

Dissecting Characteristics Nonparametrically*

Joachim Freyberger[†] Andreas Neuhierl[‡] Michael Weber[§]

Preliminary and incomplete - Please do not circulate
Comments welcome!

This version: August 2016

Abstract

We propose a nonparametric methodology to test which characteristics provide independent information for the cross section of expected returns. We use the adaptive group LASSO to select characteristics simultaneously and estimate how they affect expected returns nonparametrically. Our method can handle a large number of characteristics, allows for a flexible functional form, and is insensitive to outliers. Many of the previously identified return predictors do not provide incremental information for expected returns, and nonlinearities are important. Our proposed methodology has higher out-of-sample explanatory power compared to linear panel regressions, and increases Sharpe ratios by 70%.

JEL classification: C14, C52, C58, G12

Keywords: Cross Section of Returns, Anomalies, Expected Returns, Model Selection

*We thank Jason Chen, Gene Fama, Ken French, Bryan Kelly, Leonid Kogan, Jon Lewellen, Stefan Nagel, Stavros Panageas, Ľuboš Pástor, and seminar participants at Dartmouth College, Tsinghua University PBCSF, Tsinghua University SEM, the University of Chicago, the University of Notre Dame, and the University of Washington for valuable comments. Weber gratefully acknowledges financial support from the University of Chicago, the Neubauer Family Foundation, and the Fama–Miller Center.

[†]University of Wisconsin - Madison, Madison, WI, e-Mail: jfreyberger@ssc.wisc.edu

[‡]University of Notre Dame, Notre Dame, IN, USA. e-Mail: aneuhier@nd.edu

[§]Booth School of Business, University of Chicago, Chicago, IL, USA. e-Mail: michael.weber@chicagobooth.edu.

I Introduction

In his presidential address, Cochrane (2011) argues the “cross section” of the expected return is in disarray. Harvey et al. (2016) identify more than 300 published factors that have predictive power for the cross section of expected returns.¹ Many economic models, such as the consumption CAPM of Lucas Jr (1978), Breeden (1979), and Rubinstein (1976), instead predict that only a small number of factors can summarize cross-sectional variation in expected returns.

Researchers typically employ two methods to identify return predictors: (i) (conditional) portfolio sorts based on one or multiple characteristics such as size or book-to-market, and (ii) linear regression in the spirit of Fama and MacBeth (1973). Both methods have many important applications, but they fall short in what Cochrane (2011) calls the multidimensional challenge: “[W]hich characteristics really provide *independent* information about average returns? Which are subsumed by others?” Both methods are subject to the curse of dimensionality when the number of characteristics is large relative to the number of stocks, and linear regressions make strong functional-form assumptions and are sensitive to outliers.² Cochrane (2011) speculates, “To address these questions in the zoo of new variables, I suspect we will have to use different methods.”

We propose a nonparametric methodology to determine which firm characteristics provide independent information for the cross section of expected returns without making strong functional-form assumptions. Specifically, we use a group LASSO (least absolute shrinkage and selection operator) procedure suggested by Huang, Horowitz, and Wei (2010) for model selection and nonparametric estimation. Model selection deals with the question of which characteristics have incremental predictive power for expected returns, given the other characteristics. Nonparametric estimation deals with estimating the effect of important characteristics on expected returns without imposing a strong functional-form.³

We show three applications of our proposed framework. First, we study which

¹Figure 2 documents the number of discovered factors over time.

²We discuss these and related concerns in section II and compare current methods to our proposed framework in section III.

³In our empirical application, we estimate quadratic splines.

characteristics provide independent information for the cross section of expected returns. We estimate our model on 24 characteristics including size, book-to-market, beta, and other prominent variables and anomalies on a sample period from July 1963 to June 2015. Only eight variables, including size, idiosyncratic volatility, and return-based predictors, have independent explanatory power for expected returns for the full sample and all stocks. We find similar results when we split the sample and estimate the model in the early or later part. For stocks whose market capitalization is above the 20% NYSE size percentile, only book-to-market, investment, idiosyncratic volatility, and past returns remain significant return predictors.

Second, we compare the out-of-sample performance of the nonparametric model to a linear model. We estimate both models over a period until 1990 and select significant return predictors. We then create rolling monthly return predictions and construct a hedge portfolio going long stocks with the 10% highest predicted returns and shorting stocks with the 10% lowest predicted returns. The nonparametric model generates an average Sharpe ratio of 1.72 compared to 0.97 for the linear model.⁴ The linear model selects substantially more characteristics in sample but performs worse out of sample.

Third, we study whether the predictive power of characteristics for expected returns varies over time. We estimate the model using 120 months of data on all characteristics we select in our baseline analysis, and then estimate rolling one-month-ahead return forecasts. We find substantial time variation in the predictive power of characteristics for expected returns. As an example, momentum returns conditional on other return predictors vary substantially over time, and we find a momentum crash similar to Daniel and Moskowitz (2016) as past losers appreciated during the recent financial crisis.

A Related Literature

The capital asset pricing model (CAPM) of Sharpe (1964), Lintner (1965), and Mossin (1966) predicts an asset’s beta with respect to the market portfolio is a sufficient statistic for the cross section of expected returns. Fama and MacBeth (1973) provide empirical support for the CAPM. Subsequently, researchers identified that many variables, such as

⁴The linear model we estimate and the results are similar to Lewellen (2015).

size (Banz (1981)), the book-to-market ratio (Rosenberg et al. (1985)), leverage (Bhandari (1988)), earnings-to-price ratios (Basu (1983)) or past returns (Jegadeesh and Titman (1993)), contain additional independent information for expected returns. Sorting stocks into portfolio based on these characteristics often led to rejection of the CAPM because the spread in CAPM betas could not explain the spread in returns. Fama and French (1992) synthesize these findings and Fama and French (1993) show that a three-factor model with the market return, a size, and a value factor can explain cross sections of stocks sorted on characteristics (with the exception of momentum) that appeared anomalous relative to the CAPM. In this sense, Fama and French (1992) and Fama and French (1996) achieve a significant dimension reduction: researchers who want to explain the cross section of stock returns only have to explain the size and value factors. Daniel and Titman (1997), on the contrary, argue that characteristics have higher explanatory power for the cross section of expected returns than loadings on pervasive risk factors.

In the 20 years that followed, many researchers joined a “fishing expedition” to identify characteristics and factor exposures the three-factor model cannot explain. Harvey et al. (2016) provide an overview of this literature and list over 300 published papers that study the cross section of expected returns. They propose a t -statistic of 3 for new factors to account for multiple testing on a common data set. Figure 3 shows the suggested adjustment over time. However, even employing the higher threshold for the t -statistic still leaves approximately 150 characteristics as useful predictors for the cross section of expected returns.

The large number of significant predictors is not a shortcoming of Harvey et al. (2016), who address the issue of multiple testing. Instead, authors in this literature usually consider their proposed return predictor in isolation without conditioning on previously discovered return predictors. Haugen and Baker (1996) and Lewellen (2015) are notable exceptions. They employ Fama and MacBeth (1973) regressions to combine the information in multiple characteristics. Lewellen (2015) jointly studies predictive power of 15 characteristics and finds only a few are significant predictors for the cross section of expected returns. Although Fama-MacBeth regressions carry a lot of intuition, they do not offer a formal method to select significant return predictors. We build on this work

and provide a framework that allows for nonlinear association between characteristics and returns, provide a formal framework to disentangle significant from insignificant return predictors, and study many more characteristics.

II Current Methodology

A Expected Returns and the Curse of Dimensionality

One aim of the empirical asset pricing literature is to identify characteristics that predict expected returns, that is, find a characteristic C in period $t-1$ that predicts excess returns of firm i next period, R_{it} . Formally, we try to describe the conditional mean function,

$$E[R_{it} \mid C_{it-1}]. \quad (1)$$

We often use portfolio sorts to approximate equation (1). We typically sort stocks into 10 portfolios and compare mean returns across portfolios. Portfolio sorts are simple, straightforward, and intuitive, but they also suffer from several shortcomings. First, we can only use portfolio sorts to analyze a small set of characteristics. Imagine sorting stocks jointly into five portfolios based on CAPM beta, size, book-to-market, profitability, and investment. We would end up with $5^5 = 3125$ portfolios, which is larger than the number of stocks at the beginning of our sample.⁵ Second, portfolio sorts offer little formal guidance to discriminate between characteristics. Fama and French (2008) call this second shortcoming “awkward.” Third, we implicitly assume expected returns are constant over a part of the characteristic distribution, such as the smallest 10% of stocks, when we use portfolio sorts as an estimator of the conditional mean function. Fama and French (2008) call this third shortcoming “clumsy.”⁶ Nonetheless, portfolio sorts are by far the most commonly used technique to analyze which characteristics have predictive power for expected returns.

⁵The curse of dimensionality is a well-understood shortcoming of portfolio sorts. See Fama and French (2015) for a recent discussion in the context of the factor construction for their five-factor model. They also argue not-well-diversified portfolios have little power in asset pricing tests.

⁶Portfolio sorts are a restricted form of nonparametric regression. We will use the similarities of portfolio sorts and nonparametric regressions to develop intuition for our proposed framework below.

An alternative to portfolio sorts is to *assume* linearity of equation (1) and run linear panel regressions of excess returns on S characteristics, namely,

$$R_{it} = \alpha + \sum_{s=1}^S \beta_s C_{s,it-1} + \varepsilon_{it}. \quad (2)$$

Linear regressions allow us to study the predictive power for expected returns of many characteristics jointly, but they also have potential pitfalls. First, no a priori reason explains why the conditional mean function should be linear.⁷ Fama and French (2008) estimate linear regressions as in equation (2) to dissect anomalies, but raise concerns over potential nonlinearities. They make ad hoc adjustments and use, for example, the log book-to-market ratio as a predictive variable. Second, linear regressions are sensitive to outliers. Third, small, illiquid stocks might have a large influence on point estimates because they represent the majority of stocks. Researchers often use ad hoc techniques to mitigate concerns related to microcaps and outliers, such as winsorizing observations and estimating linear regressions separately for small and large stocks (see Lewellen (2015) for a recent example).

Cochrane (2011) synthesizes many of the challenges portfolio sorts and linear regressions face in the context of many return predictors, and suspects “we will have to use different methods.”

B Equivalence between Portfolio Sorts and Regressions

Cochrane (2011) conjectures in his presidential address, “[P]ortfolio sorts are really the same thing as nonparametric cross-sectional regressions, using nonoverlapping histogram weights.” Additional assumptions are necessary to show a formal equivalence, but his conjecture contains valuable intuition to model the conditional mean function formally. We first show a formal equivalence between portfolio sorts and regressions and then use the equivalence to motivate the use of nonparametric methods.

Suppose we observe excess returns R_{it} and a single characteristic C_{it-1} for stocks

⁷Fama and MacBeth (1973) regressions also assume a linear relationship between expected returns and characteristics. Fama-MacBeth point estimates are numerically equivalent to estimates from equation (2) when characteristics are constant over time.

$i = 1, \dots, N_t$ and time periods $t = 1, \dots, T$. We sort stocks into L portfolios depending on the value of the lagged characteristic, C_{it-1} . Specifically, stock i is in portfolio l at time t if $C_{it-1} \in I_{tl}$, where I_{tl} indicates an interval of the distribution for a given firm characteristic. For example, take a firm with lagged market cap in the 45th percentile of the firm size distribution. We would sort that stock in the 5th out of 10 portfolios in period t . For each time period t , let N_{tl} be the number of stocks in portfolio l ,

$$N_{tl} = \sum_{i=1}^{N_t} \mathbf{1}(C_{it-1} \in I_{tl}).$$

The excess return of portfolio l at time t , P_{tl} , is then

$$P_{tl} = \frac{1}{N_{tl}} \sum_{i=1}^N R_{it} \mathbf{1}(C_{it-1} \in I_{tl}).$$

The difference in average excess returns between portfolios l and l' , or the excess return $e(l, l')$, is

$$e(l, l') = \frac{1}{T} \sum_{t=1}^T (P_{tl} - P_{tl'}),$$

which is the intercept in a (time-series) regression of the difference in portfolio returns, $P_{tl} - P_{tl'}$, on a constant.⁸

Alternatively, we can run a pooled time-series–cross-sectional regression of excess returns on dummy variables, which equal 1 if firm i is in portfolio l in period t . We denote the dummy variables by $\mathbf{1}(C_{it-1} \in I_{tl})$ and write

$$R_{it} = \sum_{l=1}^L \beta_l \mathbf{1}(C_{it-1} \in I_{tl}) + \varepsilon_{it}.$$

Let \mathcal{R} be the $NT \times 1$ vector of excess returns and let X be the $NT \times L$ matrix of dummy variables, $\mathbf{1}(C_{it-1} \in I_{tl})$. Let $\hat{\beta}$ be an OLS estimate,

$$\hat{\beta} = (X'X)^{-1} X'\mathcal{R}.$$

⁸We only consider univariate portfolio sorts in this example to gain intuition.

It then follows that

$$\begin{aligned}
\hat{\beta}_l &= \frac{1}{\sum_{t=1}^T \sum_{i=1}^N \mathbf{1}(C_{it-1} \in I_{tl})} \sum_{t=1}^T \sum_{i=1}^N R_{it} \mathbf{1}(C_{it-1} \in I_{tl}) \\
&= \frac{1}{\sum_{t=1}^T N_{tl}} \sum_{t=1}^T \sum_{i=1}^N R_{it} \mathbf{1}(C_{it-1} \in I_{tl}) \\
&= \frac{1}{\sum_{t=1}^T N_{tl}} \sum_{t=1}^T N_{tl} P_{tl} \\
&= \frac{1}{T} \sum_{t=1}^T \frac{N_{tl}}{\frac{1}{T} \sum_{t=1}^T N_{tl}} P_{tl}.
\end{aligned}$$

Now suppose we have the same number of stocks in each portfolio l for each time period t , that is, $N_{tl} = \bar{N}_l$ for all t . Then

$$\hat{\beta}_l = \frac{1}{T} \sum_{t=1}^T P_{tl}$$

and

$$\hat{\beta}_l - \hat{\beta}_{l'} = \frac{1}{T} \sum_{t=1}^T (P_{tl} - P_{tl'}) = e(l, l').$$

Hence, the slope coefficients in pooled time-series–cross-sectional regressions are equivalent to average portfolio returns, and the difference between two slope coefficients is the excess return between two portfolios.

If the number of stocks in the portfolios changes over time, then portfolio sorts and regressions typically differ. We can restore equivalence in two ways. First, we could take the different number of stocks in portfolio l over time into account when we calculate averages, and define excess return as

$$e^*(l, l') = \frac{1}{\sum_{t=1}^T N_{tl}} \sum_{t=1}^T N_{tl} P_{tl} - \frac{1}{\sum_{t=1}^T N_{tl'}} \sum_{t=1}^T N_{tl'} P_{tl'},$$

in which case, we again get $\hat{\beta}_l - \hat{\beta}_{l'} = e^*(l, l')$.

