such that we can exclude the middle group and are able to compare effects for small and large gains respectively.[18] Note that the previous findings can be rewritten in terms of the model's coefficients: For instance, if the probability of selling declines away from a 0 return, we should observe $\gamma_1 \geq \gamma_2 \geq \gamma_3$ and $\beta_1 \geq \beta_2 \geq \beta_3$.

Columns (1) and (2) in table 3.2 report results with and without control variables. They are in line with the findings in the figures we presented in the preceding section. In particular, we find that the probability of selling declines with the size of returns. Small gains are more likely to be sold than medium-sized gains (the left-out category), and large gains are less likely to be sold than medium-sized gains. The same hold true for small and large losses relative to medium-sized losses.

Note also that the coefficients between quantiles differ significantly, rejecting the hypothesis that the probability of selling a security is constant across returns. The size of coefficients is in line with the hypothesized order in the preceding paragraph. Adding demographic control variables does not change results significantly.

[18] We have tried other splits into 2 to 10 groups with no qualitative change in results or interpretation.

Hartzmark (2015) shows that investors are more likely to sell stocks with more extreme returns among the stocks in a portfolio. In columns (3) and (4) of table 3.2, we therefore add two dummies as additional control variables: "'Best"' is equal to 1 if a stock is the best performing stock in an investor's portfolio in a given month, and "'Worst"' is equal to 1 if a stock is the worst performing stock in an investor's portfolio in a given month. Adding these controls does not alter our results. Given that the number of observations varies widely across specifications,[19], the stability of coefficients is actually remarkable.

Table 3.2: *Cox proportional hazard model for the probability to sell a stock given its return relative to purchase price.*

	Dependent variable: Dummy for selling			
Variable	(1)	(2)	(3)	(4)
g_l	1.613***	1.630***	1.640***	1.628***
g_h	-1.095***	-1.107***	-1.268***	-1.297***
l_l	1.332***	1.345***	1.420***	1.415***
l_h	-1.796***	-1.816***	-1.730***	-1.722***
Worst			-0.110***	-0.151***
Best			0.045***	0.027
Controls		X		X
N	4439680	1978133	1418624	578739

***, ** and * denote significance at the 1%, 5% and 10% significance level, respectively.

[19]The sample size varies for two reasons: First, demographic control variables are only available for a subset of regressors. Second, Hartzmark (2015) only includes observations that have at least five stocks in their portfolio, and we follow this approach in columns (3) and (4) of table 3.2.

3.4.3 Odean approach

Table 3.3 replicates Odean's main result. We use the proposed correction of the bias towards realized gains and losses that we discussed in section 3.3.3. As a result, all estimates are slightly lower than without bias correction, but qualitatively they are the same. We view this as evidence that the disposition effect cannot be explained by this simple estimation bias. For all months except for December, we find the disposition effect. PGR exceeds PLR in magnitude and the difference is significantly different from zero.[20] From January to November, investors realize 12% of their gains but only 6.7% of their losses. In December, the disposition effect is reversed and investors realize 9.9% of their gains but 10.5% of their losses. As discussed in Odean (1998) and Ivkovic *et al.* (2005), investors face a trade-off between realizing their losses and foregoing tax benefits. Since December is the last month for realizing tax-loss savings, investors choose more often to sell their losers in that month. Odean (1998) actually shows that the ratio of PGR and PLR declines over the year, implying that tax motives become more important in the course of the year. Note that our estimates have higher t-statistics than Odean (1998)'s results, mainly because our dataset is much larger.

Figure 3.14 goes beyond the original approach in Odean (1998) and plots the proportions of realized gains and losses as functions of stock returns for different holding periods.[21] For all panels, the proportion of realized gains exceeds the proportion of realized losses for all return quantiles, that is, the disposition effect is apparent for all sizes of returns and for all holding periods. Furthermore, for short holding periods, small gains are less likely to be sold than large gains, while for larger holding periods, stocks with small and large returns are equally likely to be sold.

---

[20] The standard error is computed in the same way as in Odean (1998), that is, the standard error of the respective difference is given by:

$$\sqrt{\frac{PGR(1-PGR)}{n_{rg}+n_{pg}}+\frac{PLR(1-PLR)}{n_{rl}+n_{pl}}}$$

where $n_{rg}, n_{pg}, n_{rl}, n_{pl}$ are the numbers of realized gains, paper gains, realized losses and paper losses.

[21] For the same reasons as in the previous section, we use quantiles rather than return intervals.

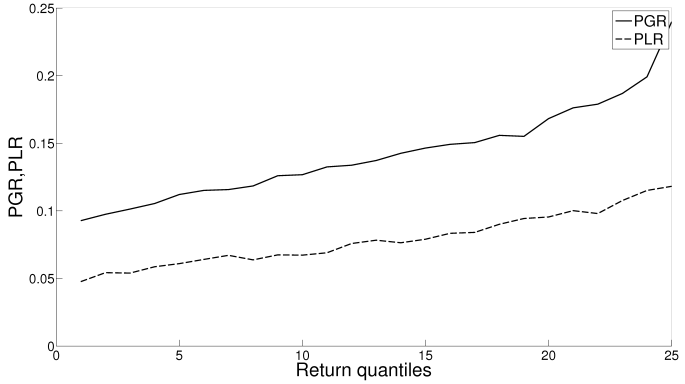
Table 3.3: *PGR and PLR for the Entire Data Set (bias corrected)*

	Entire Year	Dec	Jan-Nov	Entire Year (Odean)
PLR	0.070	0.105	0.067	0.098
PGR	0.118	0.099	0.120	0.148
Difference	-0.048	0.007	-0.053	-0.05
t-stat	-151.773	5.551	-160.598	-35

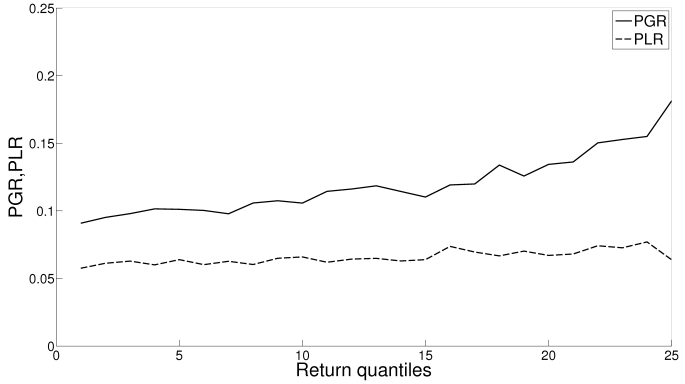
At first sight, these results appear to be quite different from our previous findings but recall that the Odean approach is biased towards trading activity, that is, the probability that is computed is not the unconditional probability to sell, but the conditional probability to sell, given some trading activity takes place in the portfolio. In the next section, we show that the Odean graphs in figure 3.14 display the behavior that one would expect if small returns are (unconditionally) more likely to be sold.

3.4.4 The propensity to trade

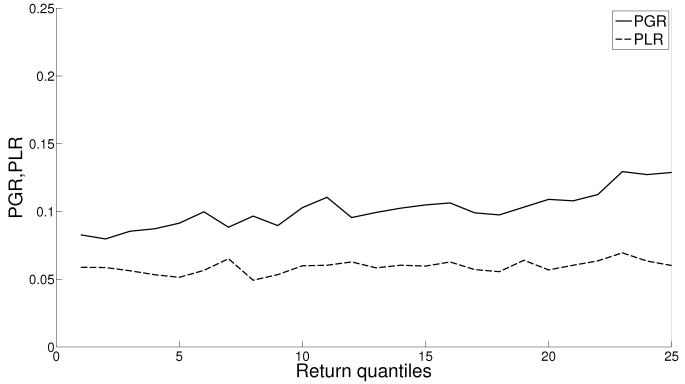
As noted in the previous section, at first sight, the results from the empirical approaches that we presented appear to lead to quite different conclusions. In this section, we show that this is actually not true. Rather, we argue that the two methods compute different objects and we show that there is a simple relationship between the two. While the duration model computes a conditional probability of selling a stock, given that one has not sold the stock before, the Odean approach estimates the probability to sell given that some activity in the portfolio takes place. Hence, the conditioning set is different in the two estimations, and this is at the heart of why the figures look so different.



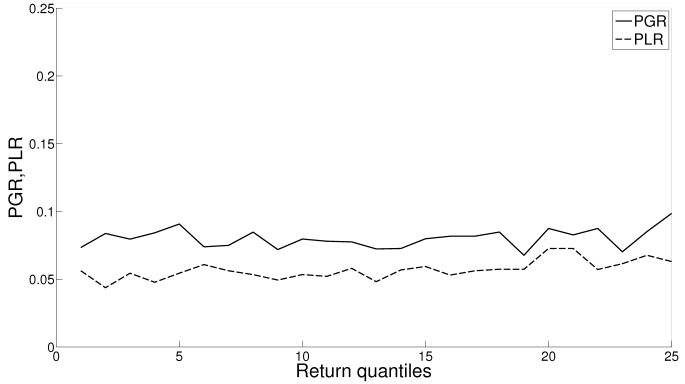
(a) *Holding period: 1 to 3 months*



(b) *Holding period: 5 to 7 months*



(c) *Holding period: 11 to 13 months*



(d) *Holding period: 22 to 24 months*

Figure 3.14: *Proportions of realized gains and losses as in Odean (1998) as a function of stock returns for different holding periods. Return quantile 1 denotes small returns and return quantile 25 denotes large returns.*

One can illustrate the difference via Bayes' rule. I.e. if $P(Sell)$ is the hazard rate, then

$$P(Sell|active) = \frac{\overbrace{P(active|Sell)}^{=1} P(Sell)}{P(active)}. \tag{3.5}$$

The left-hand side is the quantity that is estimated by the Odean approach. Therefore, e.g. if $P(Sell|active)$ is constant over returns, then $P(Sell)$ must be proportional to $P(active)$ in order to explain our findings, that is, $P(active)$ must also fall with absolute size of returns. Hence, the link between the results is given by $P(active)$ which we would like to compute in this section.

