Dissecting Characteristics Non-Parametrically*

Joachim Freyberger† Andreas Neuhierl‡ Michael Weber§

This version: May 2016

Abstract

Academic research has identified more than 300 characteristics associated with cross-sectional return premia. We propose a novel, non-parametric methodolody to test which characteristics provide independent information for the cross section of expected returns. We use the adaptive group LASSO (least absolute shrinkage and selection operator) to select characteristics and estimate splines over intervals of the characteristic distribution. The proposed method overcomes issues traditional methods face and is robust to outliers. Many of the previously identified return predictors do not provide incremental information for expected returns and non-linearities 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 Gene Fama, Bryan Kelly, Leonid Kogan, Stavros Panageas, Ľuboš Pástor, and seminar participants at the University of Chicago 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, Email: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 which 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 only a small number of factors as being important for the cross section of expected return.

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; (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, they make strong functional form assumptions, and they 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 non-parametric methodology to determine which firm characteristics provide independent information for the cross section of expected returns without imposing strong functional forms. We estimate *smooth* functions over intervals (*knots*) of the distribution of many firm characteristics to allow for non-linearities. Specifically, we use a group LASSO (least absolute shrinkage and selection operator) procedure suggested by Huang, Horowitz, and Wei (2010). This procedure achieves two goals: (i) model selection: which characteristics have incremental predictive power for expected returns, given the other characteristics; (ii) non-parametric estimation: estimating the effect of characteristics on returns non-parametrically. In our empirical application, we estimate quadratic splines.

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

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

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 8 variables, including size, idiosyncratic volatility, and return-based predictors, have independent explanatory power for expected returns for the full sample and all stocks using ten knots for interpolation. 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.

We compare the out-of-sample performance of the non-parametric 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 portfolios going long stocks with the 10% highest predicted returns and shorting stocks with the 10% lowest predicted returns. The non-parametric model generates an average Sharpe ratio of 1.72 compared to 0.97 for the linear model. The linear model selects substantially more characteristics in sample but performs worse out of sample.

We also 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 for expected returns. Momentum returns conditional on other return predictors vary substantially over time. For example, we also 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 many variables such as size (Banz (1981)), the book-to-market ratio (Rosenberg et al. (1985)), leverage (Bhandari (1988)), or earnings-to-price ratios (Basu (1983)) contain additional

independent information for expected returns. Fama and French (1992) synthesize these findings and Fama and French (1993) show a three factors model with the market return, a size, and a value factor can explain cross sections of stocks sorted on characteristics which 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.

In the following 20 years, many researchers joined a "fishing expedition" to identify characteristics and factor exposures the three-factor model cannot explain. Harvey et al. (2016) list over 300 published papers which study the cross section of expected returns and propose a t-statistic of 3 for new factors to account for multiple testing on a common dataset. Figure 3 shows the suggested adjustment over time.

II Current Methodology

A. Expected Returns and the Curse of Dimensionality

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

$$E[R_{it}|C_{it-1}]. \tag{1}$$

Portfolio sorts are a standard practice 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.[3] Second, portfolio sorts offer little formal guidance to discriminate between characteristics. Imagine jointly sorting stocks on size and book-to-market and finding a value premium only for the smallest stocks. Does book-to-market now provide independent information? Fama and French (2008) call this second shortcoming "awkward". Third, portfolio sorts impose a strong functional form on the conditional mean function estimating a constant expected return over a part of the characteristic distribution, such as the smallest 10% of stocks. Fama and French (2008) call this third shortcoming "clumsy".[4] Nonetheless, portfolio sorts are by far the most commonly used technique to analyze which characteristics have predictive power for expected returns.

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

$$R_{it} = \alpha + \sum_{s=1}^{S} \beta_j C_{s,it-1} + \varepsilon_{it+1}. \tag{2}$$

Linear regressions allow to study the predictive power for expected returns of many characteristics jointly but they also have potential pitfalls. First, there is no a priori reason why the conditional mean function should be linear.[5] Fama and French (2008) estimate linear regressions as in equation (2) to dissect anomalies, but raise concerns over potential non-linearities. They make ad-hoc adjustments and use the log book-to-market ratio as an explanatory variable, for example. Second, outliers might drive point estimates in linear regressions. Third, small, illiquid stocks might have a large influence on point estimates as 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).

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

[4]Portfolio sorts are a restricted form of non-parametric regression. We will use the similarities of portfolio sorts and non-parametric regressions to develop intuition for our proposed framework.

[5]Fama 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.

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, "[P]ortfolio sorts are really the same thing as nonparametric cross-sectional regressions, using nonoverlapping histogram weights." Additional assumptions are necessary to create a formal equivalence, but this statement contains valuable intuition to study the problem of modeling the conditional mean function more formally. We will first show a formal equivalence between portfolio sorting and regressions and then outline how this naturally motivates the use of non-parametric methods.

We use R to denote excess returns and C to denote firm characteristics.

Suppose we observe returns R_{it} and a single characteristic C_{it-1} for stocks $i = 1, \ldots, N$ and time periods $t = 1, \ldots, T$. 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 $C_{it-1} \in I_{lt}$, where I_{lt} indicates an interval of the distribution for a given firm characteristic. Take a firm with lagged market cap in the 45^{th} percentile of the firm size distribution. We would sort that stock in the 5^{th} out of 10 portfolios in period t. For each time period t, let N_{lt} be the number of stocks in portfolio l,

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

N_t is the total number of stocks in period t.

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

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

The difference in average returns between portfolios l and l', or the excess return

$e(l, l')$, is

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

which is the intercept in a (time series) regression of the difference in portfolio returns, $P_{lt} - P_{l't}$, on a constant.[6]

Alternatively, we can run a pooled time series-cross sectional regression of excess returns on a dummy variable which equals 1 if firm i is in portfolio $l \in L$ in period t, $\mathbf{1}(C_{it-1} \in I_{lt})$:

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

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

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

It then follows

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

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

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

[6]We only consider univariate portfolios sorts in this example to gain intuition.

and

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

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

If the number of stocks in a portfolio, N_{lt}, changes over time, then portfolio sorts and regressions typically differ. There are two ways to restore equivalence. 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

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

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 $\tilde{\beta}_l - \tilde{\beta}_{l'} = e(l, l')$.

We will use the relationship between portfolio sorts and regressions to develop intuition for our non-parametric estimators in Section III and show how we can interpret portfolio sorts as a special case of non-parametric estimation.

III Non-parametric Estimation

We now return to the more general case trying to understand which characteristics, C, provide independent information for expected returns. Suppose we knew the conditional

mean function $m_t(c) \equiv E(R_{it} \mid C_{it-1} = c)$.[7] Then,

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

where f is the density function of the characteristic in period $t-1$. Hence, to get the return of portfolio l we can just 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, $\mathbf{1}$,

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

Also assume the quantiles of the distribution of a characteristic, C_{it-1}, determine the intervals I_{lt} and these quantiles might change over time. For example, with two portfolios we might have

$$I_{1t} = [\min(C_{it-1}), \text{median}(C_{it-1})] \quad \text{and} \quad I_{2t} = (\text{median}(C_{it-1}), \max(C_{it-1})].$$

Now define $\tilde{C}_{it-1} = F_t(C_{it-1})$, which denotes the rank of characteristic, C_{it-1}, normalized to the unit interval, $[0, 1]$, for a fixed time period t.

Then, $\mathbf{1}(C_{it-1} \in I_{lt}) = \mathbf{1}(\tilde{C}_{it-1} \in \tilde{I}_l)$ and

$$R_{it} = \sum_{l=1}^{L} \beta_l \mathbf{1}(\tilde{C}_{it-1} \in \tilde{I}_l) + \varepsilon_{it},$$

where $\tilde{I}_l$ does not depend on t.

[7] We take the expected excess return for a fixed time period t.

For example with the two portfolios we get

$$\mathbf{1}(C_{it-1} \in I_{1t}) = \mathbf{1}(\tilde{C}_{it-1} \in [0, 0.5]) \quad \text{and} \quad \mathbf{1}(C_{it-1} \in I_{2t}) = \mathbf{1}(\tilde{C}_{it-1} \in (0.5, 1]). \tag{3}$$

In non-parametric estimation, indicator functions of the form $\mathbf{1}(\tilde{X}_{it-1} \in \tilde{I}_l)$ are called constant splines. Estimating the conditional mean function, m_t, using constant splines, means approximating it by a step function. In this sense, portfolio sorting is a special case of non-parametric regression when the number of portfolios approaches infinity. While a step function is non-smooth and has therefore undesirable theoretical properties as a non-parametric estimator, we build on this intuition to estimate m_t non-parametrically.

Figure 4 – Figure 6 illustrates the intuition behind the relationship between portfolio sorts and non-parametric 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 non-parametric estimator we propose below for simulated data.

We see in Figure 4 most of the dispersion in book-to-market ratios and returns is in the extreme portfolios. There is little variation in returns across portfolios 3 to 5 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, portfolios 1 and 5 have difficulties capturing the non-linearities we see in the data.

Figure 5 documents a non-parametric 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 portfolio means provide a better fit in the tails of the distribution once we allow for more portfolios. The predictions from the non-parametric estimator and portfolio mean returns become more comparable once the number of portfolio increases. Therefore, as the number of portfolios grows, portfolio sorts converge to a non-parametric estimator.

A. Normalization of Characteristics

We now come back to the normalization we discuss above in equation (3). Empirically, we estimate a transformation of the conditional mean function m_t non-parametrically. We now briefly discuss the transformation and the advantages.