Second, we could use the weighted least squares estimator,

$$\tilde{\beta} = (X'WX)^{-1}X'W\mathcal{R},$$

where the $NT \times NT$ weight matrix W is a diagonal matrix with the inverse number of stocks on the diagonal, $\text{diag}(1/N_{tl})$. With this estimator, we again get $\tilde{\beta}_l - \tilde{\beta}_{l'} = e(l, l')$.

III Nonparametric Estimation

We now use the relationship between portfolio sorts and regressions to develop intuition for our nonparametric estimator, and show how we can interpret portfolio sorts as a special case of nonparametric estimation. We then show how to select characteristics with independent information for expected returns within that framework.

Suppose we knew the conditional mean function $m_t(c) \equiv E[R_{it} \mid C_{it-1} = c]$.⁹ Then,

$$E[R_{it} \mid C_{it-1} \in I_{tl}] = \int_{I_{tl}} m_t(c) f_{C_{it-1} \mid C_{it-1} \in I_{tl}}(c) dc,$$

where $f_{C_{it-1} \mid C_{it-1} \in I_{tl}}$ is the density function of the characteristic in period $t-1$, conditional on $C_{it-1} \in I_{tl}$. Hence, to obtain the expected return of portfolio l , we can simply integrate the conditional mean function over the appropriate interval of the characteristic distribution. Therefore, the conditional mean function contains all information for portfolio returns. However, knowing $m_t(c)$ provides additional information about nonlinearities in the relationship between expected returns and characteristics, and the functional form more generally.

To estimate the conditional mean function, m_t , consider again regressing excess returns, R_{it} , on L dummy variables, $\mathbf{1}(C_{it-1} \in I_{tl})$,

$$R_{it} = \sum_{l=1}^L \beta_l \mathbf{1}(C_{it-1} \in I_{tl}) + \varepsilon_{it}.$$

In nonparametric estimation, we call indicator functions of the form $\mathbf{1}(C_{it-1} \in I_{tl})$

⁹We take the expected excess return for a fixed time period t .

constant splines. Estimating the conditional mean function, m_t , with constant splines, means we approximate it by a step function. In this sense, portfolio sorting is a special case of nonparametric regression when the number of portfolios approaches infinity. A step function is nonsmooth and therefore has undesirable theoretical properties as a nonparametric estimator, but we build on this intuition to estimate m_t nonparametrically.

Figures 4–6 illustrate the intuition behind the relationship between portfolio sorts and nonparametric regressions. These figures show returns on the y-axis and book-to-market ratios on the x-axis, as well as portfolio returns and the nonparametric estimator we propose below for simulated data.

We see in Figure 4 that most of the dispersion in book-to-market ratios and returns is in the extreme portfolios. Little variation occurs in returns across portfolios 2-4 in line with empirical settings (see Fama and French (2008)). Portfolio means offer a good approximation of the conditional mean function for intermediate portfolios. We also see, however, that portfolios 1 and 5 have difficulty capturing the nonlinearities we see in the data.

Figure 5 documents that a nonparametric estimator of the conditional mean function provides a good approximation for the relationship between book-to-market ratios for intermediate values of the characteristic but also in the extremes of the distribution.

Finally, we see in Figure 6 that portfolio means provide a better fit in the tails of the distribution once we allow for more portfolios. The predictions from the nonparametric estimator and portfolio mean returns become more comparable with the larger number of portfolios.

A Multiple Regression & Additive Conditional Mean Function

Both portfolio sorts and regressions theoretically allow us to look at several characteristics simultaneously. Consider small (S) and big (B) firms and value (V) and growth (G) firms. We could now study four portfolios: (SV) , (SG) , (BV) , (BG) . However, portfolio sorts quickly become unfeasible as the number of characteristics increases. For example, if we have four characteristics and partition each characteristics into five portfolios, we end up with $5^4 = 625$ portfolios. First, analyzing 625 portfolio returns would be impractical.

Second, as the number of characteristics increases, we will only have very few observations in each portfolio.

In nonparametric regressions, an analogous problem arises. Estimating the conditional mean function $m_t(c) \equiv E[R_{it} \mid C_{it} = c]$ fully nonparametrically with many regressors results in a slow rate of convergence and imprecise estimates in practice.¹⁰ Specifically, with S characteristics and N_t observations, the optimal rate of convergence in mean square is $N_t^{-4/(4+S)}$, which is always smaller than the rate of convergence for the parametric estimator of N_t^{-1} . Notice the rate of convergence decreases as S increases.¹¹ Consequently, we get an estimator with poor finite sample properties if the number of characteristics is large.

As an illustration, suppose we observe one characteristic, in which case, the rate of convergence is $N_t^{-4/5}$. Now suppose instead we have 11 characteristics, and let N_t^* be the number of observations necessary to get the same rate of convergence as in the case with one characteristic. We get,

$$(N_t^*)^{-4/15} = N_t^{-4/5} \Rightarrow N_t^* = N_t^3.$$

Hence, in the case with 11 characteristics, we have to raise the sample size to the power of 3 to obtain the same rate of convergence and finite sample properties as in the case with only one characteristic. Consider a sample size, N_t , of 1,000. Then, we would need 1 billion return observations to obtain similar finite sample properties of an estimated conditional mean function with 11 characteristics.

Conversely, suppose $S = 11$ and we have $N_t^* = 1,000$ observations. This combination yields similar properties as an estimation with one characteristic and a sample size $N_t = (N_t^*)^{1/3}$ of 10.

Nevertheless, if we are interested in which characteristics provide incremental information for expected returns given other characteristics, we cannot look at each characteristic in isolation. A natural solution in the nonparametric regression framework

¹⁰This literature refers to this phenomenon as the “curse of dimensionality” (see Stone (1982) for a formal treatment).

¹¹Note we assume the conditional mean function m_t is twice continuously differentiable.

is to assume an additive model:

$$m_t(c_1, \dots, c_S) = \sum_{s=1}^S m_{ts}(c_s),$$

where $m_{ts}(\cdot)$ are unknown functions. The main theoretical advantage of the additive specification is that the rate of convergence is always $N_t^{-4/5}$, which does not depend on the number of characteristics S (see Stone (1985), Stone (1986), and Horowitz et al. (2006)).

An important restriction of the additive model is

$$\frac{\partial^2 m_t(c_1, \dots, c_S)}{\partial c_s \partial c_{s'}} = 0$$

for all $s \neq s'$; therefore, the additive model does not allow for interactions between characteristics. Thus the predictive power of the book-to-market ratio for expected returns does not depend on firm size to give an example. One way around this shortcoming is to add certain interactions as additional regressors. For instance, we could interact every characteristic with size to see if small firms are really different. An alternative solution is to estimate the model separately for small and large stocks.

Although the assumption of an additive model is somewhat restrictive, it provides desirable econometric advantages and is far less restrictive than assuming linearity right away as we do in Fama-MacBeth regressions. Another major advantage of an additive model is that we can jointly estimate the model for a large number of characteristics, select important characteristics, and estimate the summands of the conditional mean function, m_t , simultaneously, as we explain in section *D*.

***B* Comparison of Linear & Nonparametric Models**

We now want to compare portfolio sorts and a linear model with nonparametric models in some specific numerical examples. The comparison helps us understand the potential pitfalls from assuming a linear relationship between characteristics and returns, and gain some intuition for why we might select different characteristics in a linear model in our

empirical tests in section V.

Suppose we observe excess returns R_{it} and a single characteristic, C_{it-1} distributed according to $C_{it-1} \sim U[0, 1]$ for $i = 1, \dots, N$ and $t = 1, \dots, T$ with the data generating process,

$$R_{it} = m_t(C_{it-1}) + \varepsilon_{it},$$

where $E[\varepsilon_{it} | C_{it-1}] = 0$.

Without knowing the conditional mean function m_t , we could sort stocks into portfolios according to the distribution of the characteristic. C_{it-1} predicts returns if mean returns differ significantly across portfolios. For example, we could construct 10 portfolios based on the percentiles of the distribution and test if the first portfolio has a significantly different return than the 10th portfolio.

If we knew the conditional mean function m_t , we could conclude C_{it-1} predicts returns if m_t is not constant on $[0, 1]$. Moreover, knowing the conditional mean function allows us to construct portfolios with a large spread in returns. Instead of sorting stocks based on their values of the characteristic C_{it-1} , we could sort stocks directly based on the conditional mean function $m_t(C_{it-1})$. For example, let $q_t(\alpha)$ be the α -quantile of $m_t(C_{it-1})$ and let stock i be in portfolio l at time t if $m_t(C_{it-1}) \in [q_t((l-1)/10), q_t(l/10)]$. That is, we construct 10 portfolios based on return predictions. Portfolio 1 contains the 10% of stocks with the lowest predicted returns, and portfolio 10 contains the 10% of stocks with the highest predicted returns.

If m_t is monotone and we only study a single characteristic, both sorting based on the value of the characteristic and based on predicted returns $m_t(C_{it-1})$ results in the same portfolios. However, if m_t is not monotone, the “10-1 portfolio” return is higher when we sort based on $m_t(C_{it-1})$.

As a simple example, suppose $m_t(c) = (c - 0.5)^2$. Then the expected “10-1 portfolio” return when sorting based on characteristic, C_{it-1} , is 0.

We now consider two characteristics, $C_{1,it-1} \sim U[0, 1]$, and $C_{2,it-1} \sim U[0, 1]$ and assume the following data-generating process:

$$R_{it} = m_{t1}(C_{1,it-1}) + m_{t2}(C_{2,it-1}) + \varepsilon_{it},$$

where $E[\varepsilon_{it} \mid C_{1,it-1}, C_{2,it-1}] = 0$. Again, we can construct portfolios with a large spread in predicted returns based on the value of the conditional mean function, m_t . The idea is similar to constructing trading strategies based on the predicted values of a linear model,

$$R_{it} = \beta_0 + \beta_1 C_{1,it-1} + \beta_2 C_{2,it-1} + \varepsilon_{it}.$$

We will now, however, illustrate the potential pitfalls of the linear model and how a nonparametric model can alleviate them.

Assume the following return-generating process:

$$R_{it} = -0.2 + 0.3\sqrt{C_{1,it-1}} + 0.25C_{2,it-1}^2 + \varepsilon_{it}.$$

In this example, a regression of returns R_{it} on the characteristics $C_{1,it-1}$ and $C_{2,it-1}$ yields slope coefficients of around 0.25 in large samples. Therefore, the predicted values of a linear model treat $C_{1,it-1}$ and $C_{2,it-1}$ almost identically, although they affect returns very differently.

We now compare the performance of the linear and nonparametric model for the “10-1” hedge portfolio. The table below shows monthly returns, standard deviations, and Sharpe ratios from a simulation for 2,000 stocks and 240 periods for both models.¹²

Predicted returns for the nonparametric model are slightly higher compared to the linear model, with almost identical standard deviations resulting in larger Sharpe ratios with the nonparametric method. Nevertheless, the linear model is a good approximation in this example, and the nonparametric method improves only marginally on the linear model.

¹²The numbers in the table are averages of portfolio means, standard deviations, and Sharpe ratios of 1,000 simulated data sets. We use the first 120 periods to estimate the conditional functions using the adaptive group LASSO, which we explain below, and form portfolios for each remaining period based on the estimates. Therefore, the portfolio means in the table are based on 120 time periods.

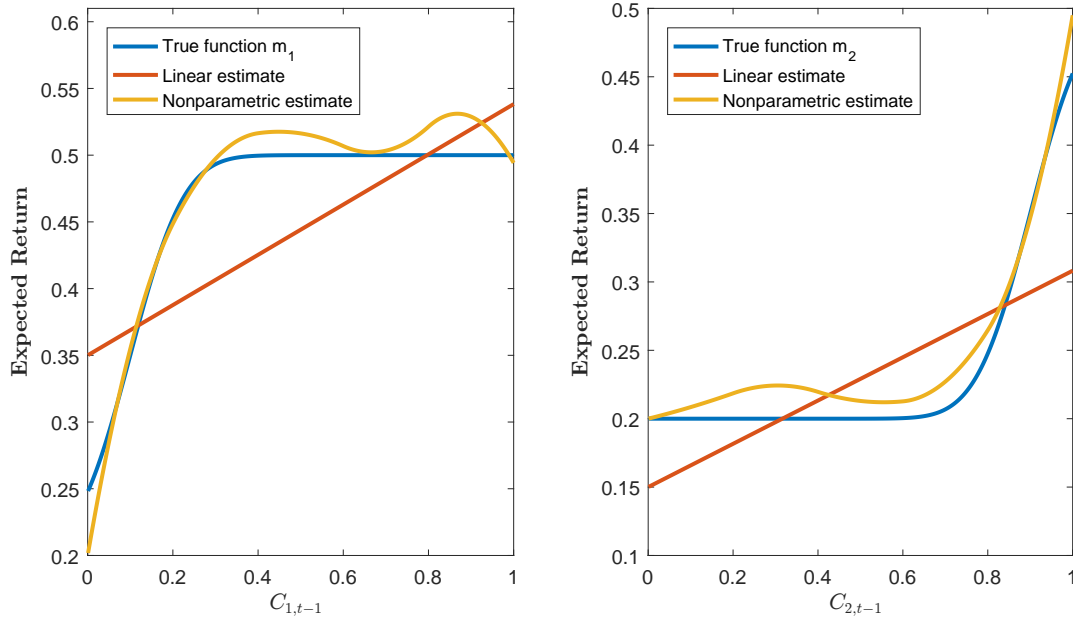


Figure 1: Regression functions and estimates

	Linear	Nonparametric
Return	0.1704	0.1734
Std	0.2055	0.2054
Sharpe Ratio	0.8329	0.8480

Instead, now we study the following data-generating process:

$$R_{it} = -0.3 + 0.3\Phi((C_{1,it-1} - 0.1)/0.1) + 0.3\Phi((C_{2,it-1} - 0.9)/0.1) + \varepsilon_{it},$$

where Φ denotes the standard normal cdf. Figure 1 plots the two functions, along with a parametric and a nonparametric estimate for a representative data set.

In this example, a regression of R_{it} on $C_{1,it-1}$ and $C_{2,it-1}$ yields two slope coefficients of around 0.15. Hence, as in the previous example, the predicted values of a linear model treat $C_{1,it-1}$ and $C_{2,it-1}$ identically.

	Linear	Nonparametric
Return	0.1154	0.1863
Std	0.1576	0.1576
Sharpe Ratio	0.7352	1.1876

The portfolio returns using the nonparametric model are now substantially higher compared to the linear model, with almost identical standard deviations, resulting in much larger Sharpe ratios. In this example, the linear model is a poor approximation of the true data-generating process.

Note that we do not know the true data-generating process, and the linear model may provide a good or poor approximation. Therefore, nonparametric methods are the natural choice.