We estimate $P(active)$ using parametric probit regressions as well as a nonparametric approach. To start, we construct stock portfolios for each investor at the investor-month level as in section 3.3.2. We then collapse the data, compute the portfolio return for each investor-month combination and construct a dummy that is equal to one if the investor traded at all (sold or bought any stock) in a particular month and that is zero otherwise. Our resulting dataset therefore has one observation for each investor and each month .[22] We then regress this dummy on the portfolio return to get an estimate of the propensity of trading as a function of the portfolio return. Since we do not have a strong prior about how this relationship should look like, we start with a nonparametric regression using the Nadaraya-Watson estimator.[23] The nonparametric model that we want to estimate is given by:

$$g(x) = E[y|X = x], \tag{3.6}$$

where x denotes the portfolio return, and y is our activity dummy.

Hence, an estimate of $g(x)$ can be obtained from the regression

$$y = g(x) + \epsilon, \quad \epsilon \sim (0, \sigma^2(x)) \tag{3.7}$$

[22]We require that an investor holds at least two stocks in a given month to be included in the sample.

[23]Our procedure follows Haerdle (1990).

Figure 3.15 shows the estimated regression function and 95% confidence bands. The propensity to trade is highest around zero returns and declines in absolute size of returns. As argued above, this is what one would expect given the conditional estimates of the Odean approach.

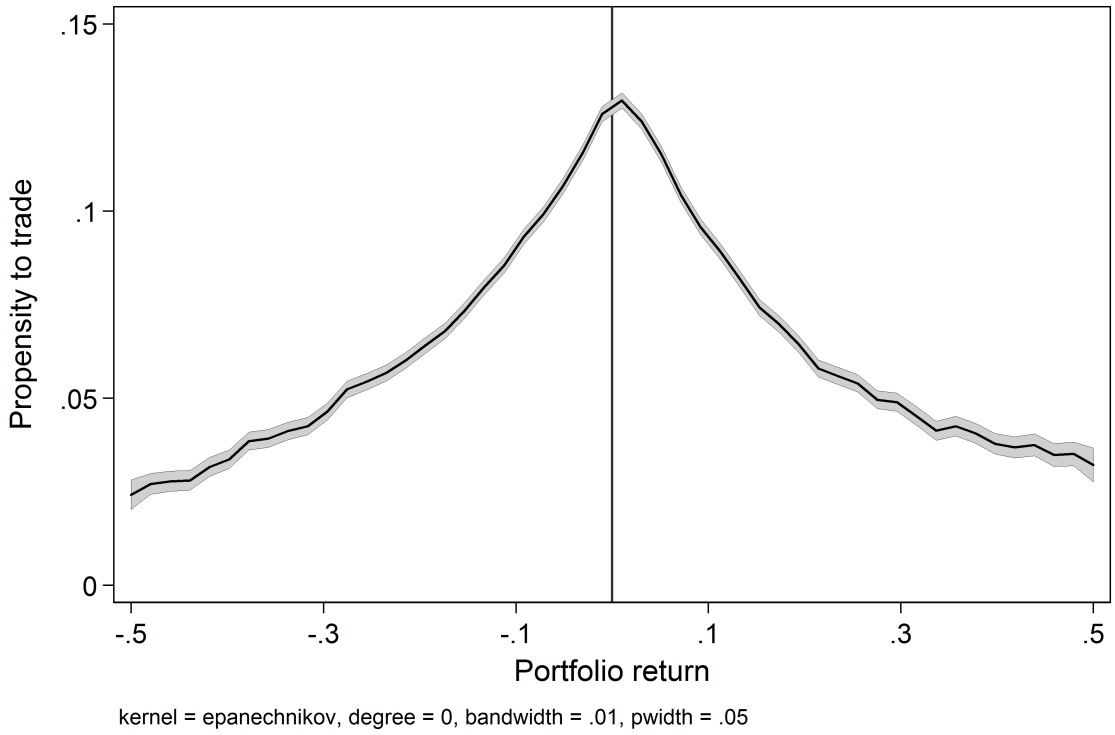


Figure 3.15: *Nonparametric estimate of propensity to trade. Grey area denotes the 95% confidence bands.*

Table 3.4 confirms this result parametrically. We regress the portfolio activity dummy on a (first and third order) polynomial of the portfolio return and we allow for a structural break in the relationship at a portfolio return of 0. The first two columns show results for a linear regression while the last two show results for a probit model. All models confirm the significant impact of the portfolio return on the propensity to trade at all, and the regression function is in line with the non-parametric model above, that is, it is upward-sloping for negative returns and downward-sloping for positive returns.

To summarize, the results of the duration model and the Odean approach can be reconciled when we take into account that the two estimates condition on different information

Table 3.4: *Propensity to trade as a function of portfolio return*

	Dependent variable: Dummy for any trade			
Regressors	Linear model		Probit model	
$I(Ret_p > 0)$	0.005***	0.007***	0.022***	0.031***
Ret_p	0.222***	0.509***	1.753***	2.600***
$I(Ret_p > 0)Ret_p$	-0.452***	-1.051***	-3.497***	-5.106***
Ret_p^2		1.026***		3.463***
Ret_p^3		0.887***		3.447***
$I(Ret_p > 0)Ret_p^2$		-0.062		-2.261**
$I(Ret_p > 0)Ret_p^3$		-1.468***		-1.553
Constant	0.115***	0.131***	-1.158***	-1.119***
N	1254918	1254918	1254918	1254918

***, ** and * denote significance at the 1%, 5% and 10% significance level, respectively.

sets. The link is given by the probability that an investor makes any trade as a function of the stocks' returns in her portfolio. We find that this probability declines in the size of absolute returns which reconciles our seemingly different results in sections 3.4.2 and 3.4.3.

3.5 Relation to Ben-David and Hirshleifer (2012)

Our approach is closely related to a recent article by Ben-David and Hirshleifer (2012) who also investigate the relation between past security returns and subsequent sales using the same data as we do. Their main result is that the probability to sell a security is "'asymmetrically V-shaped"', that is, larger returns are more likely to get sold and more so for positive returns, and that the selling probability does not have a discontinuity at a stock return of 0. While the latter is consistent with our results (recall that we find a stronger disposition effect for larger returns), the former fact seemingly stands in contrast to our reported results and we investigate the causes in this section.

In contrast to the approach in Ivkovic *et al.* (2005), Ben-David and Hirshleifer (2012) follow the holding of a stock on each *day* after it was purchased without conditioning on portfolio activity as in Odean (1998). This approach hugely expands the data set and most of our results are based on a random sample of 25% of accounts.[24] For the short holding horizons in figure 3.16 below, we are able to use the entire data set.

We start by successfully replicating the main results in Ben-David and Hirshleifer (2012). First, we follow their procedure to construct each investor's portfolio on each possible trading day. This enables us to follow each stock from the purchase day to the selling day. Table 3.5 reports the unconditional probability that a stock gets sold for different holding periods. In general, stocks are sold infrequently: The probability that a stock is sold in the first 30 (trading) days after purchase is .79 percent and it monotonically decreases for longer holding horizons.

Figure 3.16 shows our replication of the motivating results in Ben-David and Hirshleifer

[24]Ben-David and Hirshleifer (2012) follow the same strategy and base their estimates on a random sample of 10,000/77,000 $\approx$ 12% of accounts.