Define the conditional mean function m for S characteristics as

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

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

$$\tilde{m}_t(C_1, \ldots, C_S) = m_t(F_{1,t}^{-1}(C_1), \ldots, F_{S,t}^{-1}(C_S).$$

Then

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

Therefore, knowing the conditional mean function m_t is equivalent to knowing the transformed conditional mean function $\tilde{m}_t$ and using a transformation is without loss of generality. Instead of estimating m_t, we will estimate $\tilde{m}_t$. This transformation naturally relates to portfolio sorting. We are not interested in the value of a characteristic in isolation when we sort stocks in portfolios but we care 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 was a large firm in the 1960s, but is not a large firm today. Our normalization considers the relative size in the cross section rather than the absolute size similar to portfolio sorting.

Therefore, we choose $F_{s,t}(\cdot)$ to be a rank transformation of $C_{s,it-1}$ such that the cross-sectional distribution of a given characteristic lies in the unit interval, i.e., $C_{s,it-1} \in [0, 1]$. Specifically, let

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

Here, $\text{rank}(\min_{i=1\ldots,N} C_{s,it-1}) = 1$ and $\text{rank}(\max_{i=1\ldots,N} C_{s,it-1}) = N$. 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. The general econometric theory we discuss below (model selection, consistency, etc.) also applies to any other monotonic transformation or the non-transformed conditional mean function.

While knowing m_t is equivalent to knowing $\tilde{m}_t$, in finite samples, the estimates of the two typically differ,

$$\hat{\tilde{m}}_t(C_1, \ldots, C_S) \neq \hat{m}_t(F_{1,t}^{-1}(C_1), \ldots, F_{S,t}^{-1}(C_S)).$$

In numerical simulations, we found $\tilde{m}$ yields better out-of-sample predictions than m. The transformed estimator is 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, nicely relates to portfolio sorting, and works well in finite sample as it is robust to outliers, which is a concern in linear regressions as stressed in Cochrane (2011).

B. 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 4 characteristics and partition each characteristics into 5 portfolios, then we end up with $5^4 = 625$ portfolios. First, it would be impractical to investigate 625 portfolio returns. Second, as the number of characteristics increases, we will only have very few observations in each portfolio. In non-parametric regression, an analogous problem arises. Estimating the conditional mean function $m_t(c) \equiv E(R_{it} \mid C_{it} = c)$ fully non-parametrically, results in slow rates of convergence with many regressors. This is often referred to as the "curse of dimensionality" (see Stone (1982) for a formal treatment). Nevertheless, if we are interested in which characteristics provide additional information for expected returns given other characteristics, we cannot look at each characteristic in isolation.

If the number of characteristics S was small, we could estimate the conditional mean

function consistently and obtain an estimator with good finite sample properties. The optimal rate of convergence in mean square (assuming the conditional mean function m_t is twice continuously differentiable) is $N^{-4/(4+S)}$, which is always smaller than the rate of convergence for the parametric estimator of N^{-1}. The rate of convergence decreases as S increases, i.e., we get an estimator with poor finite sample properties if the number of characteristics is large.

As a simple example suppose the number of stocks is $N = 1,000$ and the number of characteristics is $S = 1$. Then $N^{-4/(4+S)} = 1,000^{-4/5}$. Now suppose, instead, we have 10 characteristics and choose the number of stocks N^* such that

$$(N^*)^{-4/(4+10)} = 1,000^{-4/5} \Rightarrow N^* = 1,000^{14/5} \approx 251,000,000.$$

Therefore, we need 251 million stocks to get similar finite sample properties of an estimated conditional mean function for 10 characteristics and a one dimensional conditional mean function with $1,000$ stocks.

Conversely suppose that $N = 1,000$ and $S = 10$. Then $N^{-4/(4+S)} = 1,000^{-4/14}$. Now find N^* such that

$$(N^*)^{-4/(4+1)} = 1000^{-4/14} \Rightarrow N^* = 1000^{5/14} < 12.$$

Therefore, estimating a 10 dimensional function with 1000 observations is similar to estimating a one dimensional function with less than 12 observations.

A natural solution in the regression framework is to assume an additive model. To get around the curse of dimensionality, we make the following assumption:

$$\tilde{m}_t(\tilde{C}_1, \ldots, \tilde{C}_S) = \sum_{s=1}^{S} \tilde{m}_{st}(\tilde{C}_s).$$

The main theoretical advantage of this specification is that the rate of convergence is $N^{-4/5}$, 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

$$\frac{\partial^2 \tilde{m}_t(\tilde{C}_1, \ldots, \tilde{C}_S)}{\partial \tilde{C}_s \partial \tilde{C}_{s'}} = 0$$

for all $s \neq s'$. The additive model does not allow for interactions between characteristics. For example, the predictive power of the book-to-market ratio for expected returns does not depend on firm size. A simple fix is to add certain interactions as additional regressors. 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.

While 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 functions, $\tilde{m}_t$, simultaneously.

C. Time Invariant Conditional Mean Function

We now discuss one last assumption before we discuss estimation. Assume the conditional mean function, $\tilde{m}_t$, as a function of the rank of the characteristic does not depend on time, t.

Assumption: $\tilde{m}_t$ does not depend on t.

This assumption is not necessary for non-parametric estimation, but speeds up estimation, facilitates the interpretation, and maintains portfolio sorts as a special case of our estimator.[8] In our empirical tests in Section V, we estimate our model over subsamples and also estimate rolling specifications, i.e., allow for variation in the conditional mean function over time. Again, the analogy to portfolio returns is apparent. We often estimate portfolio mean returns over subsamples and rolling over time.

[8]We implicitly make this assumption whenever we look at 10 portfolios sorted on size, for example.

With the assumption of a time invariant conditional mean function, we can write

$$E(R_{it} \mid C_{it-1}) = \tilde{m}(\tilde{C}_{it-1}).$$

We now briefly discuss the assumption. First, similar to portfolio sorting, we focus on the rank of a characteristic in the cross-section, rather than the numerical value. Take firm size as an example. The assumption says once we look at the relative size in the cross section of firms, the relationship between size and expected returns does not depend on time t.

Second, if the distribution of a characteristic does not depend on time, i.e., is time stationary, then the conditional mean function does not change over time. The book-to-market ratio is a good example.

Third, some characteristics might lose their predictive power for expected returns over time. The size effect is a recent example. McLean and Pontiff (2016) show for 97 return predictors predictability decresases by 58% post publication.

D. Adaptive Group LASSO

We use a group LASSO procedure suggested by Huang et al. (2010) for estimation and to select those characteristics which provide incremental information for expected returns, i.e., for model selection. The group LASSO estimates the conditional mean function non-parametrically using splines and sets the summand of the conditional mean function for a given characteristic to 0 if the characteristic does not help predict expected returns.

To recap, we are interested in modeling excess returns as a function of characteristics, i.e.,

$$R_{it} = \sum_{s=1}^{S} \tilde{m}_s(\tilde{C}_{it-1}) + \varepsilon_{it}, \tag{4}$$

where $\tilde{m}_s(\cdot)$ are unknown functions. The adaptive group LASSO is a two-step procedure to achieve model selection, that is, to discriminate between the $\tilde{m}_s$s, which

are constant, and the $\tilde{m}_s$s, which are not constant.[9]

Let $\tilde{I}_l$ for $l = 1, \ldots, L$ be a partition of the unit interval for a transformed characteristic. To estimate $\tilde{m}$, we use *quadratic* splines, i.e., we approximate $\tilde{m}$ as a quadratic function on each interval $\tilde{I}_l$. We choose these functions so that the endpoints are connected and $\tilde{m}$ is differentiable on $[0, 1]$. We will approximate each $\tilde{m}_s$ by a series expansion, i.e.,

$$\tilde{m}_s(\tilde{c}) \approx \sum_{k=1}^{L+2} b_{sk} p_k(\tilde{c}), \tag{5}$$

where $p_k(c)$ are basis splines. 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 also the number of parameters we have to estimate and hence the variance.

In the first step of the adaptive group LASSO, we obtain estimates of the coefficients as

$$\tilde{\boldsymbol{\beta}}_s = \underset{b_{sk}:s=1,\ldots,S;k=1,\ldots,L+2}{\arg\min} \left(\sum_{t=1}^{T} \sum_{i=1}^{N} R_{it} - \sum_{s=1}^{S} \sum_{k=1}^{L+2} b_{sk} p_k(\tilde{C}_{s,it-1}) \right)^2 + \lambda_1 \sum_{s=1}^{S} \left(\sum_{k=1}^{L+2} b_{sk}^2 \right)^{\frac{1}{2}}, \tag{6}$$

where $\tilde{\boldsymbol{\beta}}_s$ is an $(L + 2) \times S$ vector of b_{sk} estimates and λ_1 is a penalty parameter to minimize Bayes Information Criterion (BIC). The first part of equation (6) 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}s, it penalizes all coefficient associated with a given characteristic. Thus, we can set the point estimates of an entire expansion of $\tilde{m}$ for a given characteristic to 0 when the characteristic does not provide independent information for expected returns, i.e., the function associated with the characteristic is constant.[10] However, as in the linear model, the first step of the LASSO selects too many characteristics. Informally speaking, the LASSO gets all the non constant function right, but it does not get all the constant

[9] As in the linear model, the "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.