In a last example, we want to discuss how the linear and nonparametric models treat nonlinear transformations of variables. This example helps us understand why a linear model might select more variables in empirical settings. Consider the following data-generating process:

$$R_{it} = C_{1,it-1} + C_{2,it-1} + \varepsilon_{it},$$

with $C_{2,it-1} = C_{1,it-1}^2$; that is, the second characteristic is just the square of the first characteristic. In the linear model, both characteristics are important to describe the conditional mean function, whereas in the nonparametric model, m_t is a function of $C_{1,it-1}$ only (or alternatively, $C_{2,it-1}$ only). In section *D*, we consider model selection next to estimation, and these differences between the linear and the nonparametric model will play an important role.

C Normalization of Characteristics

We now describe a suitable normalization of the characteristics, which will allow us to map our nonparametric estimator directly to portfolio sorts. As before, define the conditional mean function m_t for S characteristics as

$$m_t(C_{1,it-1}, \dots, C_{S,it-1}) = E[R_{it} \mid C_{1,it-1}, \dots, C_{S,it-1}].$$

For each characteristic s , let $F_{s,t}(\cdot)$ be a known strictly monotone function and denote its inverse by $F_{s,t}^{-1}(\cdot)$. Define $\tilde{C}_{s,it-1} = F_{s,t}(C_{s,it-1})$ and

$$\tilde{m}_t(c_1, \dots, c_S) = m_t(F_{1,t}^{-1}(c_1), \dots, F_{S,t}^{-1}(c_S)).$$

Then,

$$m_t(C_{1,it-1}, \dots, C_{S,it-1}) = \tilde{m}_t(\tilde{C}_{1,it-1}, \dots, \tilde{C}_{S,it-1}).$$

Knowledge of the conditional mean function m_t is equivalent to knowing the transformed conditional mean function \tilde{m}_t . Moreover, using a transformation does not impose any additional restrictions and is therefore without loss of generality. Instead of estimating m_t , we will estimate \tilde{m}_t for a rank transformation that has desirable properties and nicely maps to portfolio sorting.

When we sort stocks into portfolios, we are typically not interested in the value of a characteristic in isolation, but rather about the rank of the characteristic in the cross section. Consider firm size. Size grows over time, and a firm with a market capitalization of USD 1 billion in the 1960s was considered a large firm, but today is not. Our normalization considers the relative size in the cross section rather than the absolute size, similar to portfolio sorting.

Hence, we choose the rank transformation of $C_{s,it-1}$ such that the cross-sectional distribution of a given characteristic lies in the unit interval; that is, $C_{s,it-1} \in [0, 1]$. Specifically, let

$$F_{s,t}(C_{s,it-1}) = \frac{\text{rank}(C_{s,it-1})}{N_t + 1}.$$

Here, $\text{rank}(\min_{i=1, \dots, N_t} C_{s,it-1}) = 1$ and $\text{rank}(\max_{i=1, \dots, N_t} C_{s,it-1}) = N_t$. Therefore, the α quantile of $\tilde{C}_{s,it-1}$ is α . We use this particular transformation because portfolio sorting maps into our estimator as a special case.¹³

Although knowing m_t is equivalent to knowing \tilde{m}_t , in finite samples, the estimates

¹³The general econometric theory we discuss in section *D* (model selection, consistency, etc.) also applies to any other monotonic transformation or the non-transformed conditional mean function.

of the two typically differ; that is,

$$\widehat{m}_t(c_1, \dots, c_S) \neq \widehat{m}_t(F_{1,t}^{-1}(c_1), \dots, F_{S,t}^{-1}(c_S)).$$

In numerical simulations and in the empirical application, we found \tilde{m}_t yields better out-of-sample predictions than m_t . The transformed estimator appears to be less sensitive to outliers thanks to the rank transformation, which could be one reason for the superior out-of-sample performance.

In summary, the transformation does not impose any additional assumptions, directly relates to portfolio sorting, and works well in finite samples because it appears more robust to outliers.¹⁴

D Adaptive Group LASSO

We use a group LASSO procedure suggested by Huang et al. (2010) for estimation and to select those characteristics that provide incremental information for expected returns, that is, for model selection. To recap, we are interested in modeling excess returns as a function of characteristics, that is,

$$R_{it} = \sum_{s=1}^S \tilde{m}_{ts}(\tilde{C}_{it-1}) + \varepsilon_{it}, \quad (3)$$

where $\tilde{m}_s(\cdot)$ are unknown functions and \tilde{C}_{it-1} denotes the rank-transformed characteristics.

The idea of the group LASSO is to estimate the functions \tilde{m}_{ts} nonparametrically, while setting functions for a given characteristic to 0 if the characteristic does not help predict expected returns. Therefore, the procedure achieves model selection; that is, it discriminates between the functions \tilde{m}_{ts} , which are constant, and the functions that are

¹⁴Cochrane (2011) stresses the sensitivity of regressions to outliers. Our transformation is insensitive to outliers and nicely addresses his concern.

not constant.¹⁵

Let \tilde{I}_l for $l = 1, \dots, L$ be a partition of the unit interval. To estimate \tilde{m}_t , we use *quadratic* splines, that is, we approximate \tilde{m}_t as a quadratic function on each interval \tilde{I}_l . We choose these functions so that the endpoints are connected and \tilde{m}_t is differentiable on $[0, 1]$. It turns out that we can approximate each \tilde{m}_{ts} by a series expansion with these properties, that is,

$$\tilde{m}_{ts}(\tilde{c}) \approx \sum_{k=1}^{L+2} \beta_{tsk} p_k(\tilde{c}), \quad (4)$$

where $p_k(c)$ are basis splines.¹⁶

The number of intervals L is a user-specified smoothing parameter, similar to the number of portfolios. As L increases, the precision of the approximation increases, but so does the number of parameters we have to estimate and hence the variance. Recall that portfolio sorts can be interpreted as approximating the conditional mean function as a constant function over L intervals. Our estimator is a smooth and more flexible estimator, but follows a similar idea (see again Figures 4 – 6).

We now discuss the two steps of the adaptive group LASSO. In the first step, we obtain estimates of the coefficients as

$$\tilde{\beta}_t = \arg \min_{b_{sk}: s=1, \dots, S; k=1, \dots, L+2} \sum_{i=1}^N \left(R_{it} - \sum_{s=1}^S \sum_{k=1}^{L+2} b_{sk} p_k(\tilde{C}_{s,it-1}) \right)^2 + \lambda_1 \sum_{s=1}^S \left(\sum_{k=1}^{L+2} b_{sk}^2 \right)^{\frac{1}{2}}, \quad (5)$$

where $\tilde{\beta}_t$ is an $(L+2) \times S$ vector of b_{sk} estimates and λ_1 is a penalty parameter.

The first part of equation (5) is just the sum of the squared residuals as in ordinary least squares regressions; the second part is the LASSO group penalty function. Rather than penalizing individual coefficients, b_{sk} , the LASSO penalizes all coefficients associated with a given characteristic. Thus, we can set the point estimates of an entire expansion of \tilde{m}_t to 0 when a given characteristic does not provide independent information for expected returns. In the application, we choose λ_1 in a data-dependent way to minimize a Bayes

¹⁵The “adaptive” part indicates a two-step procedure, because the LASSO selects too many characteristics in the first step and is therefore not model selection consistent unless restrictive conditions on the design matrix are satisfied (see Meinshausen and Bühlmann (2006) and Zou (2006) for an in-depth treatment of the LASSO in the linear model).

¹⁶See Chen (2007) for an overview of series estimation.

Information Criterion (BIC) proposed by Yuan and Lin (2006). Due to the penalty, the LASSO is applicable even when the number of characteristics is larger than the sample size.

However, as in a linear model, the first step of the LASSO selects too many characteristics. Informally speaking, the LASSO selects all characteristics that predict returns but also selects some characteristics that have no predictive power. A second step addresses this problem.

We first define the following weights:

$$w_s = \begin{cases} \left(\sum_{k=1}^{L+2} \tilde{\beta}_{sk}^2 \right)^{-\frac{1}{2}} & \text{if } \sum_{k=1}^{L+2} \tilde{\beta}_{sk}^2 \neq 0 \\ \infty & \text{if } \sum_{k=1}^{L+2} \tilde{\beta}_{sk}^2 = 0. \end{cases} \quad (6)$$

In the second step of the adaptive group LASSO, we solve

$$\tilde{\beta}_t = \arg \min_{b_{sk}: s=1, \dots, S; k=1, \dots, L+2} \sum_{i=1}^N \left(R_{it} - \sum_{s=1}^S \sum_{k=1}^{L+2} b_{sk} p_k(\tilde{C}_{s,it-1}) \right)^2 + \lambda_2 \sum_{j=1}^P \left(w_s \sum_{k=1}^{L+2} b_{sk}^2 \right)^{\frac{1}{2}}. \quad (7)$$

Huang et al. (2010) show the estimator from equation (7) is model selection consistent; that is, it correctly selects the non-constant functions with probability approaching 1 as the sample size grows large. We again choose λ_2 to minimize a BIC.¹⁷

Denote the estimated coefficients for a selected characteristic s by $\hat{\beta}_{ts}$. The estimator of the function \tilde{m}_{ts} is then

$$\hat{\tilde{m}}_{ts}(\tilde{c}) = \sum_{k=1}^{L+2} \hat{\beta}_{tsk} p_k(\tilde{c}).$$

If the cross section is sufficiently large, model selection and estimation can be performed period by period. Hence, the method allows for the importance of characteristics and the shape of the conditional mean function to vary over time. For example, some characteristics might lose their predictive power for expected returns.¹⁸ McLean and Pontiff (2016) show that for 97 return predictors, predictability decreases

¹⁷As a technical note, although the procedure is model-selection consistent, the resulting estimators have unfavorable statistical properties because they are not oracle efficient. Re-estimating the parameters using only the selected variables and no penalty function addresses this problem.

¹⁸The size effect is a recent example.

by 58% post publication. However, if the conditional mean function were time-invariant, pooling the data across time would lead to more precise estimates of the function and therefore more reliable predictions. In our empirical application in Section V, we estimate our model over subsamples and also estimate rolling specifications to investigate the variation in the conditional mean function over time.

E Confidence Bands

We also report uniform confidence bands for the functions \tilde{m}_{ts} . We approximate $\tilde{m}_{ts}(\tilde{c})$ by $\sum_{k=1}^{L+2} \beta_{tsk} p_k(\tilde{c})$ and estimate it by $\sum_{k=1}^{L+2} \hat{\beta}_{tsk} p_k(\tilde{c})$.

Let $p(\tilde{c}) = (p_1(\tilde{c}), \dots, p_{L+2}(\tilde{c}))'$ be the vector of spline functions and let Σ_{ts} be the $L+2 \times L+2$ covariance matrix of $\sqrt{n}(\hat{\beta}_{ts} - \beta_{ts})$. We define $\hat{\Sigma}_{ts}$ as the heteroscedasticity-consistent estimator of Σ_{ts} and define $\hat{\sigma}_{ts}(\tilde{c}) = \sqrt{p(\tilde{c})' \hat{\Sigma}_{ts} p(\tilde{c})}$.

The uniform confidence band for \tilde{m}_{ts} is of the form

$$\left[\sum_{k=1}^{L+2} \hat{\beta}_{tsk} p_k(\tilde{c}) - d_{ts} \hat{\sigma}_{ts}(\tilde{c}), \sum_{k=1}^{L+2} \hat{\beta}_{tsk} p_k(\tilde{c}) + d_{ts} \hat{\sigma}_{ts}(\tilde{c}) \right],$$

where d_{ts} is a constant.

To choose the constant, let $Z \sim N(0, \hat{\Sigma}_{ts})$ and let d_{ts} be such that

$$P \left(\sup_{\tilde{c} \in [0,1]} \left| \frac{Z' p(\tilde{c})}{\sqrt{p(\tilde{c})' \hat{\Sigma}_{ts} p(\tilde{c})}} \right| \leq d_{ts} \right) = 1 - \alpha.$$

We can calculate the probability on the left-hand side using simulations.

Given consistent model selection and under the conditions in Belloni, Chernozhukov, Chetverikov, and Kato (2015), it follows that

$$P \left(\tilde{m}_{ts}(\tilde{c}) \in \left[\sum_{k=1}^{L+2} \hat{\beta}_{tsk} p_k(\tilde{c}) - d_{ts} \hat{\sigma}_{ts}(\tilde{c}), \sum_{k=1}^{L+2} \hat{\beta}_{tsk} p_k(\tilde{c}) + d_{ts} \hat{\sigma}_{ts}(\tilde{c}) \right] \forall \tilde{c} \in [0,1] \right) \rightarrow 1 - \alpha$$

as the sample size increases.

F Interpretation of the Conditional Mean Function

In a nonparametric additive model, the locations of the functions are not identified. Consider the following example. Let α_s be constants such that

$$\sum_{s=1}^S \alpha_s = 0.$$

Then,

$$\tilde{m}_t(\tilde{c}_1, \dots, \tilde{c}_S) = \sum_{s=1}^S \tilde{m}_{ts}(\tilde{c}_s) = \sum_{s=1}^S (\tilde{m}_{ts}(\tilde{c}_s) + \alpha_s).$$

Therefore, the summands of the transformed conditional mean function, \tilde{m}_s , are only identified up to a constant. The model selection procedure, expected returns, and the portfolios we construct do not depend on these constants. However, the constants matter when we plot an estimate of the conditional mean function for one characteristic, \tilde{m}_s . We now discuss two possible normalizations.

Let \bar{c}_s be a fixed value of a given transformed characteristic s , such as the mean or the median. Then,

$$\tilde{m}_t(\tilde{c}_1, \bar{c}_2, \dots, \bar{c}_S) = \tilde{m}_{t1}(\tilde{c}_1) + \sum_{s=2}^S \tilde{m}_{ts}(\bar{c}_s),$$

which is identified and a function of \tilde{c}_1 only. This function is the expected return as a function of the first characteristic when we fix the values of all other characteristics. When we set the other characteristics to different values, we change the level of the function, but not the slope. We will report these functions in our empirical section, and we can interpret both the level and the slope of the function.

An alternative normalization is $\tilde{m}_1(0.5) = 0$, that, the conditional mean function for a characteristic takes the value of 0 for the median observation. Now, we cannot interpret the level of the function. This normalization, however, is easier to interpret when we plot the estimated functions over time in a three-dimensional surface plot. Changes in the slope over time now tell us the relative importance of the characteristic in the time series. The first normalization has the disadvantage that in years with very low overall returns, the conditional mean function is much lower. Hence, interpreting the relative importance

of a characteristic over time from surface plots is more complicated when we use the first normalization.

IV Data

Stock return data come from the Center for Research in Security Prices (CRSP) monthly stock file. We follow standard conventions and restrict the analysis to common stocks of firms incorporated in the United States trading on NYSE, Amex, or Nasdaq. Market equity (ME) is the total market capitalization at the firm level. LME is the total market capitalization at the end of the previous calendar month. $LTurnover$ is the ratio of total monthly trading volume over total market capitalization at the end of the previous month. The bid-ask spread ($spread_mean$) is the average daily bid-ask spread during the previous month. We also construct lagged returns over the previous month ($cum_return_1_0$), the previous 12 months leaving out the last month ($cum_return_12_2$), intermediate momentum ($cum_return_12_7$), and long-run returns from three years ago until last year ($cum_return_36_13$). We follow Frazzini and Pedersen (2014) in the definition of Beta ($beta$), and idiosyncratic volatility ($idio_vol$) is the residual from a regression of daily returns on the three Fama and French factors in the previous month as in Ang, Hodrick, Xing, and Zhang (2006).