Table 3.5: *Unconditional probability (in %) of selling for different numbers of holding days. Note: Days are trading days.*

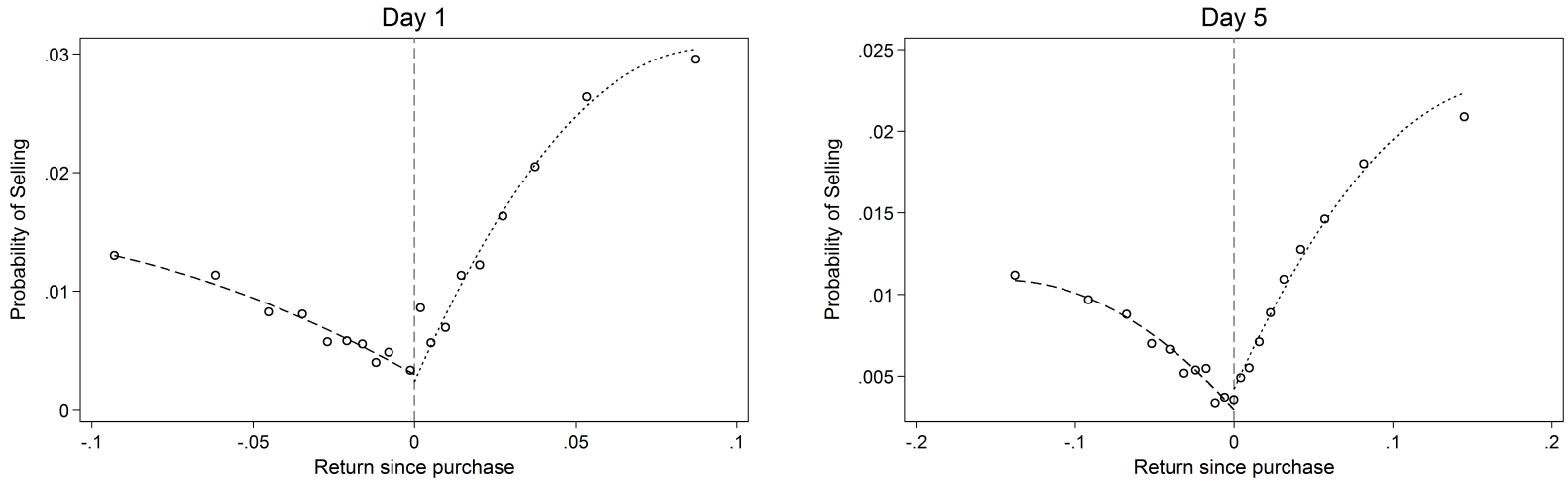
Holding days	1-30	31-60	61-90	91-120	121-150	151-180	181-210	211-240
Probability to sell	.790	.508	.391	.309	.258	.230	.201	.174

(2012) (that is, their figure 1). We document a sharp decrease of the selling probability around small stock returns and we find that the difference between small positive and small negative returns is negligible.[25] The results thus support the existence of an asymmetrically V-shaped selling schedule, that is, stocks with small returns are unlikely to be sold and stocks with higher returns are more likely to be sold, more so for positive returns. Note, however, that the results in figure 3.16 are conditional on the specific holding periods of 1 or 5 days, both of which are extremely short.

Figure 3.17 illustrates the results of the same exercise when data are pooled over different holding horizons. The left panel pools all holding periods less than 30 days and is generally in line with the results in figure 3.16, although there is an apparent small discontinuity around 0. The right panel pools all holding periods of less than 250 trading days and looks very different: The discontinuity of the selling probability around 0 is more apparent and the probability to sell appears to *decrease* with the absolute value of the return.

Table 3.6 provides a more detailed account of the relation between the discontinuity result around a return of 0 and the holding period. We regress an indicator for stock sales (multiplied by 100) on an indicator of whether a stock's return was greater than zero, on a third-order polynomial of a stock's return and on interactions of the indicator and the polynomial terms. Each column reports the regression results for a different stock holding period. For ease of interpretability, we report results for the linear probability model here but we have checked that non of the results change when a logistic regression model is

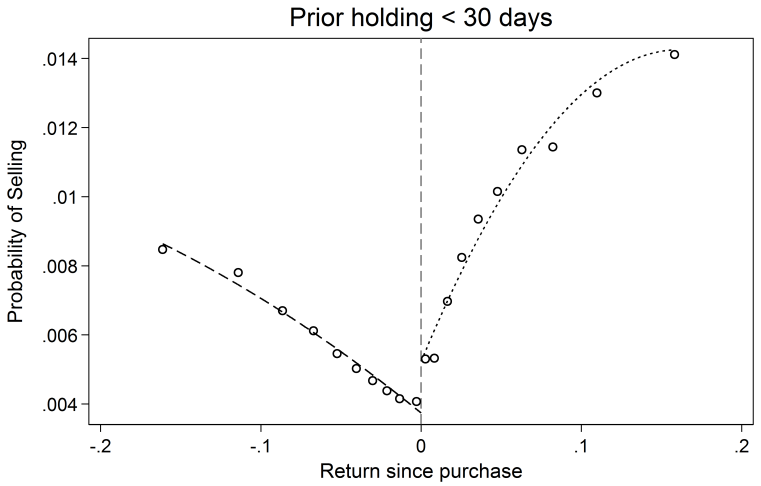
[25]Our replication is based on a quadratic polynomial in the stock return on each side of the threshold of 0, while Ben-David and Hirshleifer (2012) use higher-order polynomials. We produced additional results using nonparametric kernel density regressions (available on request) that looked even more similar to the original results.



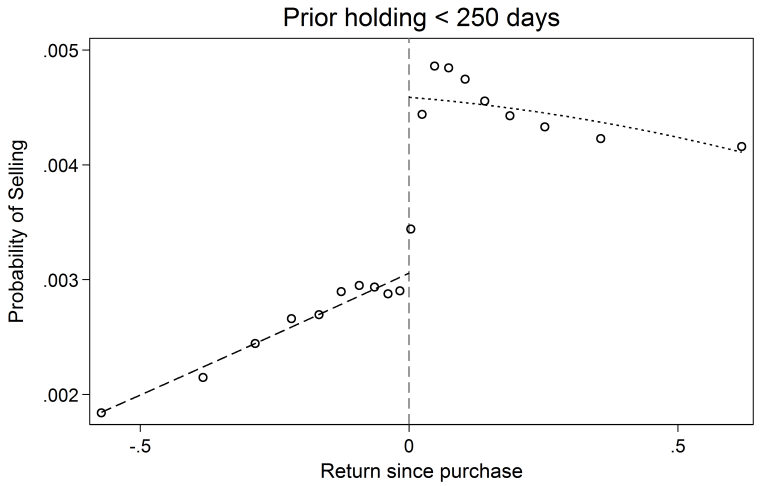
(a) *For holding period of 1 day*

(b) *For holding period of 5 days*

Figure 3.16: *Replication of Ben-David and Hirshleifer (2012)*



(a) *Holdings less than 30 days*



(b) *Holdings less than 250 days*

Figure 3.17: *Pooled selling schedules*

used instead. Note that the indicator for a positive return comes in significantly in each regression. For instance, for a holding period of up to 30 days, a positive return increases the likelihood to sell the stock by .367 percentage points. To assess whether this is large or small, we scale the coefficient by the unconditional probability of selling for each holding interval (from table 3.5 above) in the last row of the table. For shorter holding horizons, the discontinuity is about 40% of the unconditional probability to sell, while for longer holding horizons it is about 20%.

The effect of positive returns relative to purchase price on the probability to sell a stock is sizeable for all horizons within one year of stock purchase. The long-lasting effect is in line with other research that investigates how investment decisions depend on past prices. For instance, Heath *et al.* (1999) show that employees are more likely to exercise stock options when the stock price is greater than the maximum achieved over the previous year, and Baker *et al.* (2012) show that the 52-week high is an important anchor for merger offers. Here, we show that the purchase price of a stock affects trading decisions for (at least) one year, an unsurprising finding in light of the aforementioned studies.

To summarize, we document that the probability to sell is asymmetrically V-shaped only for very short holding periods. If results are pooled over different horizons, we find a pronounced discontinuity at 0 returns and we also find that the selling probability decreases in the absolute value of the return, in line with our previous findings.

It seems plausible that investors that trade at very short horizons are different from those that trade at longer horizons. Indeed, Ben-David and Hirshleifer (2012) report weaker results for infrequent traders (their figure 5). As our study generally focuses on holding periods greater than one month (and up to two years), it can also be read as a study of the behavior of those less frequent traders.

3.6 Conclusion

What drives the disposition effect? In this paper, we investigate the explanatory power of two leading explanations, prospect theory and realization utility. We derive implications

Table 3.6: *Probability of selling as a function of stock returns for different holding periods*

Holding days	1-30	31-60	61-90	91-120	121 -150	151-180	181-210	211-240
$I(Ret > 0)$	0.367***	0.211***	0.143***	0.106***	0.039***	0.055***	0.035**	0.031***
Ret	-4.892***	-1.344***	-0.048	-0.071	0.216*	0.090	0.159	0.187*
Ret^2	-14.023***	-2.689***	0.186	-0.165	0.442	0.140	0.338	0.397
Ret^3	-10.649***	-0.915	0.348	-0.104	0.378	0.073	0.277	0.314
$I(Ret > 0)Ret$	9.376***	2.710***	0.500***	0.367**	0.045	0.078	0.074	-0.086
$I(Ret > 0)Ret^2$	10.909***	1.979***	-0.222	0.141	-0.483	-0.194	-0.528	-0.417
$I(Ret > 0)Ret^3$	11.089***	1.002	-0.347	0.104	-0.377	-0.070	-0.245	-0.313
Constant	0.350***	0.278***	0.279***	0.226***	0.225***	0.192***	0.176***	0.158***
N	2808535	2857042	2869490	2888734	2781890	2719315	2712780	2598521
$\frac{\hat{\beta}_{I(Ret>0)}}{\text{Uncond P(Sell)}}$	.465	.416	.366	.343	.151	.239	.174	.178

of the theories for the probability to sell as a function of the return size, and contrast the predictions to empirical findings.

Our main finding is that, for all but very short holding horizons, investors are more likely to sell stocks with small returns (i.e. stocks with prices close to the purchase price) than to sell stocks with large returns. Since this is such a puzzling fact, we establish trust in the result by showing it using a couple of different methods. Using the duration model of Ivkovic *et al.* (2005), we find that larger absolute returns are less likely to be sold than small returns. Using the Odean (1998) methodology, we find that for small holding periods, small gains are less likely to be sold than large gains, while for larger holding periods, stocks with small and large returns are equally likely to be sold.