[10] A constant function means the characteristic has no predictive power for expected returns.

functions right. To address this problem, a second step is needed. To describe the second step, we first define the following weights

$$w_s = \begin{cases} \left(\sum_{k=1}^{L+2} \tilde{\beta}_{sk}^2\right)^{-\frac{1}{2}} & \text{if} \quad \sum_{k=1}^{L+2} \tilde{\beta}_{sk}^2 \neq 0 \\ \infty & \text{if} \quad \sum_{k=1}^{L+2} \tilde{\beta}_{sk}^2 = 0. \end{cases} \tag{7}$$

In the second step, the adaptive LASSO solves

$$\tilde{\boldsymbol{\beta}}_s = \underset{b_{sk}:s=1,\ldots,S;k=1,\ldots,L+2}{\arg\min} \left(\sum_{t=1}^{T} \sum_{i=1}^{N} R_{it} - \sum_{s=1}^{S} \sum_{k=1}^{L+2} b_{sk} p_k(\tilde{C}_{s,it-1}) \right)^2 + \lambda_2 \sum_{j=1}^{P} \left(w_s \sum_{k=1}^{L+2} b_{sk}^2 \right)^{\frac{1}{2}}, \tag{8}$$

where λ_2 is again chosen to minimize BIC. Huang et al. (2010) show the estimator from equation (8) is model selection consistent, i.e., it correctly selects the non-constant functions with probability approaching 1 as the sample size grows large.

Theoretically, the LASSO can deal with highly correlated regressors. The procedure is, therefore, well suited for our empirical application in which many firm characteristics are highly correlated. The matrix of regressors does not even have to have full rank, i.e., the number of regressors could be larger than the sample size. A linear model cannot handle these situations and a standard t-test does not work if the matrix is close to singular. In practice, the LASSO might still have difficulties to distinguish between highly correlated characteristics.

Consider two highly correlated characteristics and only one of them has predictive power for expected returns. It is possible the LASSO selects either one or none, because they are so highly correlated, but it will not select both. A linear model cannot handle this situation because the matrix of regressors is close to singular and the standard theory does not apply.

E. Confidence Bands

We estimate confidence bands for the conditional mean function $\tilde{m}(\tilde{c})$, which we approximate by $\sum_{k=1}^{L+2} b_k p_k(\tilde{c})$, and estimate by $\sum_{k=1}^{L+2} \hat{b}_k p_k(\tilde{c})$. Let $p(\tilde{c})$ be a vector of splines $p(\tilde{c}) = (p_1(\tilde{c}), \ldots, p_{L+2}(\tilde{c}))'$ and Σ be the $L+2 \times L+2$ covariance matrix of

$\sqrt{n}(\hat{b} - b)$. We define $\hat{\Sigma}$ as the heteroscedasticity-consistent estimator of Σ and define $\hat{\sigma}(\tilde{c}) = \sqrt{p(\tilde{c})'\hat{\Sigma}p(\tilde{c})}$.

The uniform confidence band is of the form

$$\left[\sum_{k=1}^{L+2} \hat{b}_k p_k(\tilde{c}) - d\hat{\sigma}(\tilde{c}) \, , \sum_{k=1}^{L+2} \hat{b}_k p_k(\tilde{c}) + d\hat{\sigma}(\tilde{c})\right].$$

We choose the constant d such that the band has the right coverage asymptotically.

Write,

$$\begin{aligned}
&P\left(\sum_{k=1}^{L+2} \hat{b}_k p_k(\tilde{c}) - \frac{d\hat{\sigma}(\tilde{c})}{\sqrt{n}} \leq m(\tilde{c}) \leq \sum_{k=1}^{L+2} \hat{b}_k p_k(\tilde{c}) + \frac{d\hat{\sigma}(\tilde{c})}{\sqrt{n}} \text{ for all } \tilde{c}\right) \\
&\quad \approx P\left(\sum_{k=1}^{L+2} \hat{b}_k p_k(\tilde{c}) - \frac{d\hat{\sigma}(\tilde{c})}{\sqrt{n}} \leq \sum_{k=1}^{L+2} b_k p_k(\tilde{c}) \leq \sum_{k=1}^{L+2} \hat{b}_k p_k(\tilde{c}) + \frac{d\hat{\sigma}(\tilde{c})}{\sqrt{n}}\right) \\
&\quad = P\left(\sup_{\tilde{c}} \left| \frac{\sum_{k=1}^{L+2} \sqrt{n}\left(\hat{b}_k - b_k\right) p_k(\tilde{c})}{\hat{\sigma}(\tilde{c})} \right| \leq d \text{ for all } \tilde{c}\right) \\
&\quad \approx P\left(\sup_{\tilde{c}} \left| \frac{Z'p(\tilde{c})}{\sqrt{p(\tilde{c})'\Sigma p(\tilde{c})}} \right| \leq d \text{ for all } \tilde{c}\right),
\end{aligned}$$

where $Z \sim N(0, \Sigma)$. The first approximation follows because $\tilde{m}(\tilde{c}) \approx \sum_{k=1}^{L+2} b_k p_k(\tilde{c})$. The underlying assumption is a standard undersmoothing condition, which says the approximation error decreases faster than the standard deviation. For the second approximation, we use the fact that $\sqrt{n}(\hat{b} - b) \to^d N(0, \Sigma)$ and that $\hat{\Sigma}$ is a consistent estimator of Σ. We now replace Σ by $\hat{\Sigma}$ and find d such that conditional on $\hat{\Sigma}$,

$$P\left(\sup_{\tilde{c}} \left| \frac{Z'p(\tilde{c})}{\sqrt{p(\tilde{c})'\hat{\Sigma}p(\tilde{c})}} \right| \leq d \text{ for all } d\right) = 0.95.$$

We can find this constant using simulations.

As a technical aside, the dimension of Z increases as the sample size increases. Nevertheless, our construction is valid (see Belloni, Chernozhukov, Chetverikov, and Kato (2015)).

F. Interpretation of the Conditional Mean Function

Let α_s be constants such that they sum to 0 across characteristics

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

Then,

$$\tilde{m}(\tilde{C}_1, \ldots, \tilde{C}_S) = \sum_{s=1}^{S} \tilde{m}_s(\tilde{C}_s) = \sum_{s=1}^{S} \left(\tilde{m}_s(\tilde{C}_s) + \alpha_s \right).$$

Therefore, the summands of 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 dependent on these constants. However, the constants matter when we plot an estimate of the conditional mean function for one characteristic, $\tilde{m}_s$.

Let $\bar{C}_{s,t-1}$ be the median of a given transformed characteristic s, $\tilde{C}_{s,it-1}$. Then,

$$\tilde{m}(\tilde{C}_1, \bar{C}_{st-1} \ldots, \bar{C}_{St-1}) = \tilde{m}_1(\tilde{C}_1) + \sum_{s=2}^{S} \tilde{m}_s(\bar{C}_{st-1}),$$

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 set all other characteristics to their median values. 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 now can interpret both the level and the slope of the function.

An alternative normalization is such that $\tilde{m}_1(0.5) = 0$, i.e., 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, might be 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.

G. Comparison of Linear & Non-parametric Models

We discussed above the relationship between portfolio sorts, linear regressions, and non-parametric estimation. We now want to compare a linear model with non-parametric models. The comparison helps us understand the potential pitfalls from assuming a linear relationship between characteristics and returns and gain some intuition why we might select a different number of characteristics in our empirical tests in Section V.

Suppose we observe excess returns R_{it} and a single characteristic, C distributed according to $C_{it-1} \sim U[0,1]$ for $i = 1, \ldots, N$ and $t = 1, \ldots, T$. Returns are generated by

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

where $E(\varepsilon_{it} \mid C_{it-1}) = 0$.

Without knowing the conditional mean function m, we could sort stocks into portfolios according to the distribution of the characteristic. Let stock i be in portfolio l at time t if $C_{it-1} \in I_{lt}$. C predicts returns if mean returns differ across portfolios, i.e., $e(l, l')$ is significantly different from 0. For example, we could construct 10 portfolios based on the intervals $I_{lt} = [(l-1)/10, l/10]$ and test if $e(1, 10)$ is significantly different from 0.

If we knew the conditional mean function m, then we could conclude C predicts returns if m varies with the characteristic. 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 $m(C_{it-1})$. For example, let $q_t(\alpha)$ be the α quantile of $m(C_{it-1})$ and let stock i be in portfolio l at time t if $m(C_{it-1}) \in [q_t((l-1)/10), q_t(l/10)]$. That is, we construct 10 portfolios based on return predictions. Portfolio 1 contains the 10% of stocks with lowest predicted returns and portfolio 10 contains the 10% of stocks with highest predicted returns.

If m is monotone, both sorting based on the value of characteristic and based on predicted returns, $m(C_{it-1})$ results in the same portfolios. However, if m is not monotone,

the "10-1 portfolio" return is higher when we sort based on m. As a simple example, suppose $m(c) = (c - 0.5)^2$. Then the "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 returns are generated by

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

where $E(\varepsilon_{it} \mid C_{1,it-1}, C_{2,it-1}) = 0$.

We are interested in whether the second characteristic provides independent information for expected returns conditional on the first characteristic. In this framework, $C_{2,it-1}$ does not provide independent information if

$$\frac{\partial m(c_1, c_2)}{\partial c_2} = 0 \text{ for all } c_1, c_2 \in [0,1],$$

which is testable if we knew $m(c_1, c_2)$.

Again, we can construct portfolios with a large spread in predicted returns based on the value of the conditional mean function, m. The idea is similar to construct 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 potential pitfalls of the linear model and how a non-parametric model can alleviate them.