Balance-sheet data are from the Standard and Poor's Compustat database. We define book equity (BE) as total stockholders' equity plus deferred taxes and investment tax credit (if available) minus the book value of preferred stock. Based on availability, we use the redemption value, liquidation value, or par value (in that order) for the book value of preferred stock. We prefer the shareholders' equity number as reported by Compustat. If these data are not available, we calculate shareholders' equity as the sum of common and preferred equity. If neither of the two are available, we define shareholders' equity as the difference between total assets and total liabilities. The book-to-market (BM) ratio of year t is then the book equity for the fiscal year ending in calendar year $t - 1$ over the market equity as of December $t - 1$. We use the book-to-market ratio then for estimation starting in June of year t until June of year $t + 1$. We use the same timing convention

unless we specify it differently.

AT are total assets, and cash (C) is cash and short-term investments over total assets. DP is depreciation and amortization over total assets. We define expenses to sales ($FC2Y$) as the sum of advertising expenses, research and development expenses, and selling, general, and administrative expenses over sales and investment expenditure ($I2Y$) as capital expenditure over sales. Operating leverage (OL) is the ratio of cost of goods sold and selling, general and administrative expenses over total assets. We define the price-to-cost margin (pcm) as sales minus cost of goods sold over sales, and gross profitability ($Prof$) as gross profits over book value of equity. The return-on-equity (ROE) is the ratio of income before extraordinary items to lagged book value of equity. Investment growth ($Investment$) is the annual growth rate in total assets. We define operating accruals (OA) as in Sloan (1996). Free cash flow ($free_cf$) is the ratio of net income and depreciation and amortization minus the change in working capital and capital expenditure over the book value of equity. We define Q (q) as total assets plus total market capitalization minus common equity and deferred taxes over total assets, and the HHI as the Herfindahl-Hirschman index of annual sales at the Fama-French 48-industry level.

We define the net payout ratio (PR) as net payout over net income. Net payout is the sum of ordinary dividends and net purchases of common and preferred stock. Return on equity (ROE) is the ratio of income before extraordinary items over lagged book equity. Sales growth ($Sales_g$) is the percentage growth rate in net sales.

To alleviate a potential survivorship bias due to backfilling, we require that a firm have at least two years of Compustat data. Our sample period is July 1963 until June 2015. Table 1 reports summary statistics for various firm characteristics and return predictors. We calculate all statistics annually and then average over time.

The online appendix contains a detailed description of the characteristics, the construction, and the relevant references.

V Results

We now study which of the 24 characteristics we describe in section IV provide independent information for expected returns, using the adaptive group LASSO for selection and estimation.

A Selected Characteristics and Their Influence

Table 2 reports average monthly returns and standard deviations of 10 portfolios sorted on the characteristics we study. Most of the 24 characteristics have individually predictive power for expected returns and result in large and statistically significant hedge portfolio returns and alphas relative to the Fama and French three-factor model (results in the online appendix). The vast majority of economic models, that is, the ICAPM (Merton (1973)) or consumption models, as surveyed in Cochrane (2007), suggest a low number of state variables can explain the cross section of returns. Therefore, all characteristics are unlikely to provide independent information for expected returns. To tackle the multi-dimensionality challenge, we now estimate the adaptive group LASSO with five and 10 knots.¹⁹

Figure 7 shows the conditional mean function, $\tilde{m}(\tilde{C}_{it-1})$, for Tobin's Q. Stocks with low Q have expected returns of around 1% per month. Returns monotonically decrease with increasing Q to a negative return of 1% per month for the firms with the highest Q. This result is consistent with our findings for portfolio sorts in Table 2. Portfolio sorts result in an average annualized hedge portfolio return of more than 14%. Tobin's Q is, however, correlated with the book-to-market ratio and other firm characteristics. We now want to understand whether Q has marginal predictive power for expected returns conditional on all other firm characteristics we study. Figure 8 plots $m(\tilde{C}_{it-1})$ for Tobin's Q conditional on all other characteristics. The conditional mean function is now constant and does not vary with Q. The constant conditional mean function implies Q has no marginal predictive power for expected returns once we condition on other firm characteristics.

¹⁹The number of knots corresponds to the smoothing parameter we discuss in section III.

The example of Tobin’s Q shows the importance to condition on other characteristics to infer on the predictive power of characteristics for expected returns. We now study this question for 24 firm characteristics using the adaptive group LASSO.

Table 4 reports the selected characteristics of the nonparametric model for different numbers of knots, sets of firms, and sample periods. We see in column (1) that the baseline estimation for all stocks over the full sample period using five knots selects 11 out of the universe of 24 firm characteristics. The lagged market cap, turnover, the book-to-market ratio, the ratio of depreciation to total assets, profitability, investment, the Herfindahl-Hirschman index, short-term reversal, intermediate momentum, momentum, and idiosyncratic volatility all provide incremental information conditional on all other selected firm characteristics.

When we allow for a finer grid in column (2), only eight characteristics provide independent incremental information for expected returns. The book-to-market ratio, the ratio of depreciation to total assets, and profitability all lose their predictive power for expected returns. The penalty function increases in the number of knots. In the nonparametric model with 10 knots, the penalty is proportional to 12 times the number of selected characteristics, which is why we select fewer characteristics with more knots.

We estimate the nonparametric model only on large stocks above the 20% size quintile of NYSE stocks in column (3). Now the book-to-market ratio again provides independent information for expected returns. Size, turnover, and the Herfindahl, instead, lose incremental predictive power for expected returns once we condition on all other firm characteristics.

Columns (4) and (5) split the sample in half and re-estimate our benchmark nonparametric model in both sub-samples separately to see whether the importance of characteristics for predicted returns varies over time. Size, turnover, the book-to-market ratio, short-term turnover, intermediate momentum, momentum, beta, and idiosyncratic volatility are significant predictors in the first half of the sample. In the second half of the sample, the book-to-market ratio, momentum, beta, and idiosyncratic volatility lose their incremental information for expected returns. The ratio of depreciation to total assets, profitability, investment, and the Herfindahl-Hirschman index, instead, gain predictive

power for expected returns.

B Time Variation in Return Predictors

Figures 9–19 show the conditional mean function for our baseline nonparametric model for all stocks and five knots over time. We estimate the model on a rolling basis over 120 months. We normalize the conditional mean function to equal 0 when the normalized characteristic equals 0.5 (the median characteristic in a given month).

We see in Figure 9 that the conditional mean function is non-constant throughout the sample period for lagged market cap. Small firms have higher expected returns compared to large firms, conditional on all other significant return predictors. Interestingly, the size effect seems largest during the end of our sample period, contrary to conventional wisdom (see Asness, Frazzini, Israel, Moskowitz, and Pedersen (2015) for a related finding). Figure 18 shows the momentum crash during the recent financial crisis. Momentum crashed due to high returns of past losers, consistent with findings in Daniel and Moskowitz (2016).

C Out-of-Sample Performance and Model Comparison

We now want to compare the performance of the nonparametric model with the linear model out of sample. We estimate the nonparametric model for a period from 1963 to 1990 and carry out model selection with the adaptive group LASSO with 10 knots, but also use the adaptive LASSO for model selection in the linear model. We select the following characteristic in the nonparametric model: lagged market cap, lagged turnover, the book-to-market ratio, investment, free cash flow, the Herfindahl-Hirschman index, momentum, intermediate momentum, short-term reversal, beta and idiosyncratic volatility.

The linear model does not select the lagged market cap and free cash flow, but instead selects the additional following seven characteristics: cash, depreciation and amortization, operating leverage, return-on-equity, Tobin’s Q, the bid-ask spread, and long-term reversal. The linear model selects in total five more characteristics than the nonparametric model.²⁰

²⁰The linear model might be misspecified and therefore select more variables (see discussion in section III).

We then create rolling monthly out-of-sample predictions for excess return using 10 years of data for estimation and form two portfolios for each method. We buy the stocks with the 10% highest expected returns and sell the stocks with the 10% lowest predicted returns. The hedge portfolio of the linear model has an out-of-sample annualized Sharpe ratio of 0.97. The out-of-sample Sharpe ratio increases by more than 70% to 1.72 for the nonparametric model.

VI Conclusion

We propose a nonparametric methodology to tackle the challenge posed by Cochrane (2011) in his presidential address: which firm characteristics provide independent information for expected returns. We use the adaptive group LASSO to select significant return predictors and to estimate the model.

We document the properties of our framework in three applications: i) Which characteristics have incremental forecasting power for expected returns? ii) Does the predictive power of characteristics vary over time? iii) How does the nonparametric model compare to a linear model out of sample?

Our results are as follows: i) Out of 24 characteristics only 6 to 11 provide independent information depending on the number of interpolation points (similar to the number of portfolios in portfolio sorts), sample period, and universe of stocks (large versus small stocks). ii) Substantial time variation is present in the predictive power of characteristics. iii) The nonparametric model selects fewer characteristics than the linear model and has a 70% higher Sharpe ratio out of sample.

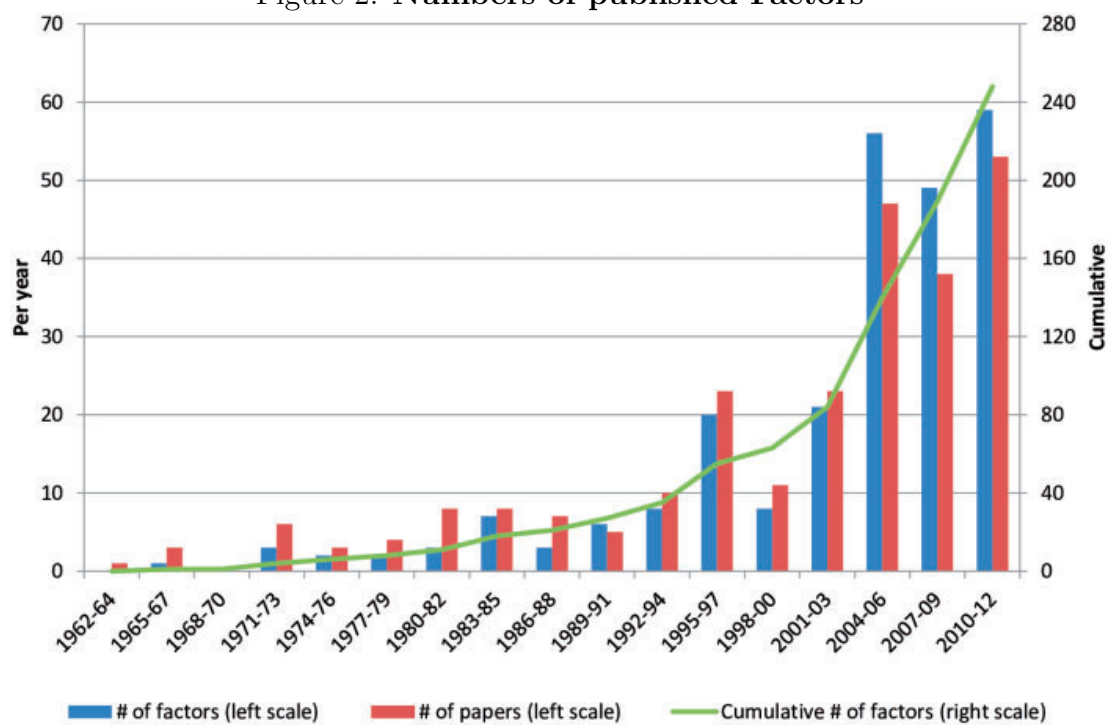
We see our paper as a starting point only and ask the following questions. Are the characteristics we identify related to factor exposures? How many factors are important? Can we achieve a dimension reduction and identify K factors that can summarize the N independent dimension of expected returns with $K < N$ similar to Fama and French (1992) and Fama and French (1993)?

References

- Ang, A., R. J. Hodrick, Y. Xing, and X. Zhang (2006). The cross-section of volatility and expected returns. *The Journal of Finance* 61(1), 259–299.
- Asness, C. S., A. Frazzini, R. Israel, T. J. Moskowitz, and L. H. Pedersen (2015). Size matters, if you control your junk. *Unpublished Manuscript, University of Chicago*.
- Banz, R. W. (1981). The relationship between return and market value of common stocks. *Journal of Financial Economics* 9(1), 3–18.
- Basu, S. (1983). The relationship between earnings’ yield, market value and return for nyse common stocks: Further evidence. *Journal of Financial Economics* 12(1), 129–156.
- Belloni, A., V. Chernozhukov, D. Chetverikov, and K. Kato (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics* 186(2), 345–366.
- Bhandari, L. C. (1988). Debt/equity ratio and expected common stock returns: Empirical evidence. *The Journal of Finance* 43(2), 507–528.
- Breeden, D. T. (1979). An intertemporal asset pricing model with stochastic consumption and investment opportunities. *Journal of Financial Economics* 7(3), 265–296.
- Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6 of *Handbook of Econometrics*, Chapter 76, pp. 5550–5623. Elsevier.
- Cochrane, J. H. (2007). Financial markets and the real economy. In R. Mehra (Ed.), *Handbook of the Equity Risk Premium*. Elsevier.
- Cochrane, J. H. (2011). Presidential address: Discount rates. *Journal of Finance* 66(4), 1047–1108.
- Daniel, K. and S. Titman (1997). Evidence on the Characteristics of Cross Sectional Variation in Stock Returns. *Journal of Finance* 52, 1–33.
- Daniel, K. D. and T. J. Moskowitz (2016). Momentum crashes. *Journal of Financial Economics* (forthcoming).
- Fama, E. F. and K. R. French (1992). The cross-section of expected stock returns. *Journal of Finance* 47(2), 427–465.
- Fama, E. F. and K. R. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33(1), 3–56.
- Fama, E. F. and K. R. French (1996). Multifactor explanations of asset pricing anomalies. *Journal of Finance* 51(1), 55–84.
- Fama, E. F. and K. R. French (2008). Dissecting anomalies. *Journal of Finance* 63(4), 1653–1678.
- Fama, E. F. and K. R. French (2015). A five-factor asset pricing model. *Journal of Financial Economics* 116(1), 1–22.
- Fama, E. F. and J. D. MacBeth (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy* 81(3), 607–636.
- Frazzini, A. and L. H. Pedersen (2014). Betting against beta. *Journal of Financial Economics* 111(1), 1–25.
- Harvey, C. R., Y. Liu, and H. Zhu (2016). ... and the cross-section of expected returns.