We show how to reconcile the seemingly contradictory empirical findings by pointing out that the two approaches estimate different probabilities. While Ivkovic *et al.* (2005) consider the probability of selling a stock with a given return in a given month (conditional on still holding it), Odean considers the probability of selling a stock, given that the investor sells or buys a stock in her portfolio. That is, when we compare the two approaches, we need to take into account that the probability of undertaking any transaction in the portfolio is a function of the individual returns of the stocks in the portfolio. We find that individuals are more likely to make transactions for small returns than for large returns.

Jointly, these findings pose yet another challenge: An investor is more likely to perform any action (i.e. sell or buy stocks, look into her portfolio?) if returns of stocks in her portfolio are small. Once the investor decides to act, however, she is more likely to sell large returns. It is hard to think of a theory that would predict this.

In any case, it should be apparent that realization utility cannot explain important features of the data. A version of prospect theory, on the other hand, that puts emphasis on "bunching" around the kink appears to be consistent with the facts. Obviously, consistency with the facts alone does not make a theory the true explanation. For instance, an explanation that combines elements of prospect theory, realization utility and overconfidence might very well be at the heart of the disposition effect. In any case, a debate-settling explanation of the

disposition effect has to account for the empirical facts that we presented.

Despite the fact that the choice of the reference point is crucial for any analysis of reference-dependent utility, the existing literature on the disposition effect has not thoroughly investigated this choice (with the exception of Meng (2013)). Instead, it solely focuses on a stock's buying price as "a noisy proxy for the investor's true reference point" (Odean, 1998). While Odean discusses the possibility of other determinants of the reference point (in particular, expectations), his focus remains on variants of the purchase price[26].

Koszegi and Rabin (2006) argue that expectations rather than the status quo play a key role in the formation of individuals' reference points. In particular, when people do not plausibly expect to maintain the status quo, "equating the reference point with expectations generally makes better predictions". For trading decisions, this suggests that the purchase price might not be a good choice of a reference point. While it may be a good proxy in times of low returns, with soaring stock prices like in the 1990s (the coverage of the data set falls within that time period)[27], investor may have higher expectations, and, hence, a higher reference point than the status quo.

Future research should take into account the possibility that investors' reference points may be driven by expectations or might generally deviate from a stock's purchase price. The application in Kleinberg *et al.* (2015), which is also part of this thesis, can be viewed as one attempt to allow for a more flexible effect of past returns on trading decisions.

[26]For investors who buy the same stock several times, Odean considers the average purchase price, the highest purchase price, the first purchase price, and the most recent purchase price

[27]Our dataset covers transaction from 1991 through 1996, a time period during which on average the S&P 500 index rose over 15% annually.

References

Alessi, L. and Detken, C. (2014). Identifying excessive credit growth and leverage, eCB Working Paper.

Asness, C., Frazzini, A., Israel, R. and Moskowitz, T. (2014). Fact, fiction and momentum investing. *Journal of Portfolio Management*.

Asness, C. S. (1997). The interaction of value and momentum strategies. *Financial Analysts Journal*, **53** (2), 29–36.

Baker, M., Pan, X. and Wurgler, J. (2012). The effect of reference point prices on mergers and acquisitions. *Journal of Financial Economics*, **106** (1), 49–71.

Bandarchuk, P. and Hilscher, J. (2012). Sources of Momentum Profits: Evidence on the Irrelevance of Characteristics. *Review of Finance*, **17** (2), 809–845.

Barber, B. and Odean, T. (2000). Trading is hazardous to your wealth: The common stock investment performance of individual investors. *Journal of Finance*, **55**, 773–806.

Barberis, N. (2012). A model of casino gambling. *Management Science*, **58** (1), 35–51.

—, Huang, M. and Thaler, R. H. (2006). Individual preferences, monetary gambles, and stock market participation: A case for narrow framing. *American Economic Review*, **96** (4), 1069–1090.

— and Xiong, W. (2009). What drives the disposition effect? an analysis of a long-standing preference-based explanation. *Journal of Finance*, **64** (2), 751–784.

— and — (2012). Realization utility. *Journal of Financial Economics*, **104** (2), 251–271.

Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, **80** (6), 2369–2429.

—, Chernozhukov, V., Fernández-Val, I. and Hansen, C. (2013). *Program evaluation with high-dimensional data*. Tech. rep., arXiv preprint arXiv:1311.2645.

—, — and Hansen, C. (2011). *Inference for high-dimensional sparse econometric models*. Tech. rep., arXiv preprint arXiv:1201.0220.

—, — and — (2014). High-dimensional methods and inference on structural and treatment effects. *The Journal of Economic Perspectives*, **28** (2), 29–50.

Ben-David, I. and Hirshleifer, D. (2012). Are investors really reluctant to realize their losses? trading responses to past returns and the disposition effect. *Review of Financial Studies*, **25** (8), 2485–2532.

Biau, G., Devroye, L. and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *The Journal of Machine Learning Research*, **9**, 2015–2033.

Breiman, L. (1996). Bagging predictors. *Machine learning*, **24** (2), 123–140.

— (2001). Random forests. *Machine Learning*, **45** (1), 5–32.

— (2002). Looking inside the black box, downloaded from www.stat.berkeley.edu/users/breiman/wald2002-2.pdf.

Brennan, M., Chordia, T. and Subrahmanyam, A. (1998). Alternative factor specifications, security characteristics, and the cross-section of expected stock returns. *Journal of Financial Economics*, **49** (3), 345 – 373.

Büchlmann, P. and Yu, B. (2002). Analyzing bagging. *Annals of Statistics*, **34** (4), 927–961.

Burke, M. and Kraut, R. (2013). Using facebook after losing a job: Differential benefits of strong and weak ties. In *Proceedings of the 2013 conference on Computer supported cooperative work*, ACM, pp. 1419–1430.

Chen, J., Hong, H. and Stein, J. C. (2002). Breadth of ownership and stock returns. *Journal of financial Economics*, **66** (2), 171–205.

Cochrane, J. H. (2011). Presidential address: Discount rates. *The Journal of Finance*, **66** (4), 1047–1108.

Criminisi, A. and Shotten, J. (2013). *Decision Forests for Computer Vision and Medical Image Analysis*. Springer.

Dahl, G. B., Loken, K. V. and Mogstad, M. (2014). Peer effects in program participation. *American Economic Review*, **104** (7), 2049–74.

Daniel, K. and Moskowitz, T. J. (2014). *Momentum Crashes*. Working Paper 20439, National Bureau of Economic Research.

— and Titman, S. (1997). Evidence on the characteristics of cross sectional variation in stock returns. *The Journal of Finance*, **52** (1), 1–33.

de Bondt, W. and Thaler, R. (1985). Does the stock market overreact? *The Journal of Finance*, **40** (3), 793–805.

de Groot, W., Huij, J. and Zhou, W. (2012). Another look at trading costs and short-term reversal profits. *Journal of Banking & Finance*, **36** (2), 371–382.

DellaVigna, S. and Pollet, J. (2009). Investor inattention and friday earnings announcements. *Journal of Finance*, **64**, 709–749.

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A. and Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PLoS ONE*, **6** (12), 1–26.

Doyle, A. C. (1892). The adventure of the cardboard box. In *The Strand Magazine*, George Newnes Ltd.

Duttagupta, R. and Cashin, P. (2011). Anatomy of banking crises in developing and emerging market countries. *Journal of International Money and Finance*, **30** (2), 354–376.

Einav, L. and Levin, J. (2014). The data revolution and economic analysis. In J. Lerner and S. Ster (eds.), *Innovation Policy and the Economy*, NBER.

Fama, E. and French, K. (1992). The cross-section of expected stock returns. *The Journal of Finance*, **XLVII** (2), 427–467.

— and — (1996). Multifactor explanations of asset pricing anomalies. *The Journal of Finance*, **51** (1), 55–84.

— and — (2008). Dissecting anomalies. *The Journal of Finance*, **63** (4), 1653–1678.

— and — (2013). A five-factor asset pricing model, unpublished manuscript.

Fama, E. F. (1965). Random walks in stock-market prices. *Financial Analysts Journal*, **21**, 55–59.

— (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *Journal of Finance*, **25** (2), 383–417.

Fan, J., Lv, J. and Qi, L. (2011). Sparse high dimensional models in economics. *Annual review of economics*, **3**, 291.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. and Lin, C.-J. (2008). LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research*, **9**, 1871–1874.

Farrell, M. H. (2013). *Robust inference on average treatment effects with possibly more covariates than observations*. Tech. rep., arXiv preprint arXiv:1309.4686.

Frazzini, A., Israel, R. and Moskowitz, T. J. (2013). Trading costs of asset pricing anomalies, unpublished manuscript.

Gelman, A. and Imbens, G. (2013). *Why ask Why? Forward Causal Inference and Reverse Causal Questions*. Working Paper 19614, National Bureau of Economic Research.

Genesove, D. and Mayer, C. (2001). Loss aversion and seller behavior: Evidence from the housing market. *Quarterly Journal of Economics*, **116** (4), 1233–1260.

Goyal, A. (2011). Empirical cross-sectional asset pricing: a survey. *Financial Markets and Portfolio Management*, **26** (1), 3–38.

— and Wahal, S. (2013). Is momentum an echo?, unpublished manuscript.