Let us assume the following return-generating process

$$R_{it} = -0.5 + 0.6\sqrt{C_{1,it-1}} + 0.5C_{2,it-1}^2 + \varepsilon_{it}.$$

A regression of returns R_{it} on the characteristics $C_{1,it-1}$ and $C_{2,it-1}$ yields slope coefficients of around 0.5. 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 non-parametric model for the "10-1" hedge portfolio. The table shows monthly returns, standard deviations, and Sharpe

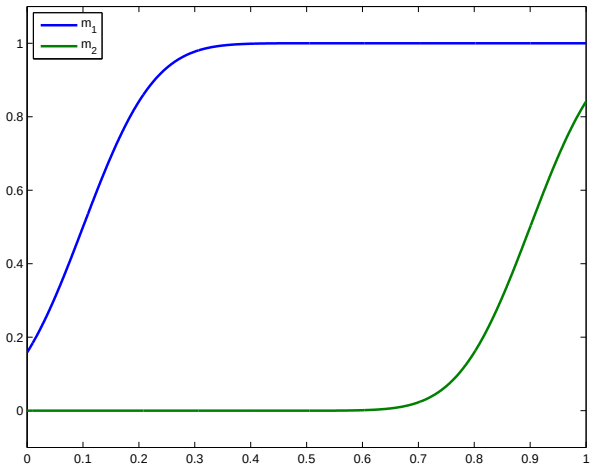


Figure 1:

ratios from a simulation for 500 stocks and 100 periods for both models.

	Linear	Non-parametric
Return	0.3506	0.3610
Std	0.0527	0.0539
Sharpe Ratio	6.6516	6.6998

The linear model and the non-parametric model result in similar predicted returns, standard deviations, and Sharpe ratios. The point estimaties for both characteristics, $C_{1,t}$ and $C_{2,t}$, are similar in both models and returns are monotone in both characteristics.

Let us now instead consider the following data-generating process:

$$R_{it} = \left(-0.5 + \Phi((C_{1,it-1} - 0.1)/0.1) + \Phi((C_{2,it-1} - 0.9)/0.1)\right)/2 + \varepsilon_{it},$$

where Φ denotes the standard normal cdf. The Figure 1 plots the two functions. In this example, a regression of R_{it} on $C_{1,it-1}$ and $C_{2,it-1}$ yields two slope coefficients of around 0.25. Hence, the predicted values of a linear model treat $C_{1,it-1}$ and $C_{2,it-1}$ identically, although they affect returns very differently. We again report the returns, standard deviations, and Sharpe ratios of the hedge portfolios:

	Linear	Non-parametric
Return	0.1988	0.3090
Std	0.0496	0.0556
Sharpe Ratio	4.0053	5.5612

Predicted returns of the non-parametric model are substantially higher compared to the linear model, with a similar standard deviations resulting in larger Sharpe ratios in the non-parametric method.

In a last example, we want to discuss how the linear and non-parametric model treat non-linear transformations of variables and why a linear model might select more variables in empirical settings. Consider returns being generated by the following process

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

with $C_{2,it-1} = C_{1,it-1}^2$, i.e., the second characteristic is just the square of the first characteristics. The linear model would select both $C_{1,it-1}$ and $C_{2,it-1}$, whereas the non-parametric model would only select $C_{1,it}$. In addition, the penalty term in the LASSO for the BIC information criterion is proportional to the number of parameters. In the non-parametric model with 10 knots, the penalty is proportional to 12 times the number of selected characteristics. In the linear model, the penalty is only proportional to 1 times the number characteristics. Taken together, we would expect the linear model to select more characteristics in sample. The out-of-sample performance of the linear model relative to the non-parametric model is unclear ex-ante and we compare the model performance in Section V.

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 twelve 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 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 has 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, e.g. the ICAPM (Merton

(1973)) or consumption models, as surveyed in Cochrane (2007) suggest that a low number of state variables can explain the cross section of returns and it is therefore unlikely all characteristics provide independent information for expected returns. To tackle the multi-dimensionality challenge, we now estimate the adaptive group LASSO with 5 and 10 knots.[11]

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 results 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, and 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.

The example of Tobin's Q shows the importance to condition on other characteristics to make inference 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 non-parametric model for different number of knots, sets of firms, and sample periods. We see in column (1) the baseline estimation for all stocks over the full sample period using 5 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

[11] The number of knots corresponds to the smoothing parameter we discuss in Section III.

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 non-parametric 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 non-parametric 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 the 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 non-parametric 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

Figure 9 to Figure 19 show the conditional mean function for our baseline non-parametric model for all stocks and 5 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 or the median characteristic in a given months.

We see in Figure 9 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 non-parametric model with the linear model out of sample. We estimate the non-parametric model for a period from 1963 to 1990 and carry out model selection with the adaptive group LASSO with ten knots but also use the adaptive LASSO for model selection in the linear model. We select the following characteristic in the non-parametric 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, 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 non-parametric model. The linear model might be misspecified and therefore select more variables.

We then create rolling monthly out-of-sample predictions for excess return using ten years of data for estimation and form two portfolios for each method. We buy the stocks with 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% for the non-parametric model to 1.72.

VI Conclusion

We propose a non-parametric methodology to tackle the challenge posed by Cochrane (2011) in his presidential address: which firm characteristics provide independent information for expected returns. We study 24 characteristics jointly and find only 6 to 11 characteristics provide independent information depending on number of interpolation

points (similar to the number of portfolios in portfolio sorts), sample period, and universe of stocks (large versus small stocks).

We compare our method to portfolio sorts and linear regressions and show it has superior out-of-sample performance and increases out-of-sample Sharpe ratios by 70%.

We see our paper as a starting point only. The next questions are: Are the characteristics we identify to provide information for expected returns related to exposures to factors? How many factors are important? Can we achieve a dimension reduction and identify K factors which can summarize the N independent dimension of expected returns we document 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.

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. 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.

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.

Lewellen, J. (2015). The cross section of expected stock returns. *Critical Finance Review (forthcoming)*.

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*, 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.

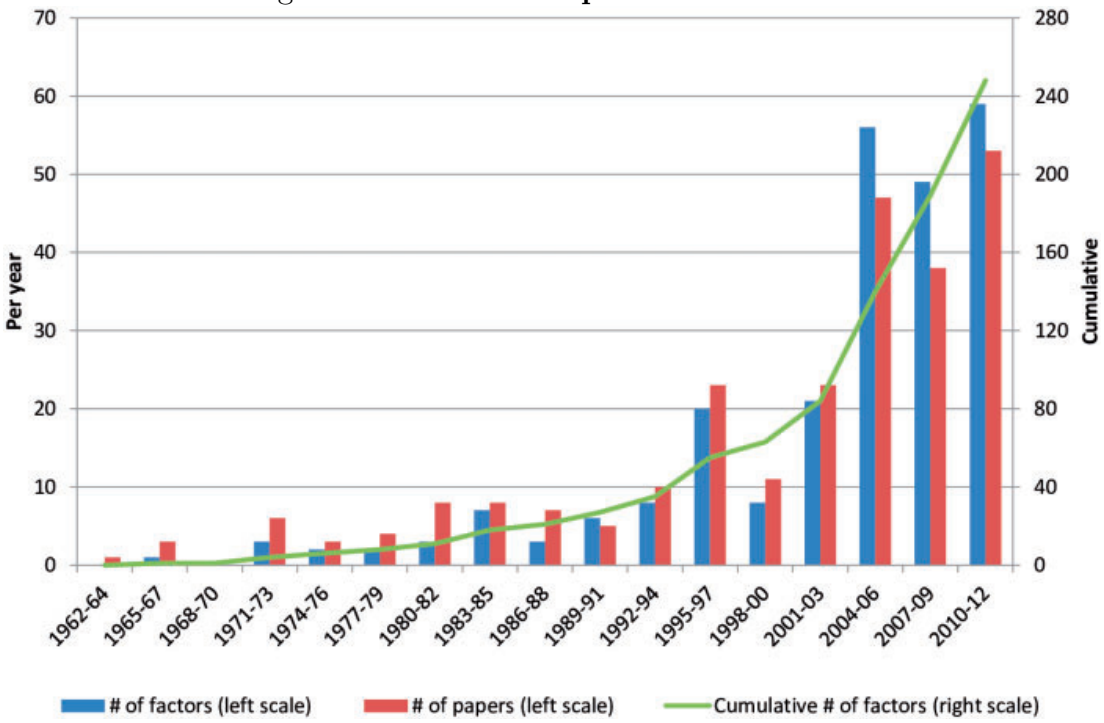
Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics 10*(4), 1040–1053.

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.

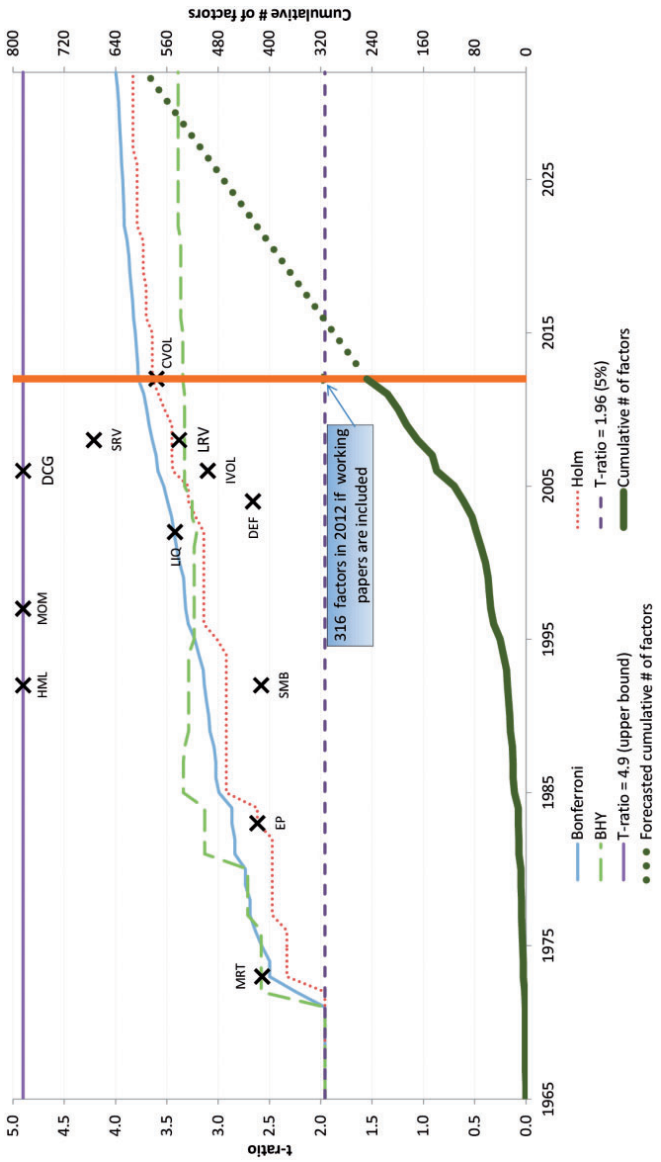
Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.