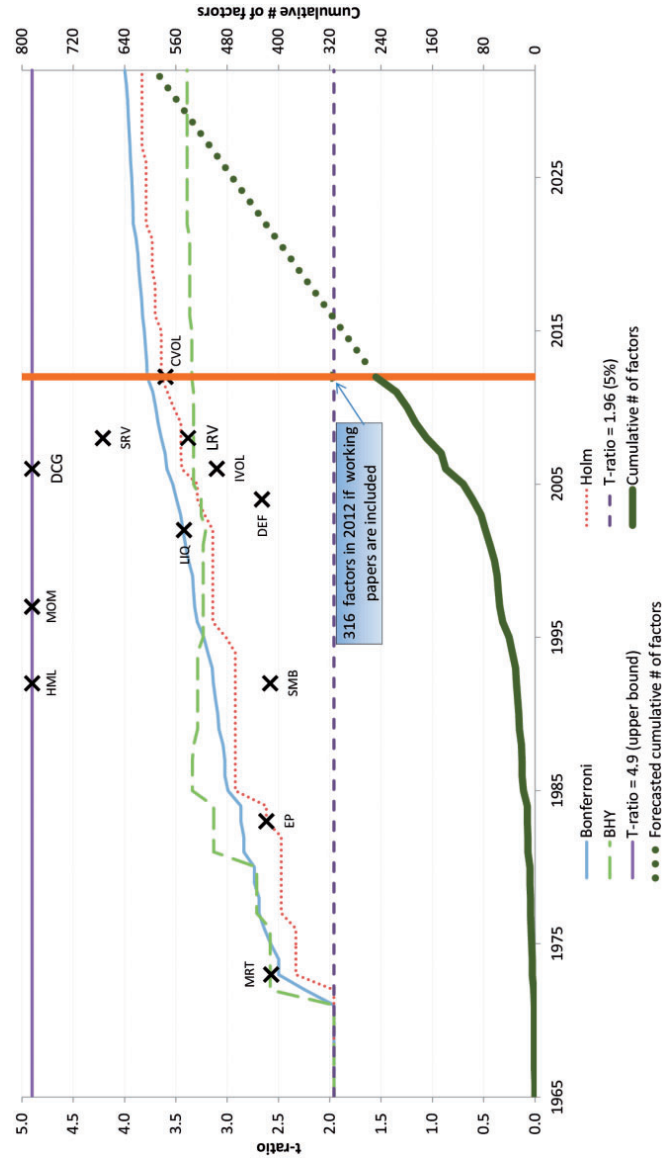
- Review of Financial Studies* 29(1), 5–68.
- Haugen, R. A. and N. L. Baker (1996). Commonality in determinants of expected stock returns. *Journal of Financial Economics* 41, 401–439.
- Horowitz, J., J. Klemelä, and E. Mammen (2006). Optimal estimation in additive regression models. *Bernoulli* 12(2), 271–298.
- Huang, J., J. L. Horowitz, and F. Wei (2010). Variable selection in nonparametric additive models. *Annals of Statistics* 38(4), 2282–2313.
- Jegadeesh, N. and S. Titman (1993). Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency. *Journal of Finance* 48, 65–91.
- Lewellen, J. (2015). The cross section of expected stock returns. *Critical Finance Review* 4(1), 1–44.
- 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.
- Lucas Jr, R. E. (1978). Asset prices in an exchange economy. *Econometrica* 46(6), 1429–1445.
- McLean, D. R. and J. Pontiff (2016). Does academic research destroy return predictability. *Journal of Finance* 71(1), 5–32.
- Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics* 34(3), 1436–1462.
- Merton, R. C. (1973). An intertemporal capital asset pricing model. *Econometrica* 41(5), 867–887.
- Mossin, J. (1966). Equilibrium in a capital asset market. *Econometrica* 34(4), 768–783.
- Rosenberg, B., K. Reid, and R. Lanstein (1985). Persuasive evidence of market inefficiency. *The Journal of Portfolio Management* 11(3), 9–16.
- Rubinstein, M. (1976). The valuation of uncertain income streams and the pricing of options. *The Bell Journal of Economics* 7(2), 407–425.
- Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance* 19(3), 425–442.
- Sloan, R. (1996). Do stock prices fully reflect information in accruals and cash flows about future earnings? *Accounting Review* 71(3), 289–315.
- Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics* 10(4), 1040–1053.
- Stone, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics* 13(2), 689–705.
- Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *The Annals of Statistics* 14(2), 590–606.
- Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68(1), 49–67.
- Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* 101(476), 1418–1429.

Figure 2: Numbers of published Factors



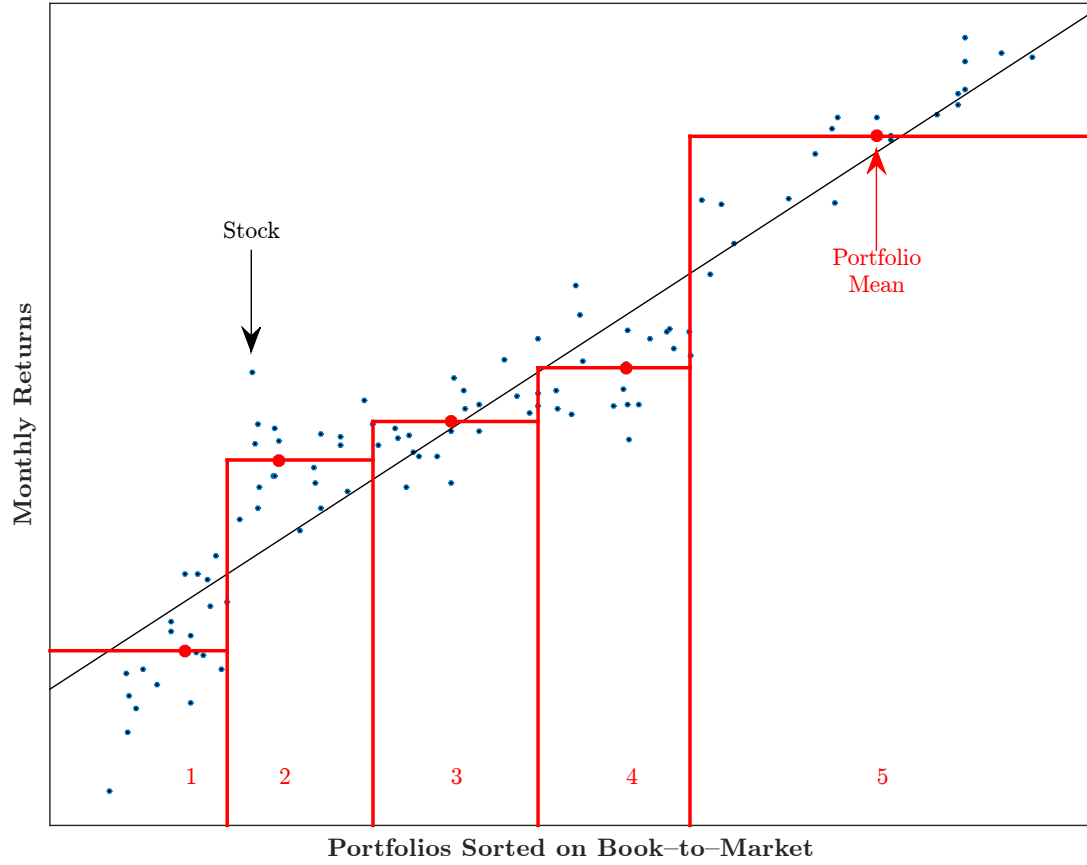
Number of discovered factors over time. Source: Figure 2 of Harvey, Liu, and Zhu (2016).

Figure 3: Suggested t-stats adjustment



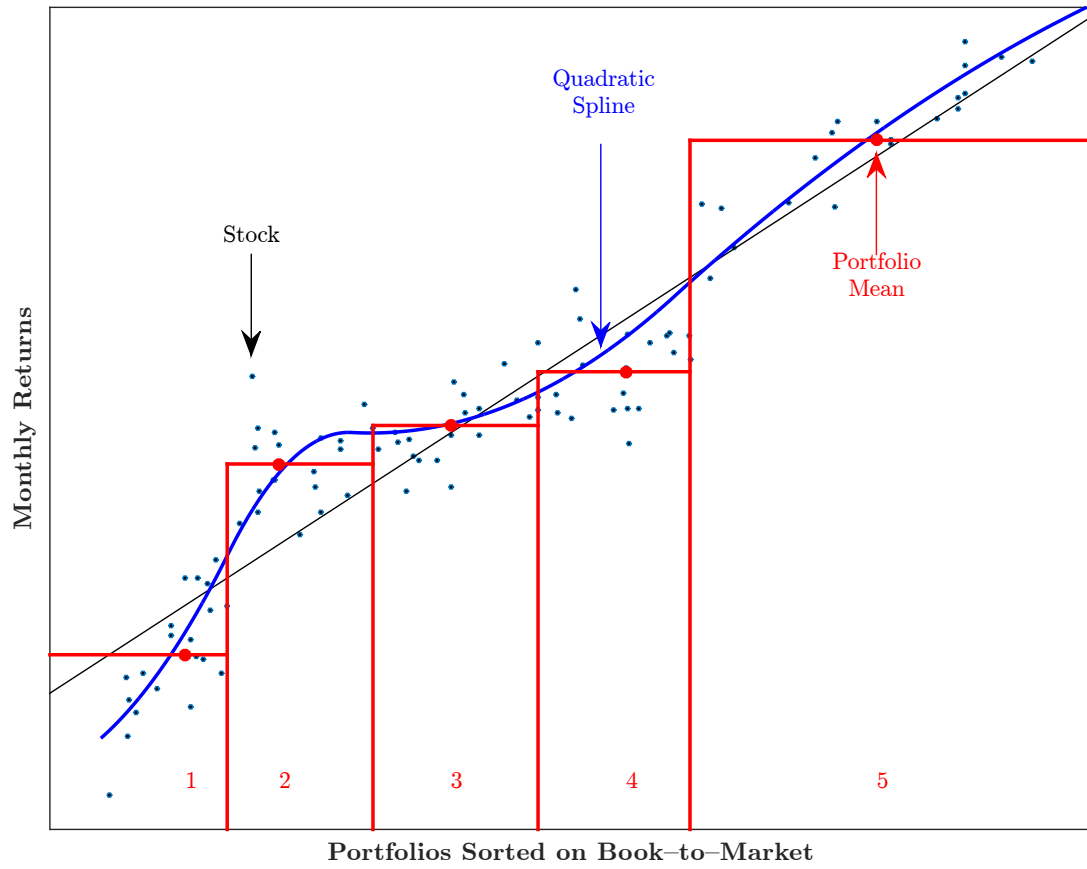
Recommendation to adjust t -statistics for multiple testing problem. Source: Figure 3 of Harvey, Liu, and Zhu (2016).

Figure 4: **5 Portfolios sorted on Book-to-Market**



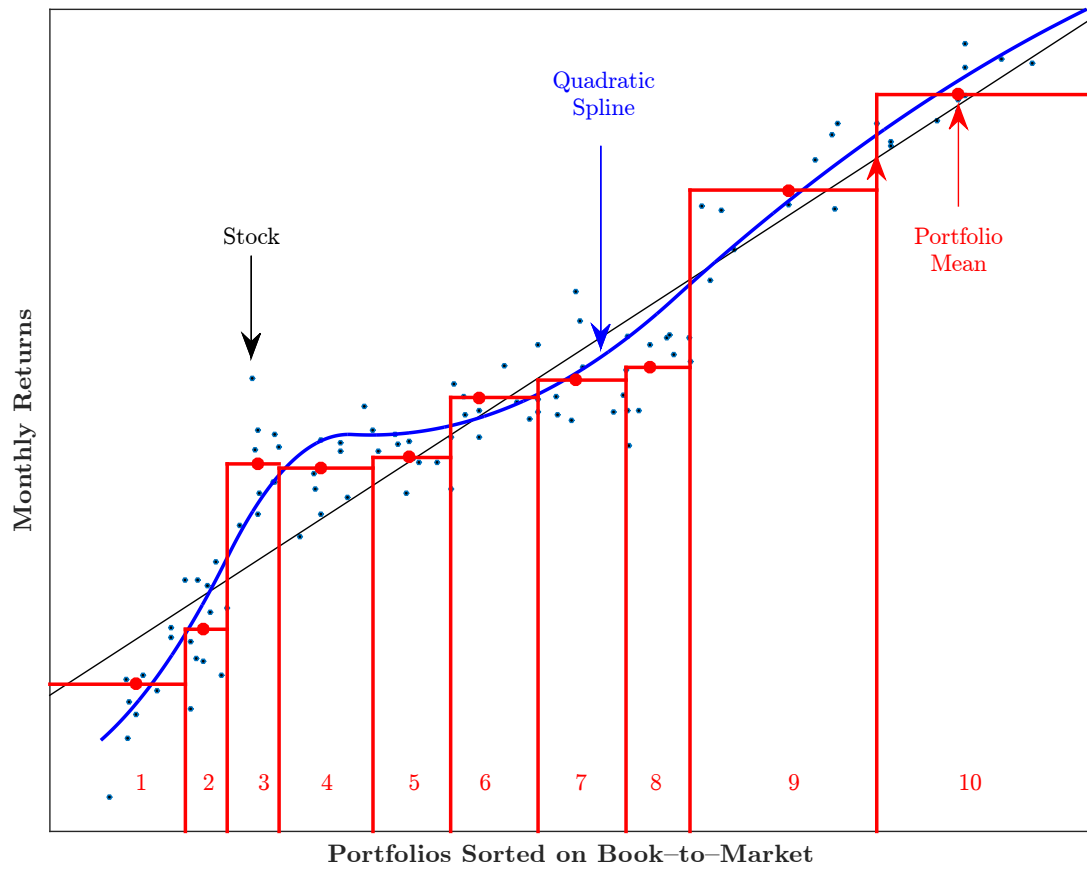
This figure plots expected returns on the y-axis against the book-to-market ratio on the x-axis as well as portfolio mean returns for simulated data.