Green, J., Hand, J. and Zhang, F. (2013). The supraview of return predictive signals. *Review of Accounting Studies*, **18** (3), 692–730.

—, Hand, J. R. M. and Zhang, F. (2014). The remarkable multidimensionality in the cross section of expected u.s. stock returns, unpublished manuscript.

Grinblatt, M. and Moskowitz, T. J. (2004). Predicting stock price movements from past returns: The role of consistency and tax-loss selling. *Journal of Financial Economics*, **71** (3), 541–579.

Haerdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press.

Han, Y., Yang, K. and Zhou, G. (2011). A new anomaly: The cross-sectional profitability of technical analysis, unpublished manuscript.

— and Zhou, G. (2013). Trend factor: A new determinant of cross-section stock returns, unpublished manuscript.

Hartzmark, S. M. (2015). The worst, the best, ignoring all the rest: The rank effect and trading behavior. *Review of Financial Studies*, **28** (4), 1024–1059.

Harvey, C. R., Liu, Y. and Zhu, H. (2013). ... and the cross-section of expected returns, unpublished manuscript.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York Inc.

Haugen, R. A. and Baker, N. L. (1996). Commonality in the determinants of expected stock returns. *Journal of Financial Economics*, **41** (3), 401–439.

Heath, C., Huddart, S. and Lang, M. (1999). Psychological factors and stock option exercise. *The Quarterly Journal of Economics*, **114** (2), 601–627.

Hens, T. and Vlcek, M. (2011). Does prospect theory explain the disposition effect? *Journal of Behavioral Finance*, **12** (3), 141–157.

Heston, S. L. and Sadka, R. (2008). Seasonality in the cross-section of stock returns. *Journal of Financial Economics*, **87** (2), 418–445.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *Pattern Analysis and Machine Intelligence, IEEE*, **20** (8), 832–844.

Huerta, R., Corbacho, F. and Elkan, C. (2013). Nonlinear support vector machines can systematically identify stocks with high and low future returns. *Algorithmic Finance*, **2**, 45–58.

Hyafil, L. and Rivest, R. L. (1976). Constructing optimal binary decision trees is np-complete. *Information Processing Letters*, **5** (1), 15 – 17.

Ivkovic, Z., Poterba, J. and Weissbenner, S. (2005). Tax-motivated trading by individual investors. *American Economic Review*, **95(5)**, 1605–1630.

Jacobs, B. I. and Levy, K. N. (1989). The complexity of the stock market. *The Journal of Portfolio Management*, **16** (1), 19–27.

Jegadeesh, N. (1990). Evidence of predictable behavior of security returns. *The Journal of Finance*, **45** (3), 881–898.

— and Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance*, **48** (1), 65–91.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, **47**, 263–291.

Kaminsky, G. L. (2006). Currency crises: Are they all the same? *Journal of International Money and Finance*, **25** (3), 503–527.

Kaustia, M. (2010). Prospect theory and the disposition effect. *Journal of Financial and Quantitative Analysis*, **45** (03), 791–812.

Keim, D. B. and Madhavan, A. (1997). Transactions costs and investment style: an inter-exchange analysis of institutional equity trades. *Journal of Financial Economics*, **46** (3), 265 – 292.

Kleinberg, E. (1990). Stochastic discrimination. *Annals of Mathematics and Artificial intelligence*, **1** (1), 207–239.

— (1996). An overtraining-resistant stochastic modeling method for pattern recognition. *The annals of statistics*, **24** (6), 2319–2349.

Kleinberg, J., Mullainathan, S., Tan, C. and Zimmermann, T. (2015). Inductive theory testing: A framework and application to the disposition effect, mimeo.

Kogan, L. and Tian, M. (2012). Firm characteristics and empirical factor models: a data-mining experiment, international Finance Discussion Papers.

Koszegi, B. and Rabin, M. (2006). A model of reference-dependent preferences. *Quarterly Journal of Economics*, **121(4)**, 1133–1166.

Lee, C. M. and Swaminathan, B. (2000). Price momentum and trading volume. *The Journal of Finance*, **55** (5), 2017–2069.

Lehmann, B. (1990). Fads, martingales, and market efficiency. *The Quarterly Journal of Economics*, **105** (1), 1–28.

Lewellen, J. (2013). The cross section of expected returns, unpublished manuscript.

Lintner, J. (1965). The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets. *The Review of Economics and Statistics*, **47** (1), 13–37.

Manasse, P. and Roubini, N. (2009). "rules of thumb" for sovereign debt crises. *Journal of International Economics*, **78** (2), 192–205.

McLean, R. D. and Pontiff, J. E. (2012). Does Academic Research Destroy Stock Return Predictability?, unpublished working paper.

Meng, J. (2013). Can prospect theory explain the disposition effect? a new perspective on reference points, mimeo.

Mossin, J. (1966). Equilibrium in a capital asset market. *Econometrica*, **34**, 768–783.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT Press.

Novy-Marx, R. (2012). Is momentum really momentum? *Journal of Financial Economics*, **103** (3), 429–453.

— (2013). The other side of value: The gross profitability premium. *Journal of Financial Economics*, **108** (1), 1–28.

— and Velikov, M. (2014). Anomalies and their trading costs, unpublished working paper.

Odean, T. (1998). Are investors reluctant to realize their losses? *Journal of Finance*, **53**, 1775–1798.

Pang, B., Lee, L. and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, Association for Computational Linguistics, pp. 79–86.

Patton, A. J. and Timmermann, A. (2010). Monotonicity in asset returns: New tests with applications to the term structure, the CAPM, and portfolio sorts. *Journal of Financial Economics*, **98** (3), 605–625.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

Popper, K. (1934). *Logik der Forschung*. Mohr Siebeck.

Sharpe, W. F. (1964). Capital Asset Prices: A Theory Of Market Equilibrium Under Conditions Of Risk. *Journal of Finance*, **19** (3), 425–442.

Shefrin, H. and Statman, M. (1985). The disposition to sell winners too early and ride losers too long: Theory and evidence. *Journal of Finance*, **40**, 777–790.

Subrahmanyam, A. (2010). The Cross-Section of Expected Stock Returns: What Have We Learnt from the Past Twenty-Five Years of Research? *European Financial Management*, **16** (1), 27–42.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58** (1), 267–288.

Tsai, C.-f., Lin, Y.-c., Yen, D. C. and Chen, Y.-m. (2011). Predicting stock returns by classifier ensembles. *Applied Soft Computing Journal*, **11** (2), 2452–2459.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, **27** (11), 1134–1142.

Watkins, B. (2003). Riding the wave of sentiment: An analysis of return consistency as a predictor of future returns. *The Journal of Behavioral Finance*, **4** (4), 37–41.

Weber, M. and Camerer, C. F. (1998). The disposition effect in securities trading: an experimental analysis. *Journal of Economic Behavior and Organization*, **33** (2), 167 – 184.

Zhang, C. and Ma, Y. (2012). *Ensemble Machine Learning: Methods and Applications*. Springer.

Appendix A

Appendix to Chapter 1

A.1 Illustration of a conditional portfolio sort

In our illustration, we consider conditional portfolio sorts that are each based on two of the following variables: short-term reversal, momentum, intermediate momentum, size, gross profitability, and book-to-market. Our results complement Fama and French (2008) who sort stocks into three size portfolios first and then sort each portfolio subsequently on a further firm characteristic. We follow the same approach with a few modifications:

Each month, we sort stocks into one of three portfolios based on the value of a sorting variable from the following list: short-term reversal, momentum, intermediate momentum, size, gross profitability, and book-to-market. Short-term reversal is defined as the return over the most recent prior month, momentum is the return over the past twelve months (excluding the most recent month) and intermediate momentum is the return over the past twelve months excluding the most recent six months. Accounting-based variables are constructed in a standard fashion (see appendix A.3.1).

Each portfolio is then further divided into ten portfolios based on a second sorting variable from the same list and we compute the equal weighted hedge return based on the second sorting within each of the three portfolios. Table A.1 reports the equal weighted hedge returns, their associated t-statistics, and the test of Patton and Timmermann (2010) for

the monotonicity of returns over the decile sort.[1] Columns labeled "Low" contain estimates based on firms in the lowest tercile of the first sorting variable, "Middle" and "High" denote the next two terciles, and "All" uses all observations without a sort on the first sorting variable for comparison.[2]

First, note that all of the sorting variables achieve significant returns in the equal-weighted hedge portfolios and pass the monotonicity test of Patton and Timmermann. To delve into the details of the results: When firms are sorted on short-term reversal first, momentum, intermediate momentum and value still manage to pass the t-test and the monotonicity test within each short-term reversal tercile portfolio. Size does not work in the top tercile, and gross profitability passes the t-test, but fails to provide monotone returns throughout all terciles. A similar picture emerges when returns are sorted on momentum first. Value and short-term reversal work throughout all terciles, while intermediate momentum yields (weakly) significant t-statistics in each tercile but does not pass the monotonicity test for the low and middle groups of momentum-sorted returns. Size passes all t-tests, but fails to provide monotone returns in the lowest two terciles. Interestingly, momentum sorts continue to work well when firms are sorted on intermediate momentum first but the reverse is not true: Intermediate momentum sorts do not consistently give a significant hedge return (only in low momentum stocks) or monotone returns (only in the middle tercile of momentum stocks). Initial sorts on value or size leave the monotonicity of return sorts intact, but interfere with the monotonicity and t-tests of gross profitability. When firms are sorted on gross profitability first, equal-weighted hedge returns are significant for all variables in all terciles, but the returns to medium gross profitability firms is not monotone when sorted by value.