Figure 2: **Numbers of published Factors**



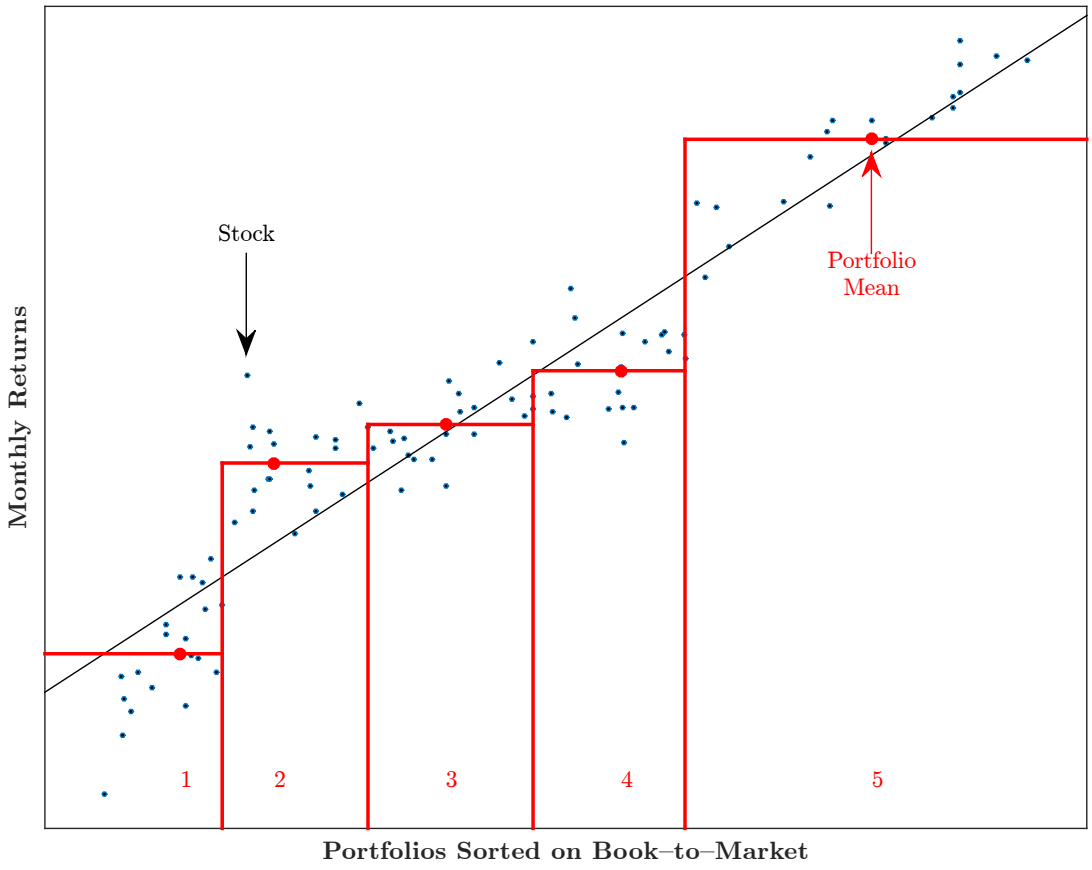
Number of discovered factors over time. Source: Figure 2 of Harvey, Liu, and Zhu (2016).

Figure 3: **Suggested t-stats adjustment**



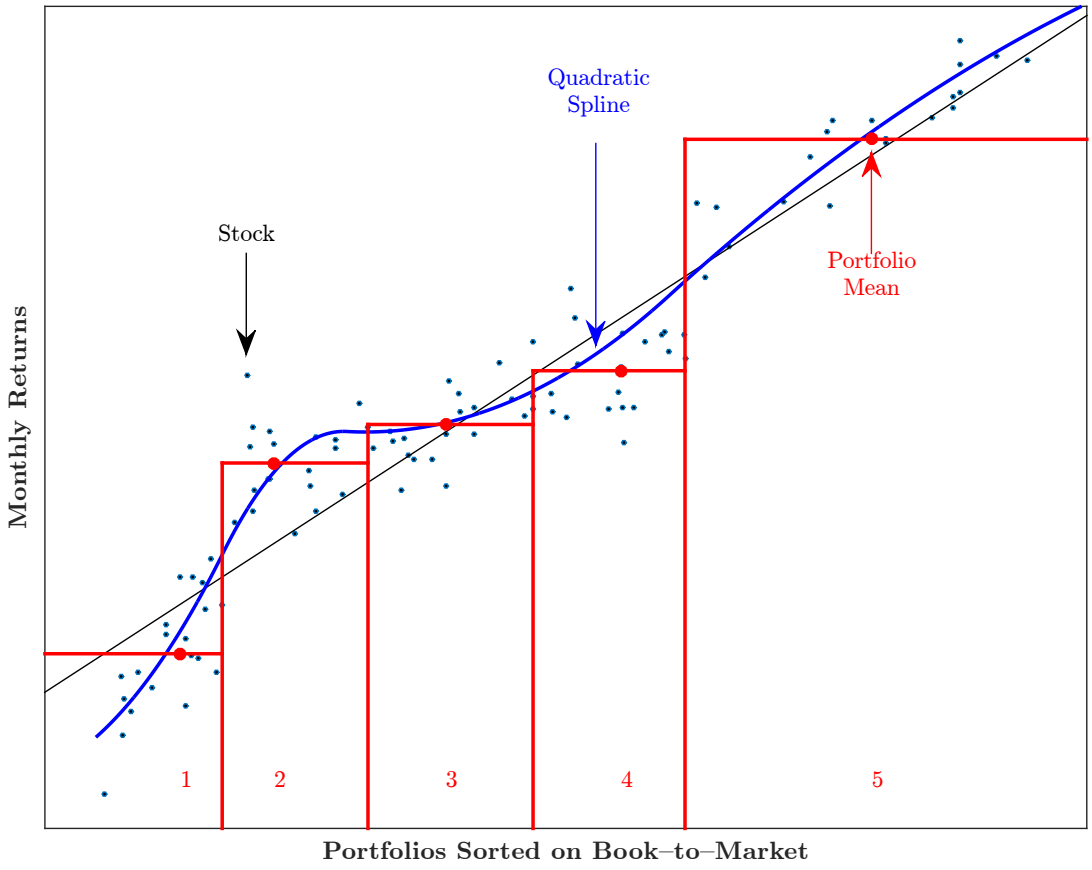
Recommendation to adjust t-statistics for multiple testing problem. Source: Figure 3 of Harvey, Liu, and Zhu (2016).

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



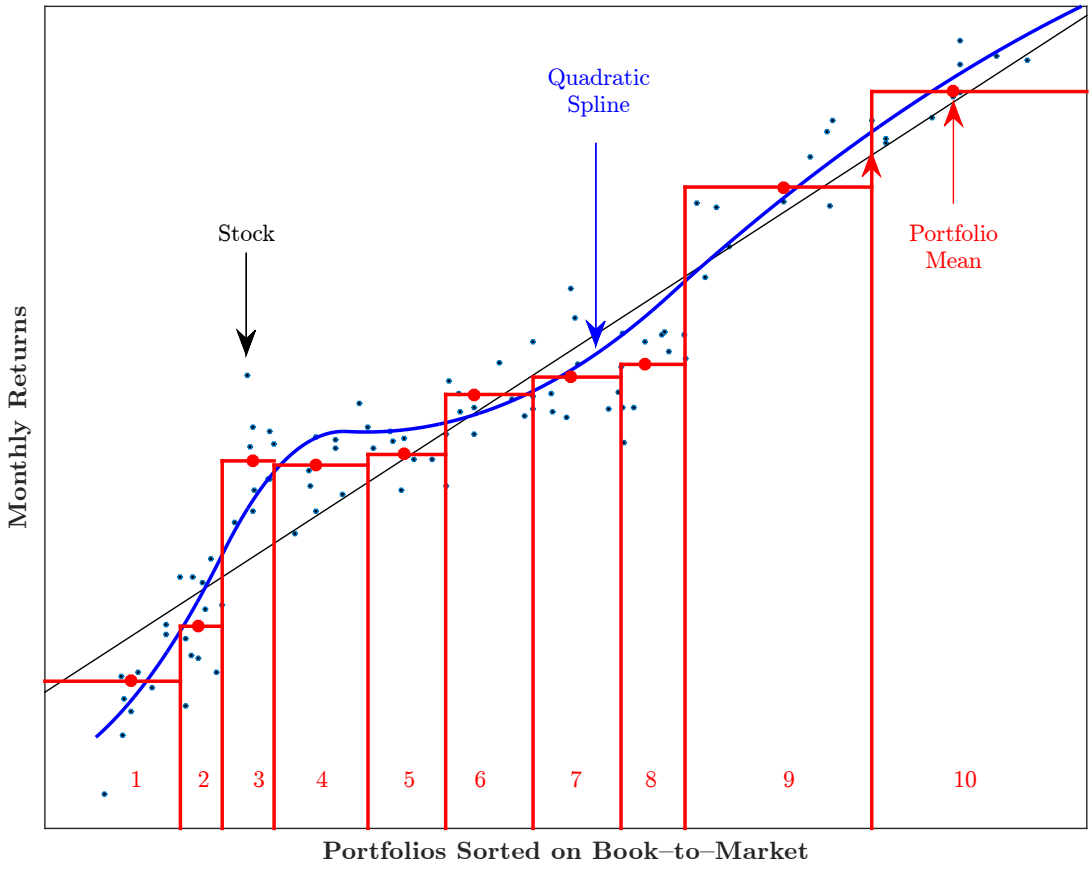
This figure plots expected returns on the y-axis against the book-to-market ratio on the x-axis as well as portfolio mean returns for simulated data.