Figure 5: **5 Portfolios** sorted on Book-to-Market and non-parametric Estimator



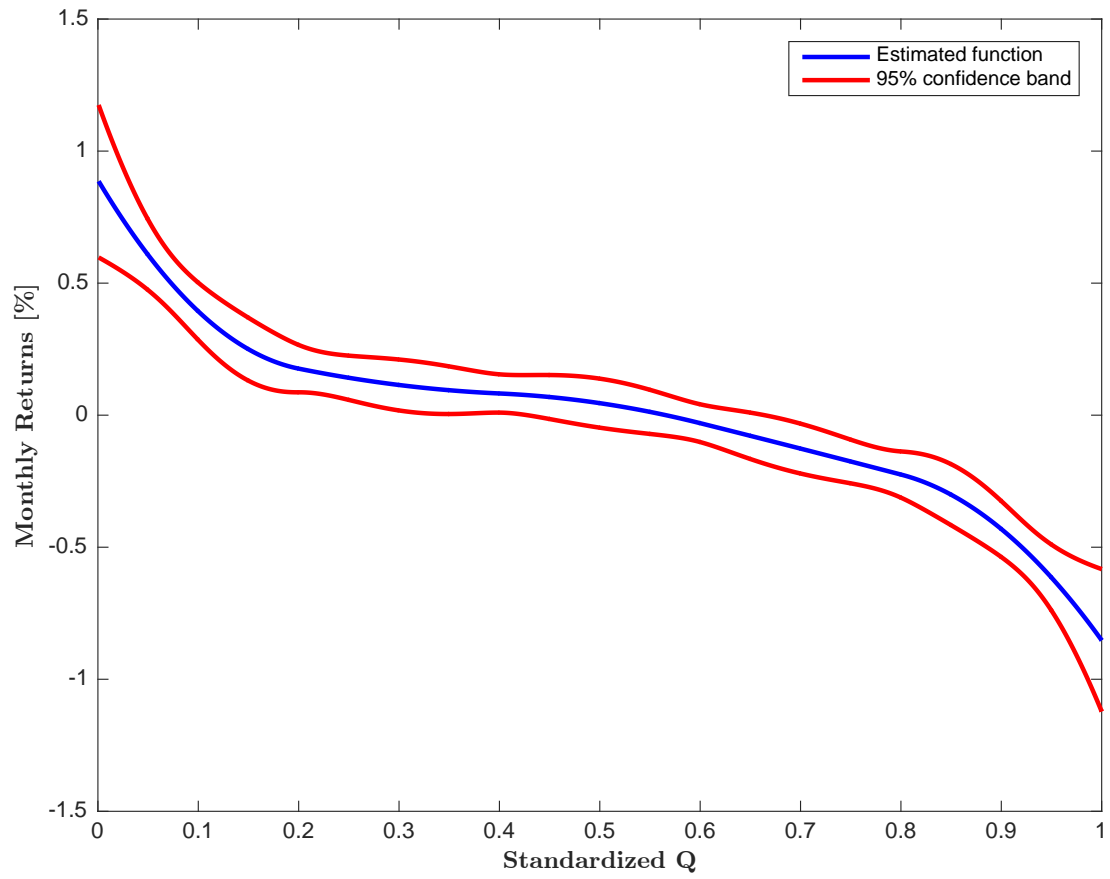
This figure plots expected returns on the y-axis against the book-to-market ratio on the x-axis as well as portfolio mean returns and a non-parametric conditional mean function for simulated data.

Figure 6: 10 Portfolios sorted on Book-to-Market and nonparametric Estimator



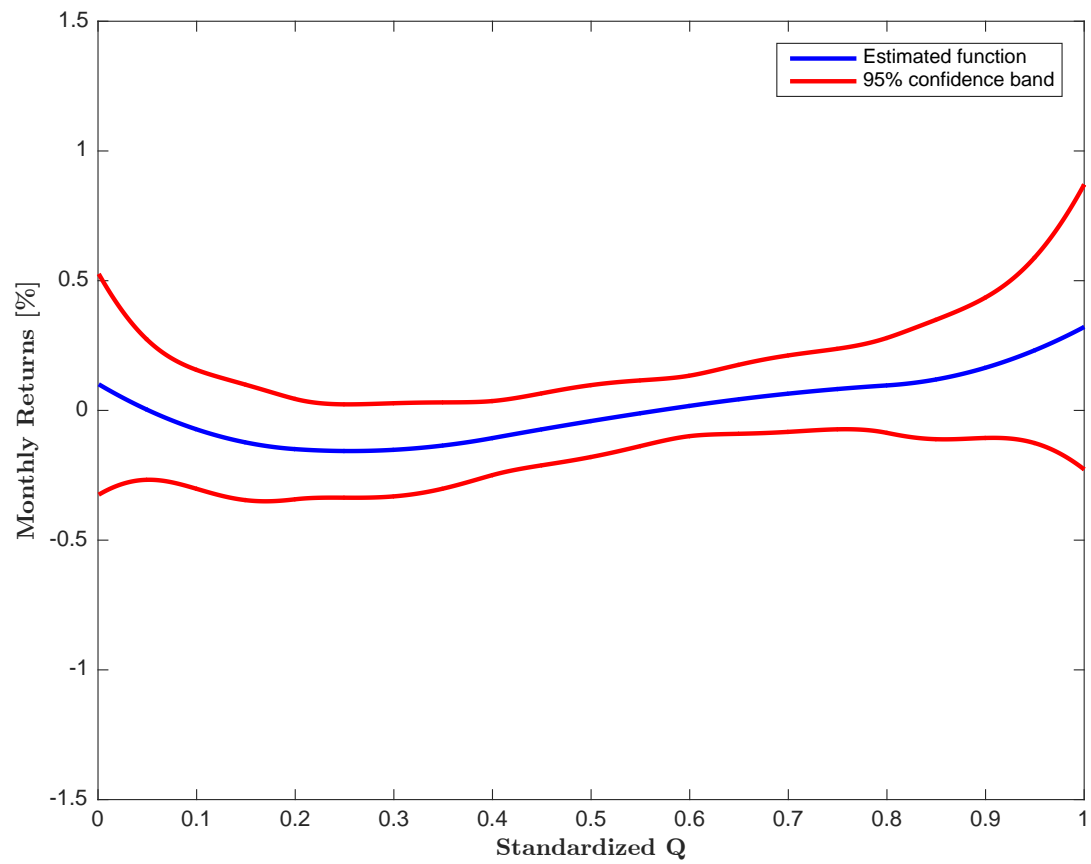
This figure plots expected returns on the y-axis against the book-to-market ratio on the x-axis as well as portfolio mean returns and a non-parametric conditional mean function for simulated data.

Figure 7: **Conditional Mean Function: Q (unconditional)**



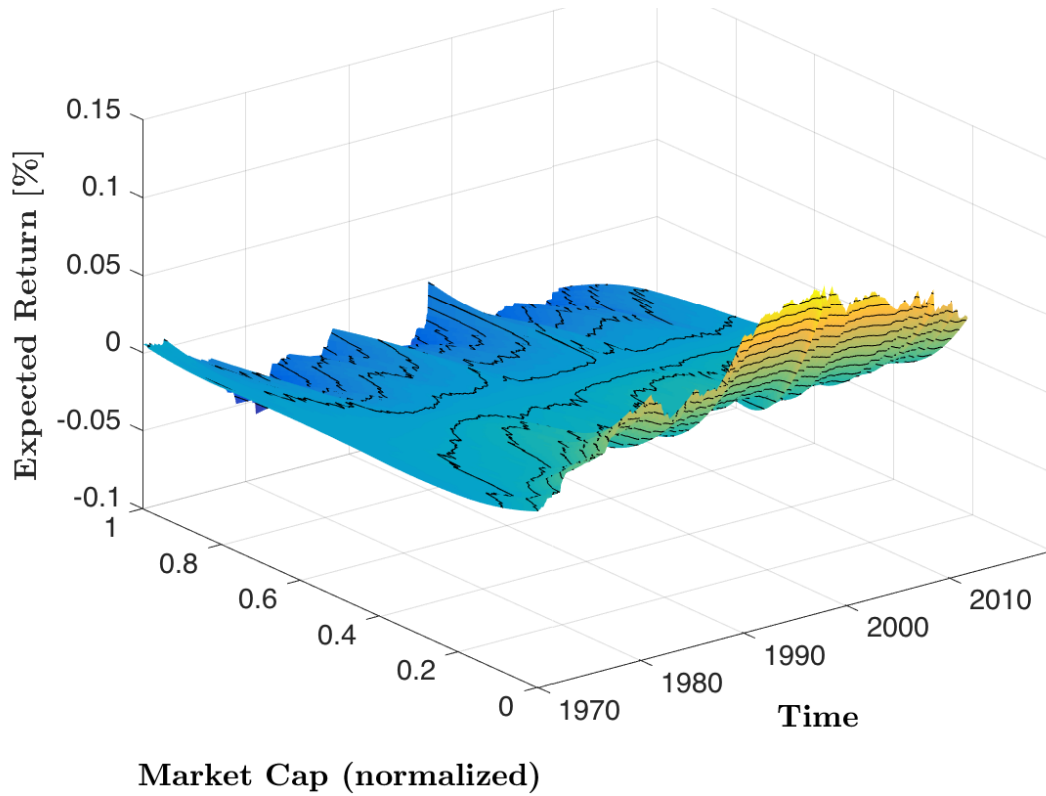
Effect of Tobin's Q on average returns.

Figure 8: **Conditional Mean Function: Q** (conditional on other characteristics)



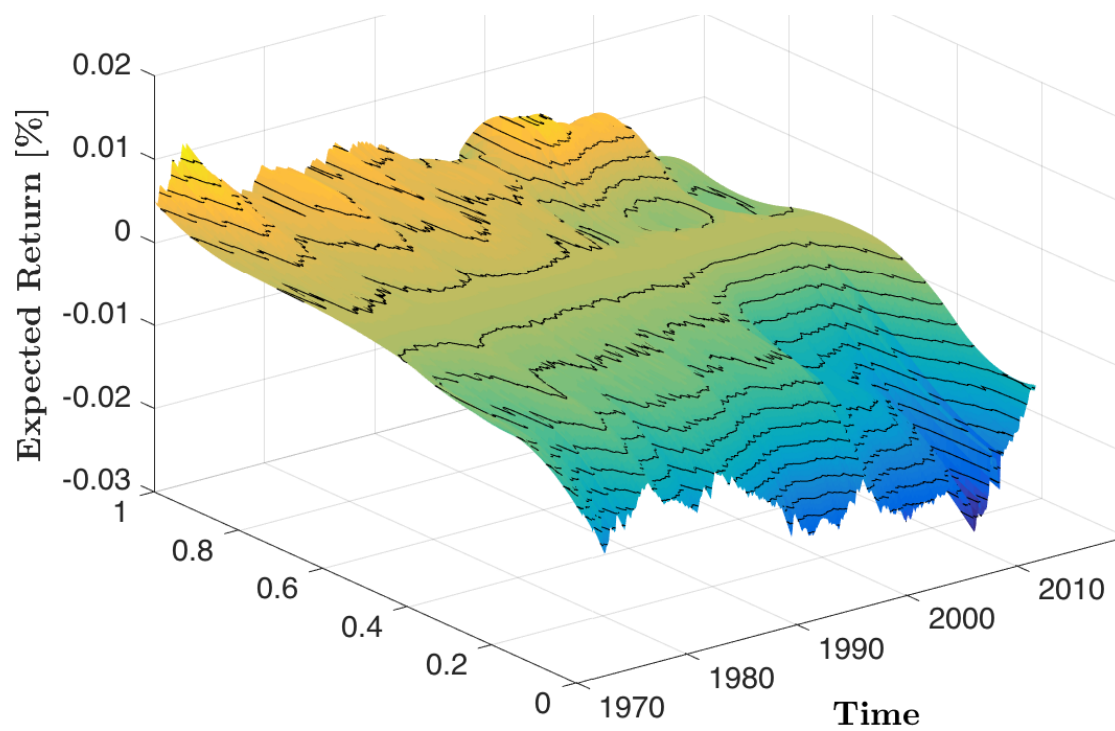
Effect of Tobin's Q on average returns (considering all the other selected characteristics).

Figure 9: **Time-varying Conditional Mean Function: Market Cap**



Effect of lagged market cap on average returns (conditional all other selected characteristics).

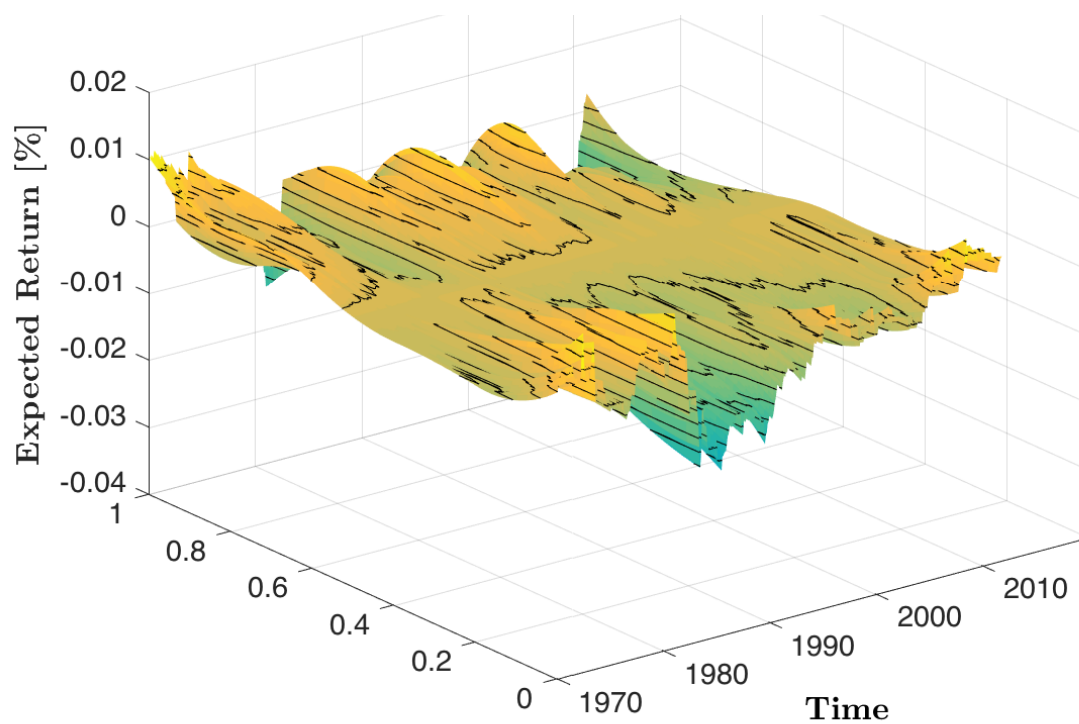
Figure 10: **Time-varying Conditional Mean Function: Turnover**



Turnover (normalized)

Effect of lagged turnover on average returns (conditional all other selected characteristics).

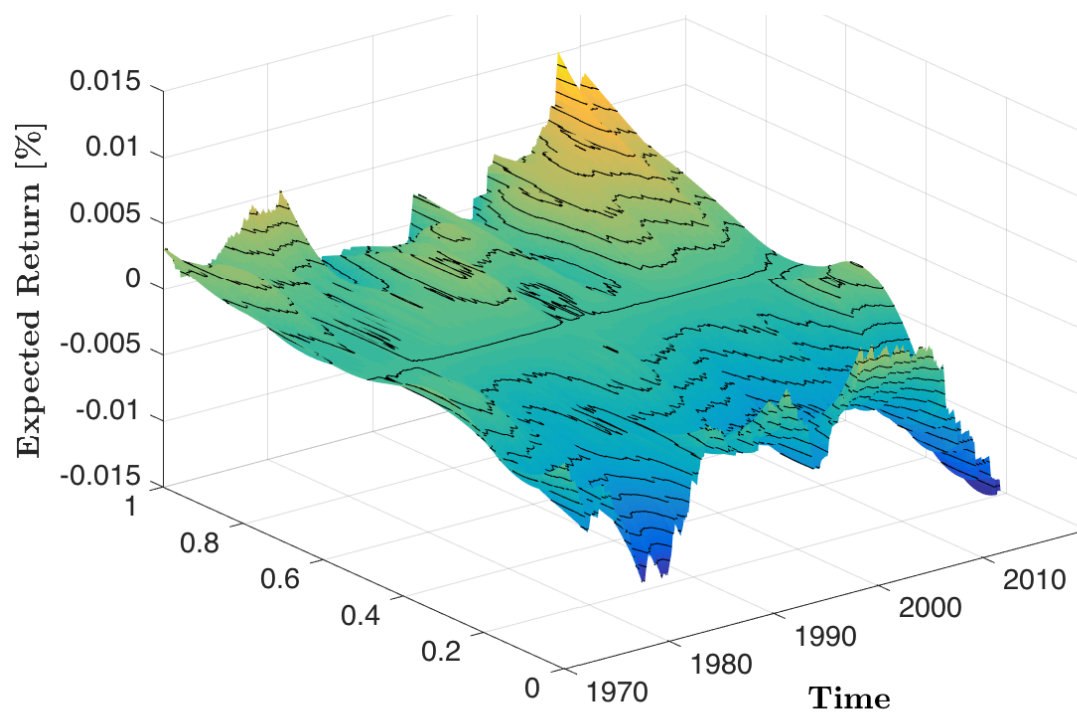
Figure 11: **Time-varying Conditional Mean Function: Book-to-Market Ratio**



Book-to-Market Ratio (normalized)

Effect of book-to-market ratio on average returns (conditional on all other selected characteristics).