The overall picture that emerges is that of return sorts being relatively stable while

[1]Since the Patton and Timmermann test for monotonicity is not (yet) standard, here is a brief summary: It computes the pairwise difference between the average returns of adjacent decile portfolios, and then tests whether the minimum of these differences is greater than zero (if the research hypothesis is that returns are increasing over deciles). If the test rejects, this provides support for the research hypothesis.

[2]In other words, the column "All" gives the results for an unconditional sort on the row variable.

accounting-based sorts are less robust to initial sorts on some other return- or accounting-based variable. The results illustrate the potential relevance of correlated return- and accounting-based characteristics, and the necessity to consider conditional returns when the objective is to evaluate the importance of a new candidate predictor variable. Variable interactions can also be relevant as is evident from the fact conditional sorts often work only in some of the tercile portfolios.[3]

[3]It is also possible to condition on more than one variable in this setting by first doubly sorting all stocks on two variables into, say, three categories each for a total of nine portfolios. Within each portfolio, one could then compute the same statistics as above, and discuss the effects of conditioning on levels and interactions of variables. While, in principle, feasible for a few variables, the approach does not lend itself to an easy interpretation in higher dimensions.

Table A.1: *Conditional portfolio sorts: Average returns, t-statistics and p-values of monotonicity tests*

	First sort on: R(0,1)											
	Average return				t-statistics				p-value PT test			
Sorting on	Low	Mid	High	All	Low	Mid	High	All	Low	Mid	High	All
R1 11	0.59	1.51	2.78	1.56	2.88	7.46	12.95	8.54	0.13	0.00	0.00	0.00
R6 6	0.71	1.13	1.74	1.19	3.98	6.32	9.48	7.69	0.00	0.00	0.00	0.00
size	-1.50	-0.73	-0.21	-1.28	-5.84	-2.87	-0.86	-5.37	0.00	0.02	0.57	0.00
gross profitability	0.60	0.32	0.77	0.48	4.24	2.32	5.52	3.99	0.39	0.59	0.94	0.09
booktomarket	1.27	1.09	1.00	1.09	7.00	6.02	4.96	6.36	0.00	0.01	0.00	0.00
	First sort on: R(1,11)											
R0 1	-3.29	-1.02	-0.85	-1.79	-15.17	-6.01	-4.79	-10.60	0.00	0.04	0.00	0.00
R6 6	0.37	0.27	0.25	1.19	2.65	1.90	1.52	7.69	0.97	0.07	0.43	0.00
size	-0.75	-1.04	-0.95	-1.28	-3.07	-4.41	-4.04	-5.37	0.62	0.67	0.00	0.00
gross profitability	0.66	0.36	0.57	0.48	4.03	2.82	4.32	3.99	0.05	0.06	0.13	0.09
booktomarket	1.39	1.23	0.67	1.09	7.50	7.25	3.52	6.36	0.00	0.00	0.00	0.00
	First sort on: R(6,6)											

Continued on next page

Table A.1: *(continued)*

	Average return				t-statistics				p-value PT test			
Sorting on	Low	Mid	High	All	Low	Mid	High	All	Low	Mid	High	All
R0 1	-2.58	-1.46	-1.38	-1.79	-12.62	-8.58	-7.15	-10.60	0.00	0.00	0.00	0.00
R1 11	1.35	1.03	1.24	1.56	7.15	6.43	6.88	8.54	0.09	0.02	0.01	0.00
size	-0.92	-1.03	-0.78	-1.28	-3.72	-4.34	-3.38	-5.37	0.08	0.05	0.00	0.00
gross profitability	0.68	0.38	0.50	0.48	4.32	3.22	3.75	3.99	0.00	0.04	0.68	0.09
booktomarket	1.44	1.14	0.83	1.09	7.65	6.12	4.97	6.36	0.00	0.00	0.14	0.00
First sort on: Book to Market												
R0 1	-1.74	-1.84	-2.10	-1.79	-9.09	-10.19	-10.33	-10.60	0.12	0.00	0.00	0.00
R1 11	1.98	1.35	1.10	1.56	9.39	6.54	6.11	8.54	0.00	0.00	0.00	0.00
R6 6	1.60	1.00	0.88	1.19	9.17	5.75	5.37	7.69	0.03	0.00	0.08	0.00
size	-1.03	-0.73	-1.33	-1.28	-3.49	-3.21	-5.14	-5.37	0.22	0.08	0.07	0.00
gross profitability	0.79	0.66	0.37	0.48	4.69	4.53	2.85	3.99	0.32	0.00	0.19	0.09
First sort on: Gross Profitability												
R0 1	-1.69	-1.95	-1.93	-1.79	-9.81	-9.74	-10.67	-10.60	0.00	0.00	0.02	0.00
R1 11	1.75	1.40	1.47	1.56	8.37	6.72	8.00	8.54	0.00	0.00	0.00	0.00

Continued on next page

Table A.1: *(continued)*

	Average return				t-statistics				p-value PT test			
Sorting on	Low	Mid	High	All	Low	Mid	High	All	Low	Mid	High	All
R6 6	1.48	0.99	1.02	1.19	8.28	5.45	6.49	7.69	0.00	0.00	0.01	0.00
size	-1.30	-1.35	-1.18	-1.28	-4.64	-5.55	-4.40	-5.37	0.08	0.06	0.04	0.00
booktomarket	1.49	1.11	1.00	1.09	7.03	6.36	5.44	6.36	0.00	0.27	0.00	0.00
First sort on: Size												
R0 1	-3.09	-1.74	-1.02	-1.79	-11.92	-9.13	-6.58	-10.60	0.00	0.00	0.09	0.00
R1 11	1.44	1.85	1.08	1.56	8.58	8.87	4.82	8.54	0.00	0.00	0.30	0.00
R6 6	0.96	1.27	1.04	1.19	5.71	7.20	5.42	7.69	0.06	0.00	0.00	0.00
gross profitability	0.14	0.68	0.40	0.48	0.87	4.43	2.95	3.99	0.06	0.23	0.23	0.09
booktomarket	0.63	1.05	0.53	1.09	3.74	5.44	2.81	6.36	1.00	0.02	0.00	0.00

A.2 Greedy algorithm

To illustrate estimation of a deep conditional portfolio sort start with the conditional portfolio sort in figure 1.3. Consider the portfolio S_1 in that figure which is defined by variable $R(g^{(1)}, 1)$ being less than threshold $\tau^{(1)}$ and variable $R(g^{(2a)}, 1)$ being smaller than threshold $\tau^{(2a)}$. Other portfolios can be defined similarly by their relations between sorting variables and associated thresholds. Within each portfolio S_l, the predicted expected return is just the average return, μ_l, of all firms in the portfolio, that is,

$$\hat{\mu}_l = \text{Mean}(r_{i,t+1} | \text{Firm i } \in S_l \text{ in period t}) \tag{A.1}$$

In other words, analogous to linear regression, we are interested in approximating the conditional mean of the outcome variable at a value of the regressor by the average of the outcome variable over observations with close values of the regressors. The conditional portfolio sort therefore generates subsets of firm observations that are more homogenous. Suppose for a moment that we have found such a homogenous allocation of firms into portfolios. The prediction function could then be written as

$$\hat{r}_{i,t+1} = \sum_{l=1}^{L} \hat{\mu}_l \mathbb{1}(\text{Firm i } \in S_l \text{ in period t}), \tag{A.2}$$

giving a portfolio-specific expected return prediction for each observation. What we have described so far is nothing more than a formal definition of the common conditional sorting methodology that we carried out in the previous section.

Of course, the conditional sort does not need to end after two levels but can be computed at greater depth. We consider the case in which the depth of the conditional sort, the sorting variables and associated thresholds are not pre-selected but need to be identified from the data. Finding the optimal solution to this problem requires solving an optimization problem that is NP complete (see (Hyafil and Rivest (1976))), that is, there does not exist a computationally fast solution to optimizing over both portfolios and predictions.

Instead, we adopt a *greedy algorithm* from the machine learning literature that proceeds in a step-wise fashion. Let $S_1(g, \tau)$ and $S_2(g, \tau)$ be two portfolios that are defined by a

firm's past return decile ranking $R(g,1)$ and a threshold value τ such that, as before, all observations for which $R(g,1) \leq \tau$ are in portfolio S_1, and all observations for which $R(g,1) > \tau$ are in portfolio S_2. At each node, all observations that are members of that node are split into two such portfolios. The greedy algorithm finds the past return characteristic $R(g,1)$ and the threshold value τ such that

$$(g^*, \tau^*) = \arg\min_{g,\tau} SC(g,\tau), \tag{A.3}$$

where $SC(g,\tau)$ is a *split criterion function* which we adopt from the related machine learning literature. The split criterion function selects the predictor variable and the associated threhold that minimize the sum of mean squared errors in the resulting portfolios with respect to the expected returns, that is,

$$SC(g,\tau) = \min_{\mu_1} \left(\sum_{R_{it}(g,1) \in S_1(g,\tau)} (r_{i,t+1} - \mu_1)^2 \right) + \min_{\mu_2} \left(\sum_{R_{it}(g,1) \in S_2(g,\tau)} (r_{i,t+1} - \mu_2)^2 \right) \tag{A.4}$$

and the inner minimizations are solved by equation (1.7).This algorithm reduces a complex non-linear estimation problem into subsets of simpler linear ones. The problem is solved in a brute-force fashion where the value of the split criterion function is computed for each firm characteristic and each threshold value. The optimization is repeated in each of the resulting portfolios until a. the number of observations in a node gets too small for further splits, or b. no variable provides a sufficient improvement of the mean squared error in equation (A.4). The result is a conditional portfolio sort with many levels.[4]

A.3 Robustness

We start by adding a set of eighty-six additional firm characteristics to the estimation and show that, again, the most recent returns are discovered as the most important ones. The

[4]The question of when to stop adding new levels to the conditional sort relates to a standard bias-variance trade-off. Using many levels potentially results in overfitting which would worsen the predictive power of equation (1.8) out of sample. Estimating only a few levels might miss important aspects of the data leading to bias. Within this sphere the number of levels can be chosen. We stop when the number of firms in a portfolio is smaller than 100 and make sure to validate all our estimates out of samples as described in section 1.3.2.

same holds true when we consider a much larger set of correlated past return variables; results are presented in section A.3.2. In all cases, we find that the derivatives and interactions are similar to our main results.