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



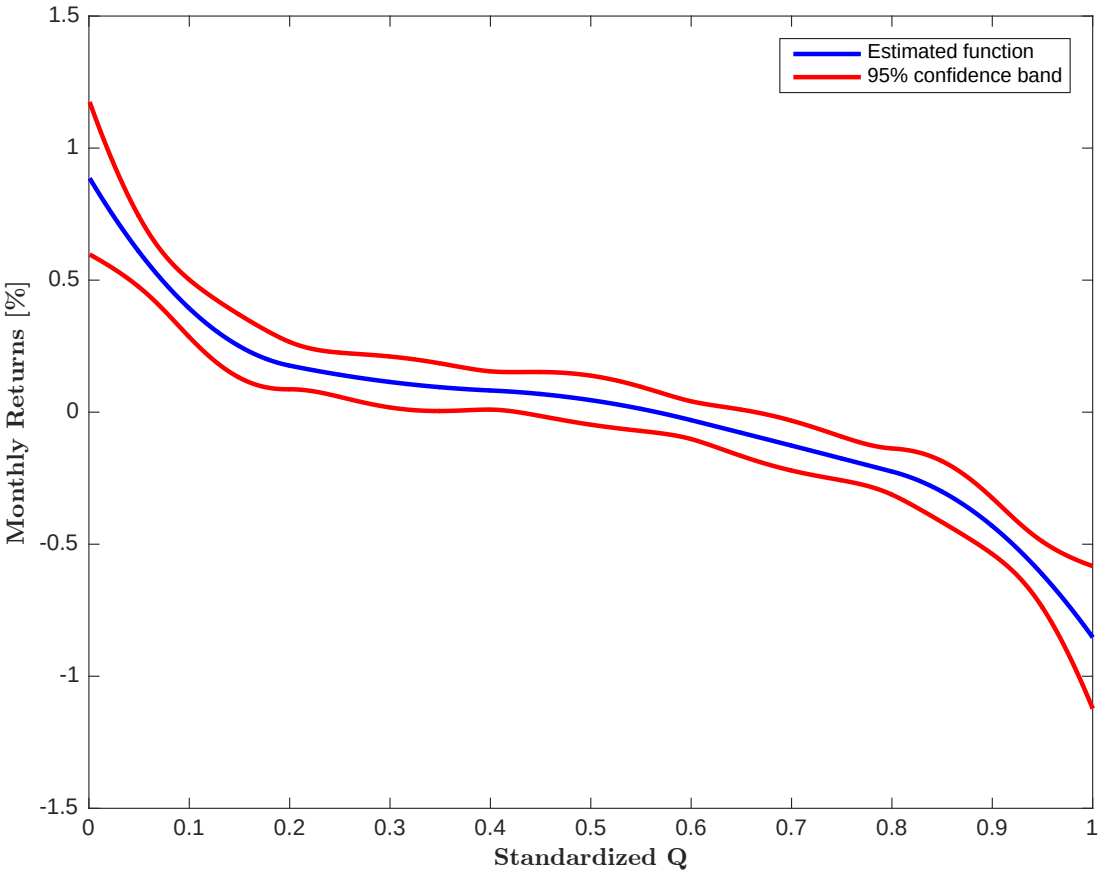
This figure plots expected returns on the y-axis against the book-to-market ratio on the x-axis as well as portfolio mean returns and a non-parametric conditional mean function for simulated data.

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



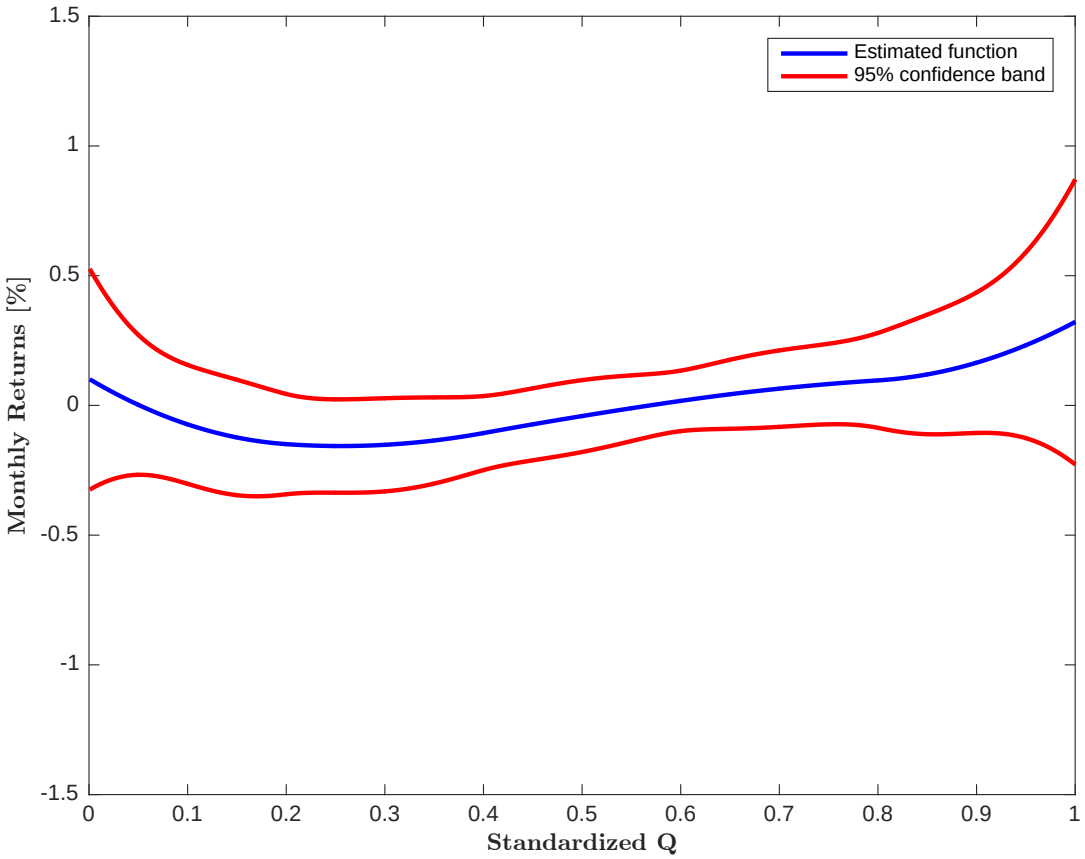
This figure plots expected returns on the y-axis against the book-to-market ratio on the x-axis as well as portfolio mean returns and a non-parametric conditional mean function for simulated data.

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



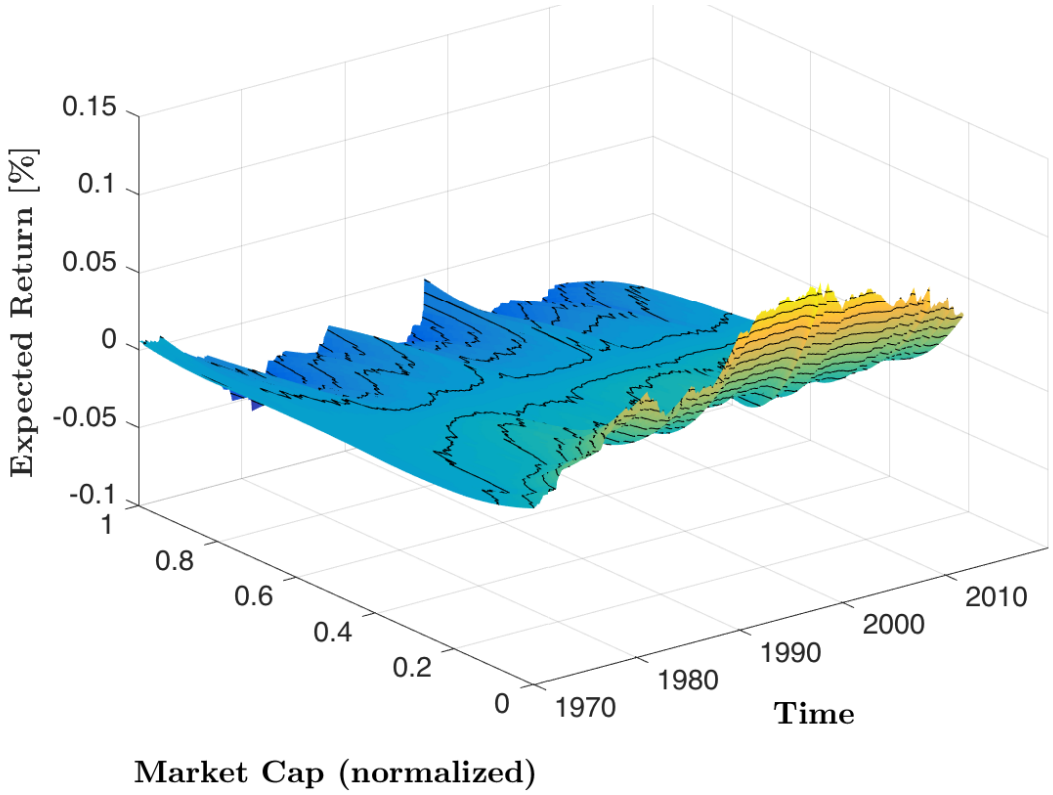
Effect of Tobin's Q on average returns.