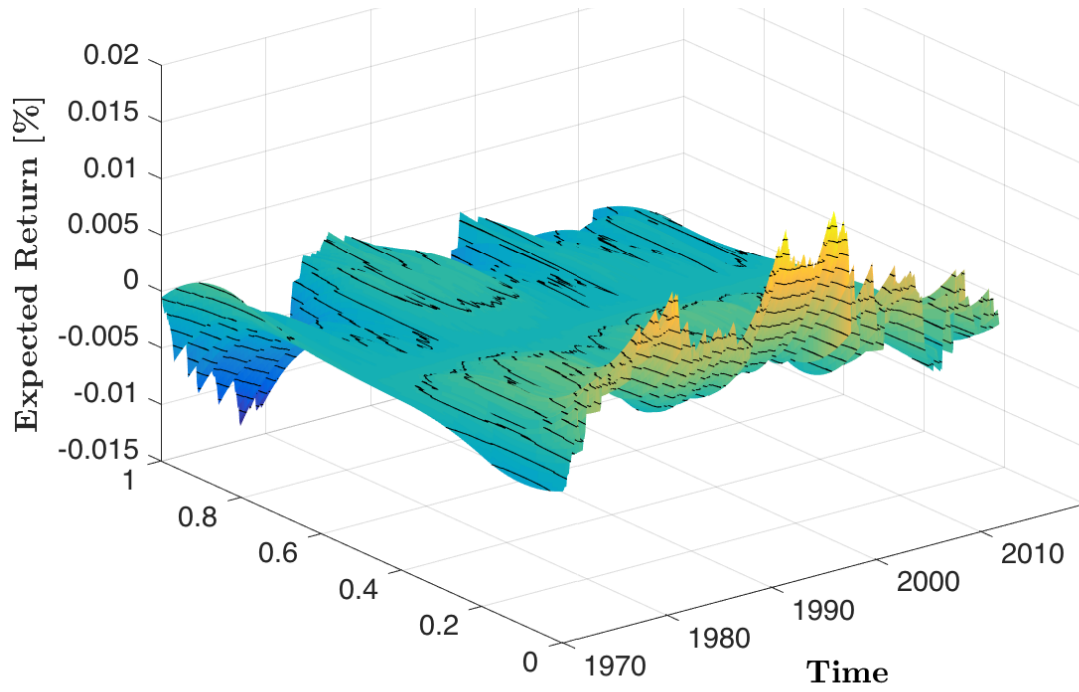
Figure 12: **Time-varying Conditional Mean Function: Depreciation-to-Assets**



Depreciation-to-Assets (normalized)

Effect of depreciation-to-assets on average returns (conditional on all other selected characteristics).

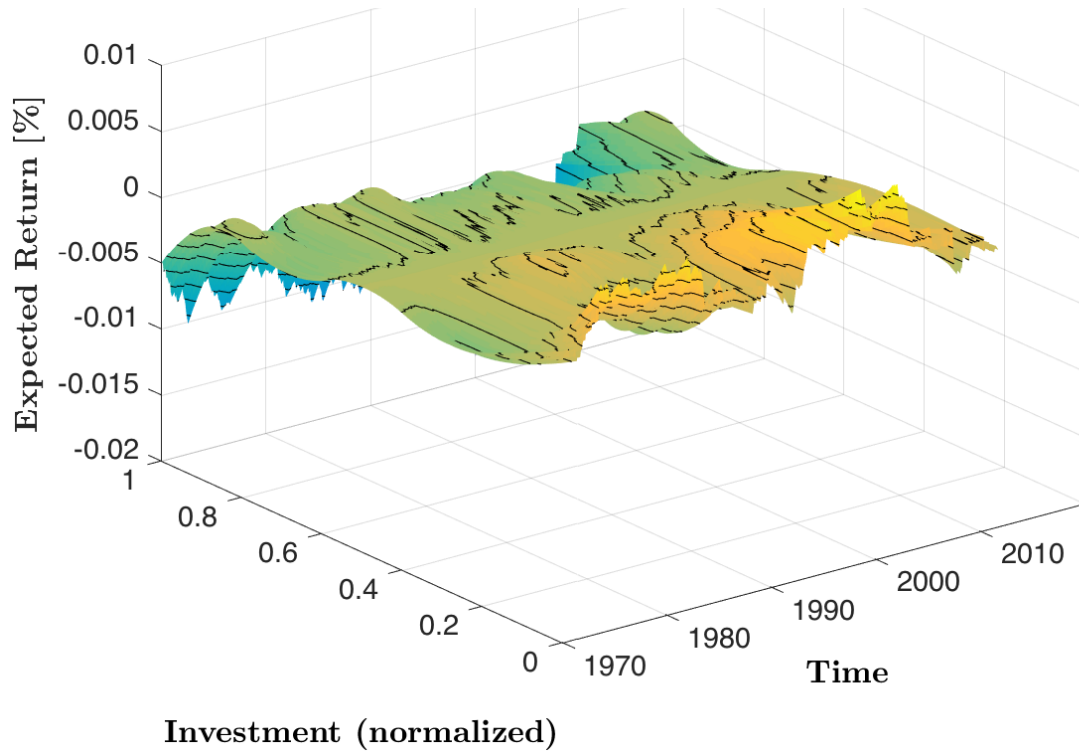
Figure 13: **Time-varying Conditional Mean Function: Profitability**



Profitability (normalized)

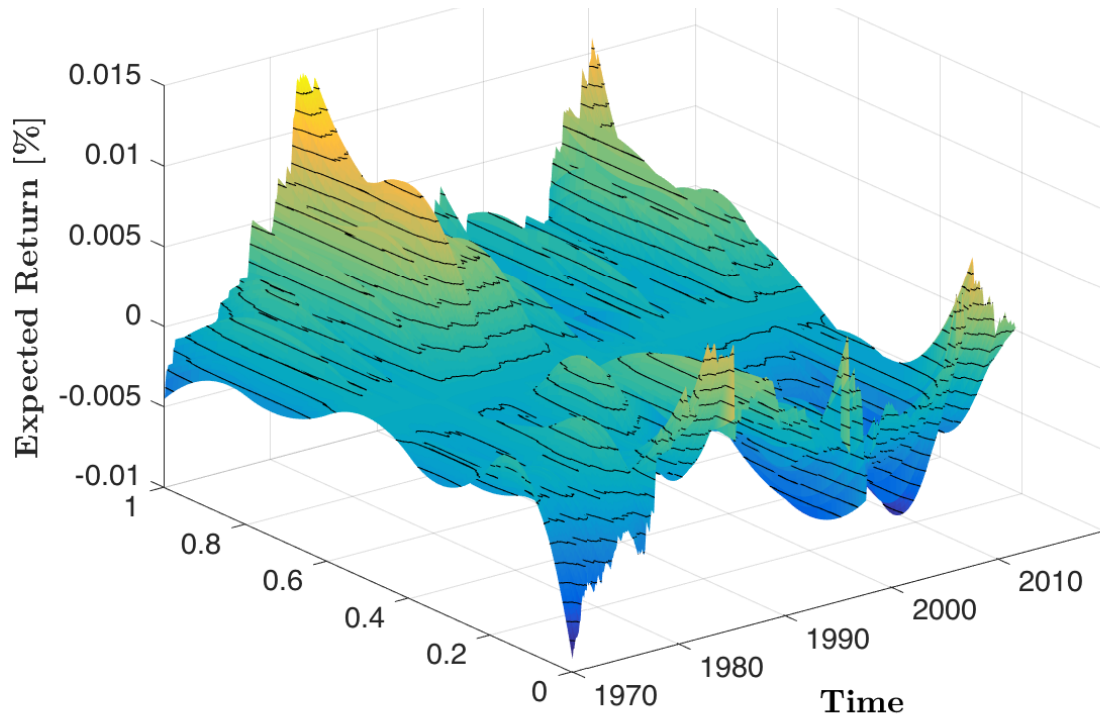
Effect of profitability on average returns (conditional on all other selected characteristics).

Figure 14: **Time-varying Conditional Mean Function: Investment**



Effect of investment on average returns (conditional on all other selected characteristics).

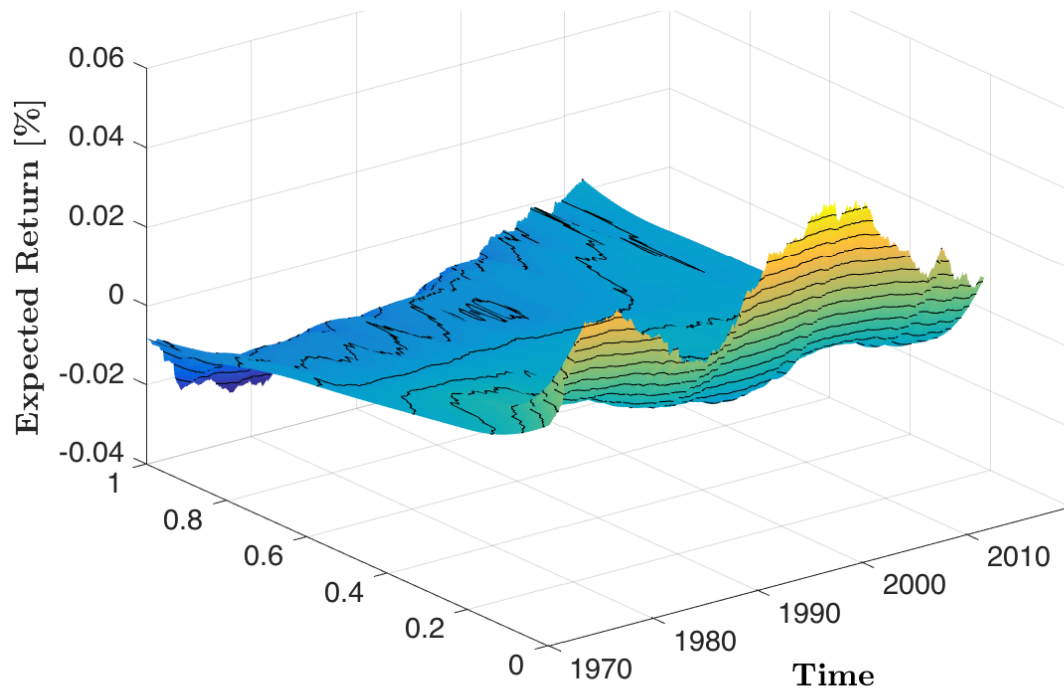
Figure 15: **Time-varying Conditional Mean Function: HHI**



HHI (normalized)

Effect of Herfindahl-Hirschman index on average returns (conditional on all other selected characteristics).

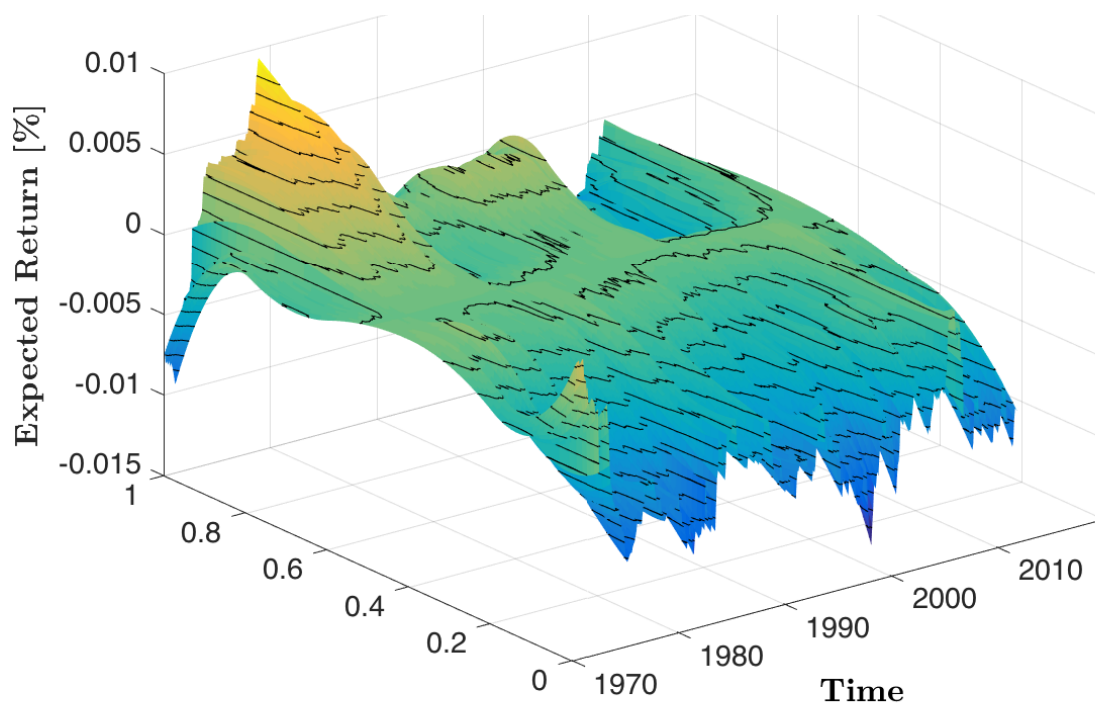
Figure 16: **Time-varying Conditional Mean Function: Short-Term Reversal**



One month lagged return (normalized)

Effect of short-term reversal on average returns (conditional on all other selected characteristics).