We then turn to the question of how our return term structure result varies across firm size categories. We repeat the analysis for three groups of stocks that are sorted by firm size first in section A.3.3.

A.3.1 Including firm characteristics

We investigate whether our results on the structural relation between future and past returns are robust to including other firm characteristics. For this paper, we focus on the changing nature of the return term structure result although the effect of firm characteristics (and the question of which of them can be found by our agnostic procedure) is an interesting one in itself.[5]

Going back to our original set of one-month return functions, we add eighty-six common firm characteristics, including size, book-to-market, gross profitability, earnings surprises, leverage and many more. The full set is described in detail in the appendix to Green *et al.* (2014); we generate the same set of variables from annual and quarterly data on firm fundamentals from Compustat, daily and monthly stock price data from CRSP, and earnings expectations and firm recommendations data from IBES.

Table A.2 mirrors table 1.4 and shows that the strategy returns are slightly higher when accounting variables are included in the deep conditional sorts. The out-of-sample performance is not better on all dimensions though: The information ratio decreases in this setting. The characteristics-augmented strategy appears to load higher on the size factor than the returns-only strategy in table 1.4 but loads similarly on other factors.

Table A.3 reports the top ten return-based predictor variables. The top ten return functions for rolling and entire period optimization, again, evolve around the most recent

[5]In ongoing work, a companion paper focuses exclusively on a large set of (mostly accounting- and earnings-based) firm characteristics.

Table A.2: *Strategy factor loadings: Including firm characteristics*

	(1)	(2)	(3)	(4)
Intercept	2.56	2.46	2.46	2.28
	(15.65)	(14.25)	(14.27)	(11.99)
MKT		0.12	0.06	0.10
		(2.75)	(1.61)	(2.44)
SMB			0.27	0.28
			(3.80)	(4.53)
HML			0.01	0.07
			(0.11)	(0.96)
UMD				0.18
				(3.44)
R^2		0.03	0.09	0.15
IR		2.66	2.75	2.63

This table shows time-series regressions of strategy returns on factors. Returns are specified in percent per month. Strategies are based on the predictions of a deep conditional portfolio sort that relates future returns to past decile sorts of returns and 86 firm fundamentals. Past return sorts include decile rankings R(g,l) with length l equal to 1 and gap g between 0 and 24 months (i.e. all one-month returns over the two years before portfolio formation). Strategies go long the highest predicted return decile and go short the lowest predicted return decile. The sample period covers 1968 to 2012, and all results are based on rolling out-of-sample estimates of the models. MKT is the market return, SMB and HML are the Fama-French factors for size and value, and UMD is the momentum factor. SR is the Sharpe ratio and IR is the information ratio. T-statistics are in parentheses, and standard errors were clustered using Newey-West's adjustment for serial correlation.

returns, and, apart from some changes in the exact order of predictor variables, are largely unaffected by the inclusion of other firm characteristics.

When comparing the top ten variables to the top ten in table 1.6, we see that nine out of ten show up in either list, with the exact ordering sometimes slightly altered. Recent returns are again the most important predictors.

Table A.3: *Most important predictor variables: Including firm characteristics*

Rolling optimization		Entire period optimization	
R(0,1)	1	R(0,1)	1
R(1,1)	0.5	R(1,1)	0.53
R(2,1)	0.45	R(2,1)	0.47
R(5,1)	0.44	R(3,1)	0.44
R(8,1)	0.41	R(11,1)	0.38
R(4,1)	0.4	R(5,1)	0.36
R(3,1)	0.39	R(6,1)	0.35
R(11,1)	0.39	R(7,1)	0.34
R(6,1)	0.38	R(8,1)	0.34
R(7,1)	0.37	R(4,1)	0.33

This table shows the most important return functions for the deep conditional portfolio sorts that use all one-month returns over the two years before portfolio formation and 86 additional firm characteristics. Results are shown for both the rolling model estimation and for optimization over the entire horizon. For rolling estimates, return functions are sorted by their median importance over forty-five years. Variable importance is measured as described in section 1.3.2.

The second row of average partial derivatives for the most recent one-month return

functions in figure 1.9 mirror the patterns from the first row that did not include firm characteristics. In particular, we observe stable linear relationships between past performance and prediction for returns that are further than four months in the past and for the most recent past return, but we also observe non-monotone or non-linear relationships for recent past returns in between.

Figure A.1 shows double partial derivatives for return characteristics when firm characteristics are included and corresponds to figure 1.10. In both cases, we observe patterns that are qualitatively very similar and only differ in details, e.g. the interaction between R(1,1) and R(3,1) is somewhat more pronounced.

Overall, we conclude that the discovered structure among return characteristics is largely unaffected by the inclusion of additional firm characteristics.

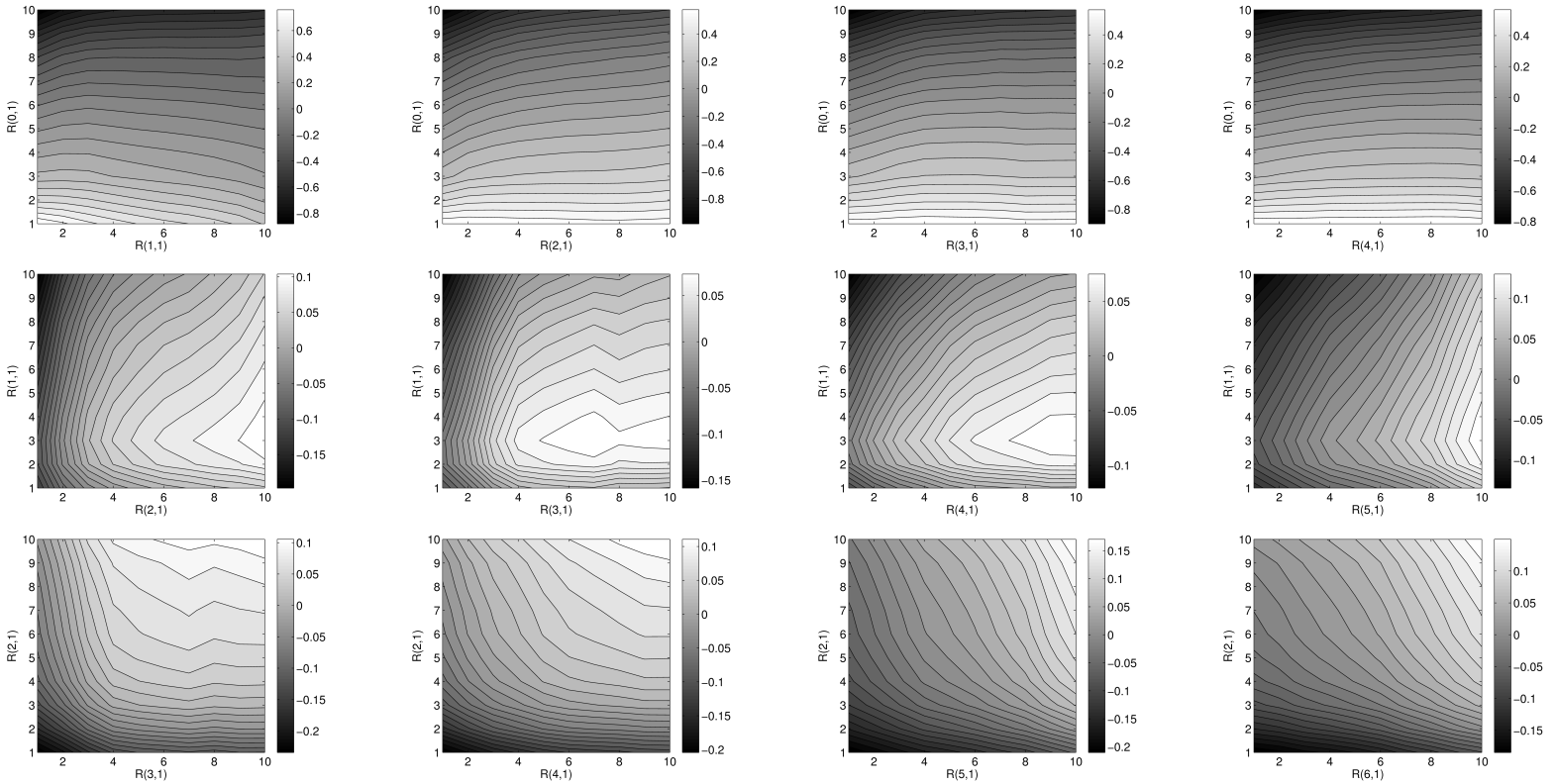


Figure A.1: *Average double partial derivatives: Firm characteristics included*

Notes: The figure shows the average prediction when two characteristics are counterfactually varied from low to high values. The figure shows results for return functions when 86 additional firm characteristics are included in the deep conditional portfolio sort. Results are based on rolling optimization of the model and predictions are averaged over the sample period. Details are in section 1.3.2.