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



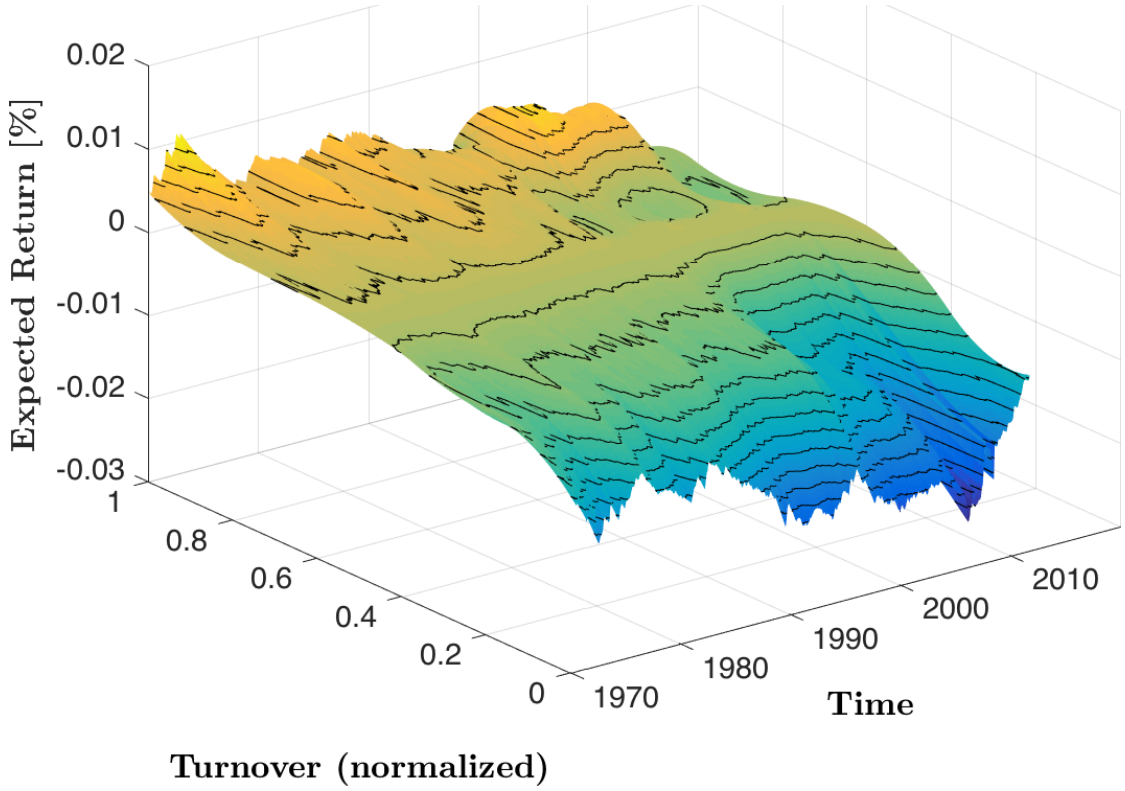
Effect of Tobin's Q on average returns (considering all the other selected characteristics).

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



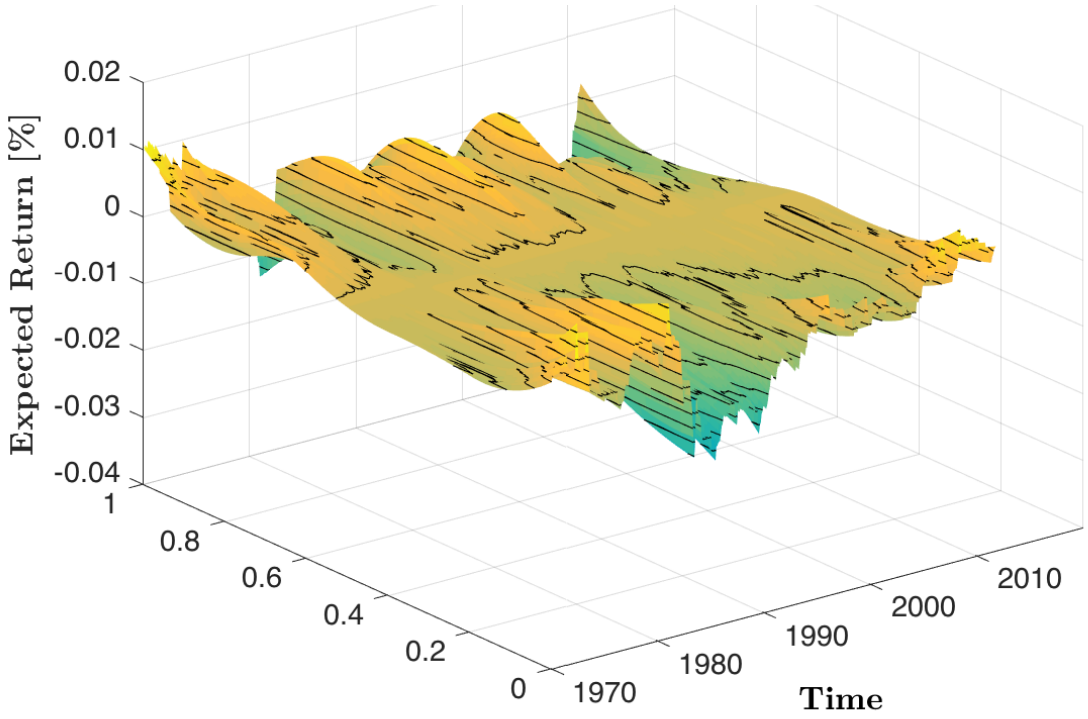
Effect of lagged market cap on average returns (conditional all other selected characteristics).

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



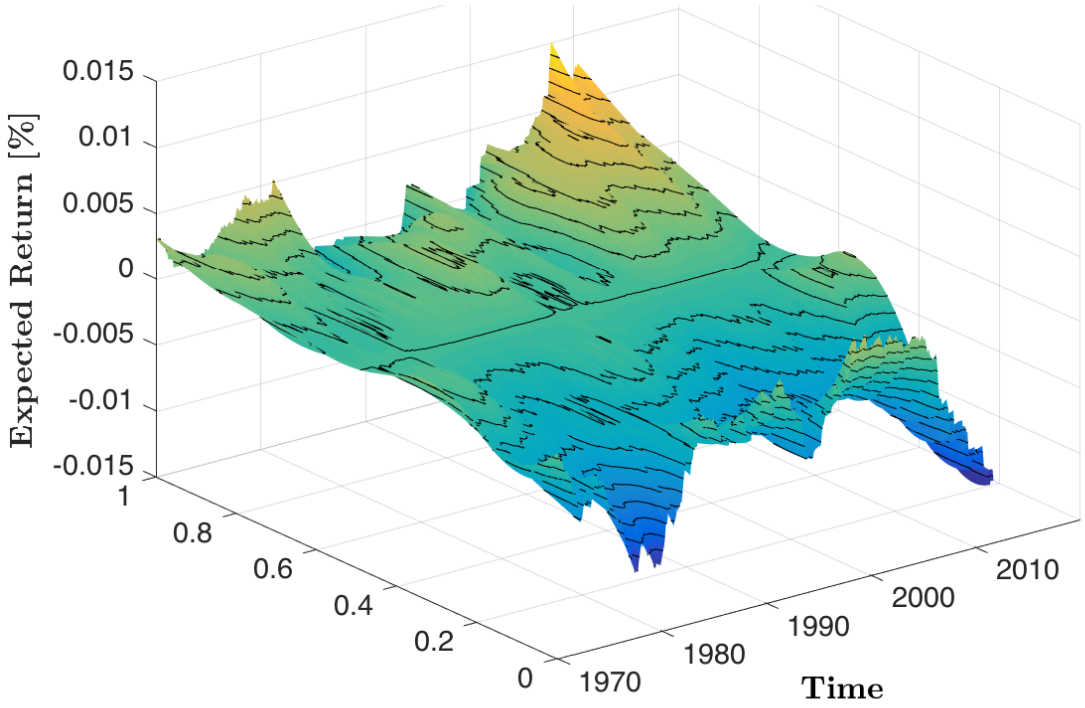
Effect of lagged turnover on average returns (conditional all other selected characteristics).