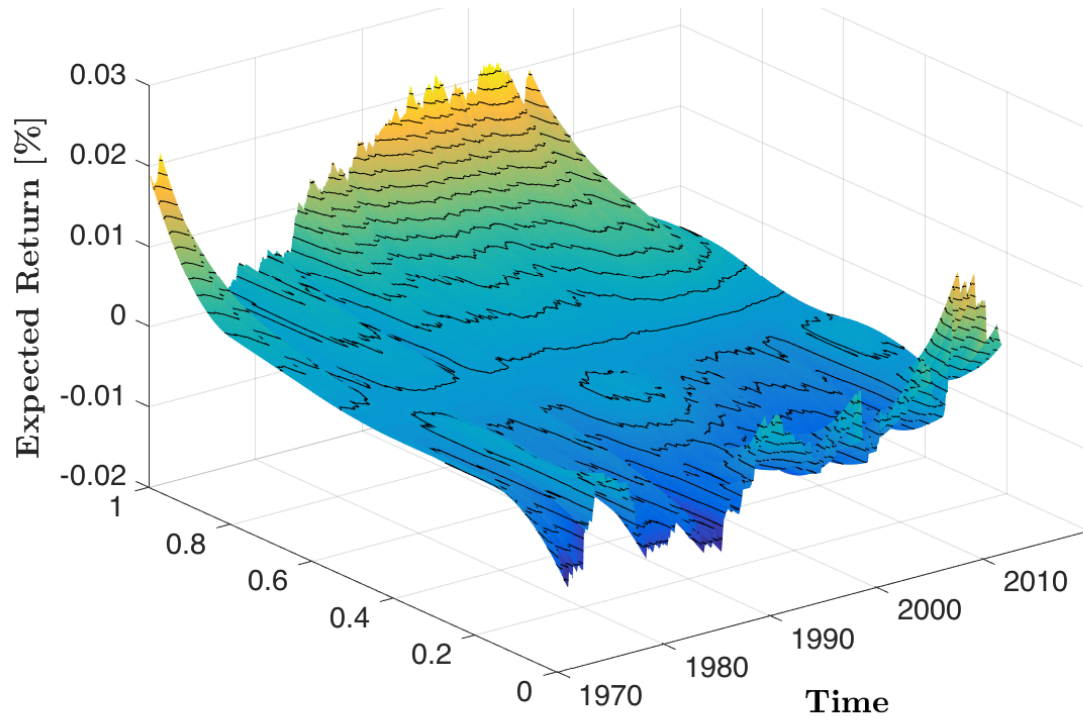
Figure 17: **Time-varying Conditional Mean Function: Intermediate Momentum**



Intermediate Momentum (normalized)

Effect of intermediate momentum on average returns (conditional on all other selected characteristics).

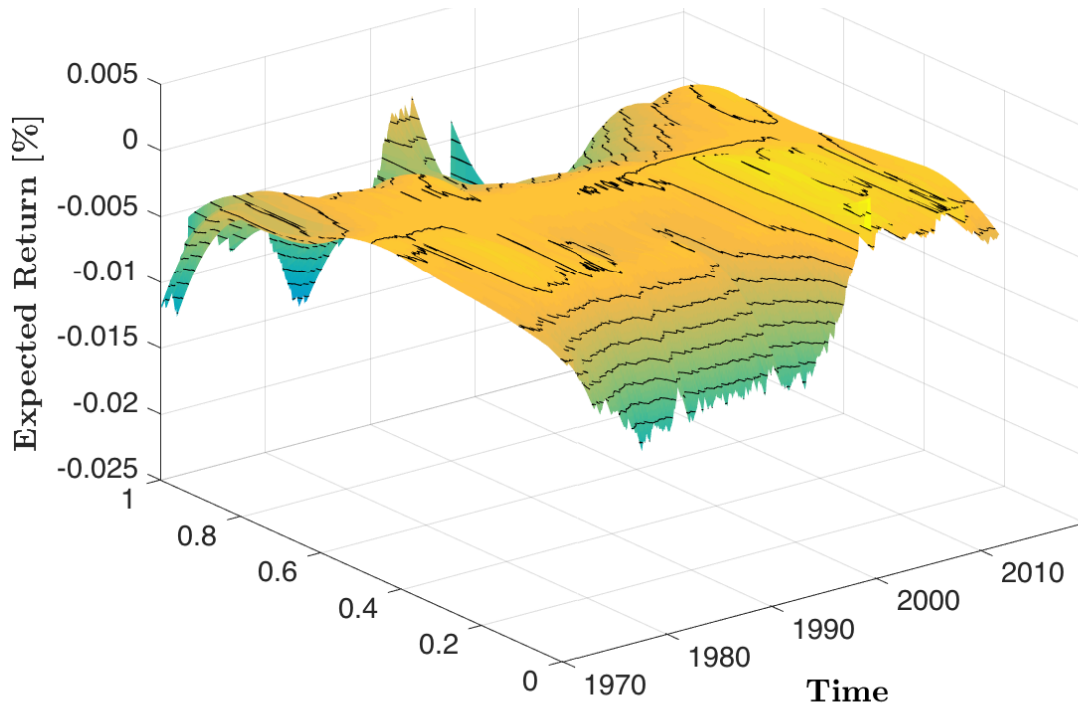
Figure 18: **Time-varying Conditional Mean Function: Momentum**



Momentum (normalized)

Effect of momentum on average returns (conditional on all other selected characteristics).

Figure 19: **Time-varying Conditional Mean Function: Idiosyncratic Volatility**



Idiosyncratic Volatility (normalized)

Effect of idiosyncratic volatility on average returns (conditional on all other selected characteristics).

Table 1: Descriptive Statistics for Firm Characteristics

This table reports average returns, time series standard deviations, and number of observations for the firm characteristics discussed in section IV. The sample period is July 1983 to June 2014.

	r_t	Beta	ME_{t-1}	Book-to Market	Profitability	Investment	Idio Vol	r_{12-2}	r_{12-7}	r_{1-0}	r_{36-13}	Cash
Mean	0.01	1.03	1.32	0.91	1.14	0.90	0.03	0.15	0.08	0.01	0.35	0.14
Std	(0.15)	(0.57)	(6.33)	(0.86)	1(8.68)	3(9.91)	(0.02)	(0.57)	(0.38)	(0.15)	(1.00)	(0.17)
Nobs	3,797	3,262	3,822	3,644	3,626	3,671	3,825	3,805	3,813	3,822	3,212	3,719
	Total Assets	D&A	Fixed Costs	Free CF	Capex	$Turnover_{t-1}$	Oper Accruals	Oper Leverage	P-C-M	Tobin's Q	ROE	bid-ask Spread
Mean	2,593.64	0.04	1.45	-0.25	0.41	0.08	-16.64	1.09	-0.94	1.87	0.00	0.04
Std	(185.68)	(0.04)	(31.53)	(11.21)	(11.85)	(0.14)	(1,078.33)	(1.00)	(39.61)	(2.99)	(14.05)	(0.06)
Nobs	3,756	3,625	3,347	3,639	3,397	3,622	3,493	3,737	3,704	3,756	3,519	3,825

Table 2: Returns of 10 Portfolios Sorted on Characteristics

This table reports average returns and time-series standard deviations for 10 portfolios sorted on firm characteristics discussed in Section IV. The sample period is July 1983 to June 2014.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P10-P1
Beta	16.95 (14.61)	16.89 (16.34)	17.41 (17.25)	17.01 (18.12)	16.40 (19.11)	16.44 (20.70)	17.17 (22.89)	15.09 (24.93)	15.54 (28.49)	14.07 (34.93)	-2.88 (25.65)
ME_{t-1}	33.06 (29.82)	16.34 (24.63)	14.95 (23.23)	14.61 (23.23)	14.74 (22.95)	15.47 (22.07)	14.70 (21.44)	14.01 (20.43)	13.30 (18.86)	11.73 (16.46)	-21.33 (25.03)
Book-to-Market	9.20 (24.15)	12.09 (22.41)	13.20 (21.15)	14.84 (21.07)	15.80 (20.26)	17.03 (20.39)	18.06 (20.40)	18.83 (20.85)	20.09 (21.46)	23.81 (24.20)	14.61 (16.02)
Profitability	14.99 (25.34)	14.22 (20.54)	15.39 (20.18)	15.43 (20.20)	16.17 (20.61)	16.21 (20.50)	16.56 (20.76)	17.43 (21.45)	18.33 (21.64)	18.22 (23.01)	3.23 (12.53)
Investment	22.44 (28.79)	20.78 (22.57)	19.02 (20.21)	17.73 (18.66)	16.52 (18.52)	16.19 (18.77)	15.49 (19.37)	14.17 (20.35)	12.62 (22.38)	7.98 (25.22)	-14.46 (12.85)
Idio Vol	12.85 (13.61)	14.96 (16.12)	15.92 (17.51)	16.85 (19.05)	17.75 (20.36)	17.46 (22.13)	18.12 (23.90)	16.37 (25.30)	15.88 (28.21)	16.80 (32.85)	3.95 (25.89)
r_{12-2}	14.77 (33.49)	12.70 (24.91)	13.24 (22.23)	13.54 (19.98)	14.62 (19.29)	15.14 (18.44)	16.57 (18.48)	18.71 (19.01)	20.12 (20.78)	23.52 (25.04)	8.75 (25.14)
r_{12-7}	13.30 (29.42)	12.90 (23.32)	13.39 (21.05)	15.32 (19.88)	16.22 (19.10)	16.03 (18.99)	16.80 (19.06)	17.81 (19.77)	19.71 (21.35)	21.45 (24.95)	8.15 (17.19)
r_{1-0}	36.62 (32.12)	21.40 (24.44)	18.78 (21.95)	16.84 (20.61)	15.69 (19.47)	14.70 (18.70)	12.86 (18.44)	11.72 (18.78)	9.97 (20.03)	4.07 (23.86)	-32.55 (21.69)
r_{36-13}	23.75 (31.28)	19.54 (24.48)	18.84 (21.97)	15.96 (19.76)	16.04 (19.19)	15.44 (18.44)	15.13 (18.68)	13.92 (19.12)	13.53 (20.69)	10.78 (24.68)	-12.97 (20.44)
Cash	15.52 (20.59)	15.13 (20.64)	15.33 (20.01)	15.41 (20.58)	16.70 (21.14)	16.47 (21.11)	17.33 (21.64)	17.16 (22.07)	16.30 (23.34)	17.60 (24.78)	2.07 (14.44)
Total Assets	21.62 (27.43)	19.25 (25.15)	17.65 (23.77)	16.09 (22.73)	15.75 (21.60)	15.74 (21.09)	15.70 (20.60)	14.51 (20.40)	13.95 (19.61)	12.67 (17.57)	-8.95 (21.57)
D&A	12.77 (22.73)	12.68 (21.43)	15.09 (20.88)	15.61 (21.17)	16.65 (20.72)	17.03 (21.04)	16.67 (20.29)	18.65 (20.44)	18.62 (21.19)	19.17 (23.81)	6.41 (11.26)

continued on next page

Table 3: Continued from Previous Page

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P10-P1
Fixed Costs	16.09 (19.95)	15.34 (19.92)	15.28 (20.43)	16.90 (20.53)	14.59 (20.00)	15.66 (20.50)	15.94 (20.84)	17.32 (22.15)	18.55 (24.40)	17.25 (29.11)	1.16 (19.51)
Free CF	15.18 (29.79)	15.95 (25.06)	16.82 (22.27)	16.48 (21.12)	16.19 (20.31)	16.85 (19.41)	15.91 (18.89)	16.12 (18.60)	16.53 (18.92)	16.91 (20.19)	1.73 (15.28)
Capex	18.85 (23.01)	18.31 (21.88)	17.14 (21.20)	16.90 (20.88)	17.09 (20.84)	16.75 (20.75)	16.39 (20.74)	14.74 (20.71)	13.94 (21.27)	12.85 (23.07)	-6.00 (12.43)
$Turnover_{t-1}$	12.48 (15.69)	14.89 (17.96)	15.66 (19.16)	16.35 (19.86)	17.95 (20.84)	17.74 (21.45)	17.71 (22.35)	17.08 (24.15)	18.06 (26.34)	15.04 (29.89)	2.56 (21.95)
Oper Accruals	17.15 (25.65)	18.53 (21.47)	18.21 (20.07)	18.13 (19.61)	17.39 (19.70)	17.09 (19.77)	16.14 (20.07)	15.62 (20.79)	14.42 (21.30)	10.28 (24.36)	-6.87 (8.59)
Oper Leverage	12.51 (20.42)	13.44 (21.50)	15.25 (21.42)	16.33 (21.72)	15.84 (21.56)	17.18 (21.46)	16.93 (21.38)	18.29 (21.99)	18.24 (22.15)	18.96 (21.36)	6.46 (12.10)
P-C-M	18.14 (25.74)	15.84 (21.59)	16.00 (20.73)	16.17 (21.36)	16.27 (20.96)	15.97 (20.52)	15.86 (20.76)	15.84 (20.58)	16.25 (20.85)	16.62 (21.45)	-1.52 (10.51)
Tobin's Q	23.20 (22.46)	19.63 (21.47)	18.62 (21.06)	17.81 (21.03)	16.86 (21.14)	16.28 (20.86)	14.94 (21.07)	14.34 (21.53)	12.11 (21.99)	9.15 (24.14)	-14.05 (14.99)
ROE	17.76 (32.18)	19.08 (26.37)	18.42 (22.08)	17.69 (20.30)	16.59 (18.93)	15.31 (18.50)	15.15 (18.33)	15.23 (19.03)	14.54 (19.74)	13.16 (21.83)	-4.60 (18.76)
bid-ask Spread	13.89 (18.50)	15.31 (18.52)	15.41 (19.05)	15.87 (20.14)	15.79 (21.01)	16.24 (22.28)	16.03 (22.99)	15.50 (23.98)	16.36 (25.28)	22.55 (29.14)	8.66 (21.73)

Table 4: Selected Characteristics in Non-parametric Model

This table reports the selected characteristics from the universe of 25 firm characteristics discussed in section IV for different numbers of knots. Large firms are all firms above the median firm size, and “early” corresponds to the first half of the sample. The sample period is July 1983 to June 2014.

Number of selected Characteristics	All Stocks Full Sample 5 Knots (1)	All Stocks Full Sample 10 Knots (2)	Large Stocks Full Sample 10 Knots (3)	All Stocks Early Sample 5 Knots (4)	All Stocks Late Sample 5 Knots (5)
1	$Market\ Cap_{t-1}$	$Market\ Cap_{t-1}$	$Book - to - Market$	$Market\ Cap_{t-1}$	$Market\ Cap_{t-1}$
2	$Turnover_{t_1}$	$Turnover_{t_1}$	$Investment$	$Turnover_{t_1}$	$Turnover_{t_1}$
3	$Book - to - Market$	$Investment$	r_{1-0}	$Book - to - Market$	$Depreciation$
4	$Depreciation$	HHI	r_{12-7}	r_{1-0}	$Profitability$
5	$Profitability$	r_{1-0}	r_{12-2}	r_{12-7}	$Investment$
6	$Investment$	r_{12-7}	$Idiosyncratic\ Vol$	r_{12-2}	HHI
7	HHI	r_{12-2}		$Beta$	r_{1-0}
8	r_{1-0}	$Idiosyncratic\ Vol$		$Idiosyncratic\ Vol$	r_{12-7}
9	r_{12-7}				
10	r_{12-2}				
11	$Idiosyncratic\ Vol$				