A.3.2 Expanded set of return functions

In this section, we directly give the algorithm access to standard notions of momentum, R(1,11) in our notation, and other past return functions. More precisely, we define an *expanded* set of past return functions that includes return-based characteristics $\{R(g,l)\}, g = 0,\ldots,6; l = 1,\ldots,18$, that is, the set includes a total of 126 return-based predictor variables that are often highly correlated. Our main findings show that the algorithm derives its predictive power from optimally using the variation in relatively short-term returns. Is the algorithm just trying to re-create standard momentum? Or is there more information in the individual returns than in a summary return like R(1,11)?

We repeat all of the calculations above for an algorithm that has access to this expanded set of return functions. The first four columns of table A.4 show excess returns and factor loadings for the implementable trading strategy based on rolling estimation of deep conditional portfolio sorts. Excess returns are about as high as in table 1.4 where we used the smaller set of past return functions. The strategy's loadings on the value and size factors are similar to the loadings in table 1.4 while the loading on the momentum factor is somewhat higher. The four-factor model explains around 20% of the variation in the strategy return. The information ratio is again much higher than what is usually reported in the literature that employs methods that do not comprehensively deal with characteristics' interactions and non-linearities.

For comparison, the last four columns of table A.4 show (non-implementable) strategy returns on the hold-out sample for optimization over the entire horizon. Factor loadings are generally similar (with the exception of the value loading in the four factor model), and excess returns are higher which, again, should not be surprising given that optimization over the entire period uses contemporary information on cross-sectionally correlated stocks.

We have also looked at the factor loadings of the individual decile portfolios. Since the results are very similar to our previous results, we only briefly describe them here (available on request). Throughout all specifications, predicted return portfolios are positively correlated with market beta, with no apparent relation between decile portfolio and beta

Table A.4: *Strategy factor loadings: Expanded set of return functions*

	Rolling optimization				Entire Period optimization			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Intercept	2.37	2.31	2.36	2.10	3.13	3.04	3.05	2.84
	(17.65)	(17.22)	(18.51)	(17.08)	(22.48)	(21.05)	(21.12)	(19.28)
MKT		0.07	0.04	0.09		0.10	0.08	0.12
		(2.22)	(1.26)	(3.35)		(2.69)	(2.28)	(3.34)
SMB			0.06	0.07			0.07	0.07
			(0.91)	(1.36)			(1.20)	(1.19)
HML			-0.09	-0.01			-0.02	0.05
			(-1.36)	(-0.12)			(-0.33)	(0.90)
UMD				0.27				0.21
				(8.06)				(4.99)
R^2		0.02	0.03	0.23		0.02	0.03	0.11
IR		3.09	3.18	3.16		3.55	3.56	3.47

This table shows time-series regressions of strategy returns on factors. Returns are specified in percent per month. Strategies are based on the predictions of a deep conditional portfolio sort that relates future returns to past decile sorts of returns. Past returns include return-based functions R(g,l) with length $g = 1, \ldots, 18$ and gaps $g = 0, \ldots, 6$. The sample period covers 1968 to 2012. Results in columns 1-4 are based on rolling out-of-sample estimates of the model, results in column 5-8 are based on optimizing over the entire horizon. The strategy return used in columns 5-8 is computed on a hold-out sample of 30% of the data. MKT is the market return, SMB and HML are the Fama-French factors for size and value, and UMD is the momentum factor. IR is the information ratio. T-statistics are in parentheses, and standard errors are clustered using Newey-West's adjustment for serial correlation.

value. In the four-factor model, the loading on the market factor is around one for all deciles. Loadings on size seem to be non-monotonic but stronger in the extreme portfolios. Interestingly, only the portfolios based on entire sample predictions display a significant monotone relation with the value factor. We also find again the by now familiar result that return deciles are monotonically related to the momentum factor, which holds for all decile portfolios except for the highest one. In general, loadings on the momentum factor are low (around -.1) but significant.

Turning to predictor variable importance in table A.5, the deep conditional portfolio sorts recover recent past returns as the most important ones. The rolling optimization in column 1 yields that the ten most important return functions are all related to the most recent six months of returns and, what is more, the top seven return functions are returns of length one that, taken together, summarize the most recent six month return. Notice that all one-month return-based functions in the expanded set actually show up as the most important functions. The same result holds for optimization over the entire period in the second column. The momentum term structure result appears to hold on average over the entire period.

Recall that the return R(0,6), that is, the total return over the most recent six months, could have been chosen by the algorithm in the expanded set. The fact that this return is not chosen but its components are, illustrates that using the return over the previous year alone (and not the one-month returns that it is based on) leads to a loss of relevant information.

A.3.3 Estimation by size categories

We re-estimate the model for three separate size categories of firms. Following Fama and French (2008), we divide the sample of firms into three size categories based on NYSE breakpoints. Micro stocks are defined as the smallest 20% of companies by market value, small companies are the next 30% of companies, and the upper 50% make up the category of large firms. We repeat our analysis within each size category, and compute the most relevant predictor variables for both our standard set of one-month return variables and

Table A.5: *Most important predictor variables: Expanded set of return functions*

Rolling optimization		Entire period optimization	
R(0,1)	1	R(1,1)	1
R(1,1)	0.88	R(2,1)	0.8
R(2,1)	0.69	R(3,1)	0.76
R(6,1)	0.69	R(4,1)	0.74
R(3,1)	0.66	R(6,1)	0.72
R(4,1)	0.66	R(0,1)	0.62
R(5,1)	0.61	R(5,1)	0.62
R(0,2)	0.52	R(6,13)	0.43
R(1,2)	0.45	R(5,10)	0.42
R(1,3)	0.43	R(6,12)	0.42

This table shows the most important return functions for a deep conditional portfolio sort that uses past returns functions R(g,l) with length $g = 1, \ldots, 18$ and gaps $g = 0, \ldots, 6$. Results are shown for both the rolling model estimation and for optimization over the entire horizon. For rolling estimates, return functions are sorted by their median importance over forty-five years. Variable importance is measured as described in section 1.3.2.

for the expanded set of return functions from appendix A.3.2. Table A.6 shows that the most important predictor variables are remarkably consistent across size categories, apart from some variations in exact rank of each predictor variable. Furthermore, most predictor variables in both sets relate to relatively recent returns.

Table A.6: *Most important predictor variables: Within size category*

One-month returns						Expanded set					
Micro		Small		Big		Micro		Small		Big	
R(0,1)	1	R(0,1)	1	R(0,1)	0.99	R(0,1)	1	R(0,1)	1	R(1,1)	1
R(2,1)	0.54	R(2,1)	0.75	R(3,1)	0.72	R(1,1)	0.72	R(1,1)	0.65	R(2,1)	0.7
R(5,1)	0.51	R(4,1)	0.6	R(4,1)	0.6	R(3,1)	0.63	R(0,2)	0.65	R(3,1)	0.66
R(1,1)	0.5	R(1,1)	0.59	R(8,1)	0.6	R(5,1)	0.62	R(2,1)	0.63	R(5,1)	0.66
R(3,1)	0.5	R(3,1)	0.59	R(1,1)	0.51	R(2,1)	0.61	R(5,1)	0.56	R(6,1)	0.66
R(6,1)	0.49	R(5,1)	0.58	R(2,1)	0.5	R(6,1)	0.59	R(6,1)	0.56	R(4,1)	0.6
R(4,1)	0.42	R(6,1)	0.55	R(5,1)	0.5	R(0,2)	0.56	R(3,1)	0.52	R(0,1)	0.57
R(11,1)	0.4	R(7,1)	0.52	R(9,1)	0.5	R(4,1)	0.51	R(4,1)	0.5	R(0,2)	0.42
R(7,1)	0.39	R(8,1)	0.47	R(11,1)	0.5	R(1,2)	0.42	R(2,2)	0.43	R(6,4)	0.34
R(8,1)	0.38	R(23,1)	0.43	R(18,1)	0.5	R(0,3)	0.38	R(3,2)	0.43	R(6,6)	0.34

This table shows predictor variable importance in portfolios that are first sorted on size. Micro stocks are defined as the smallest 20% of companies by market value, small companies are the next 30% of companies, and the upper 50% make up the category of large firms. One-month returns include all one-month returns over the two years before portfolio formation. The expanded set consists of 126 return-based characteristics R(g,l) over the two years before portfolio formation with length $g = 1, \ldots, 18$ and gaps $g = 0, \ldots, 6$. Results are shown for both the rolling model estimation and for optimization over the entire horizon. For rolling estimates, return functions are sorted by their median importance over forty-five years. Variable importance is measured as described in section 1.3.2.