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



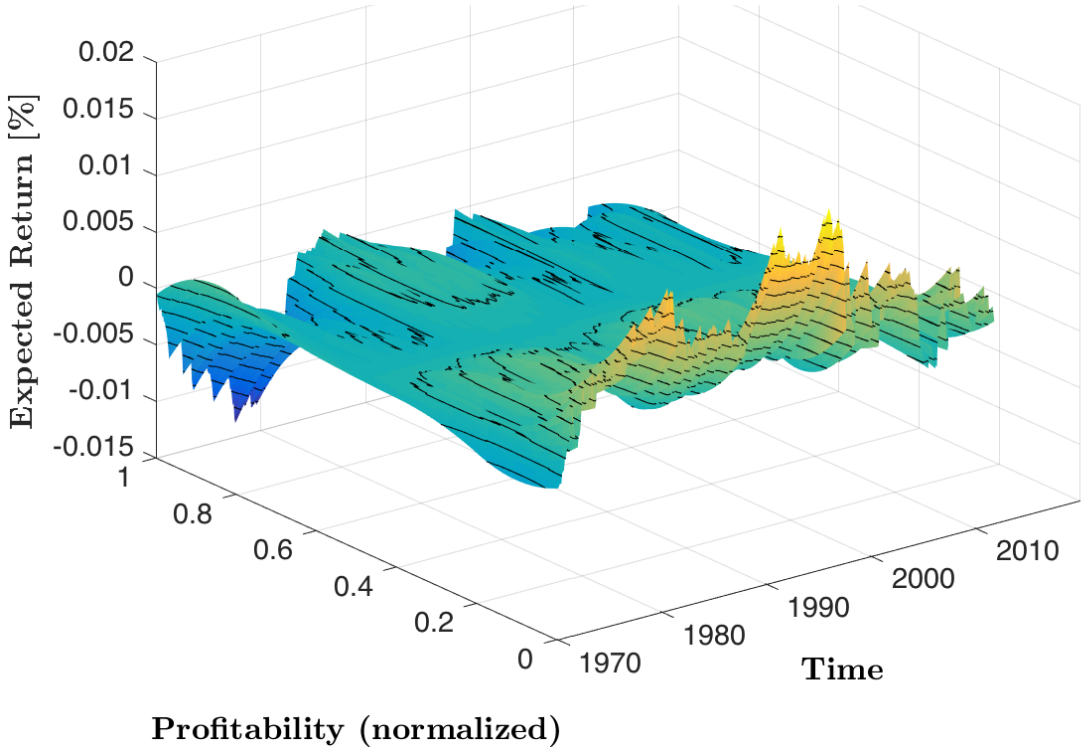
Effect of book-to-market ratio on average returns (conditional on all other selected characteristics).

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



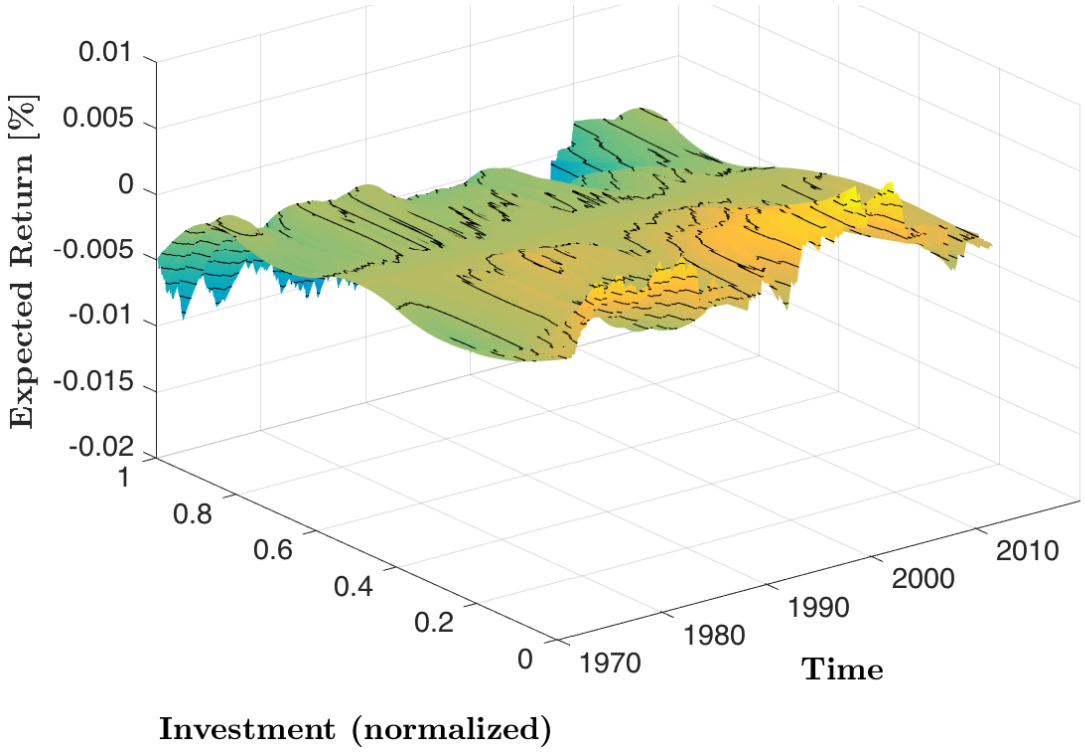
Effect of depreciation-to-assets on average returns (conditional on all other selected characteristics).

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



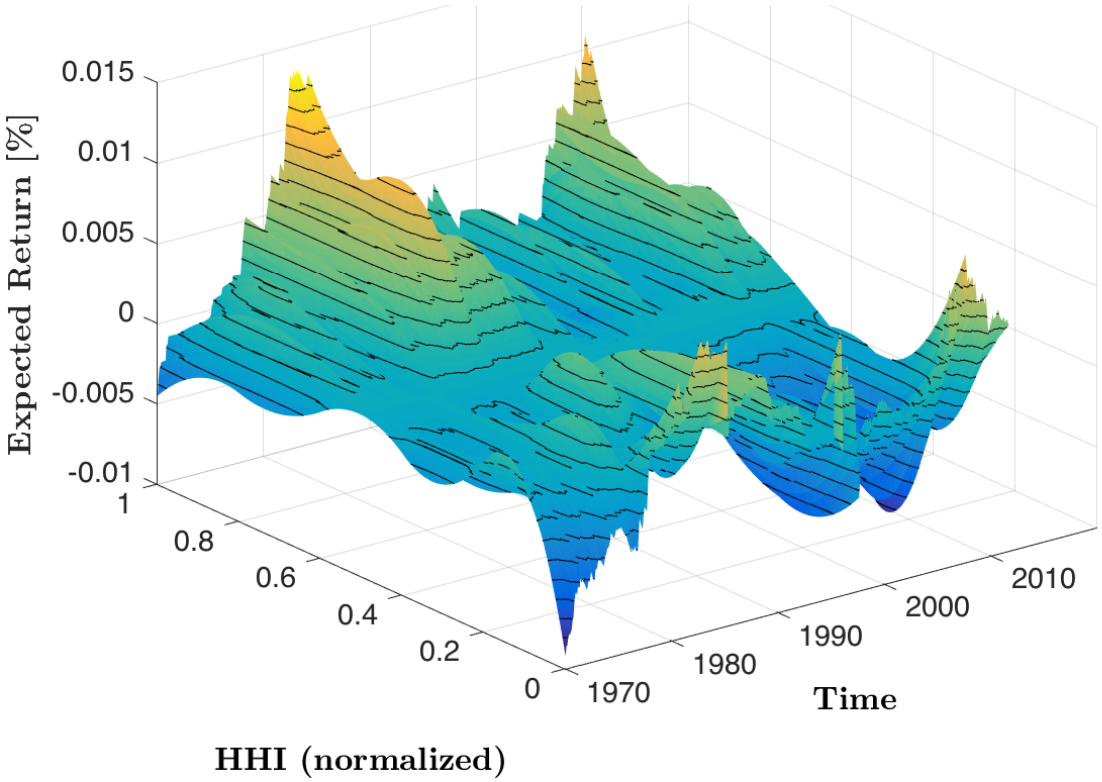
Effect of profitability on average returns (conditional on all other selected characteristics).

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



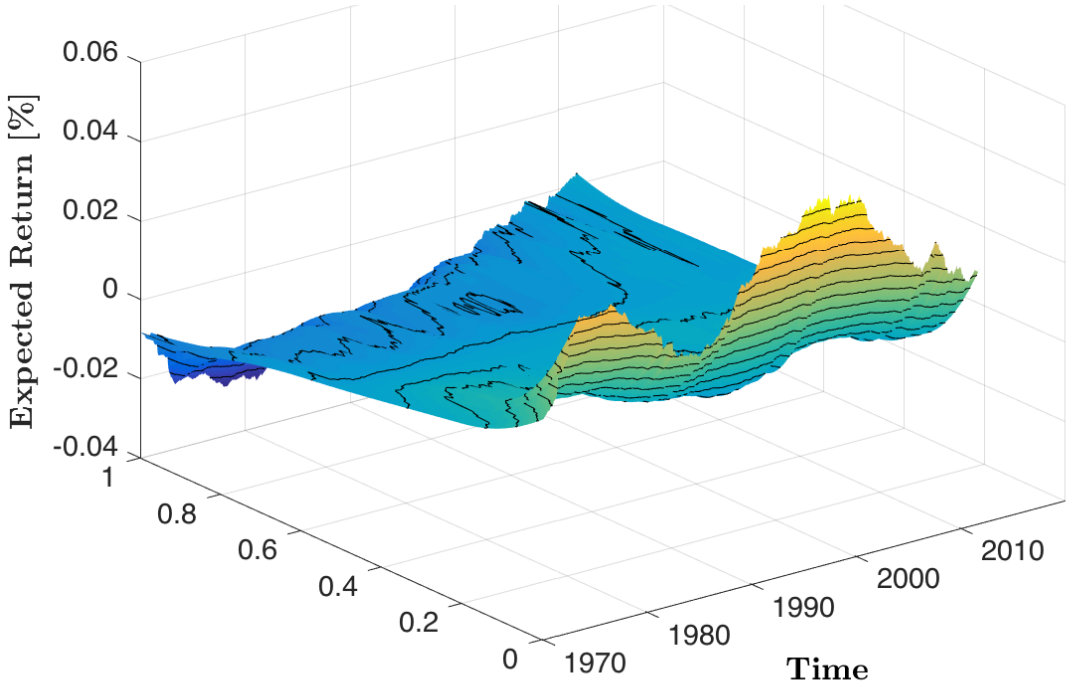
Effect of investment on average returns (conditional on all other selected characteristics).

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



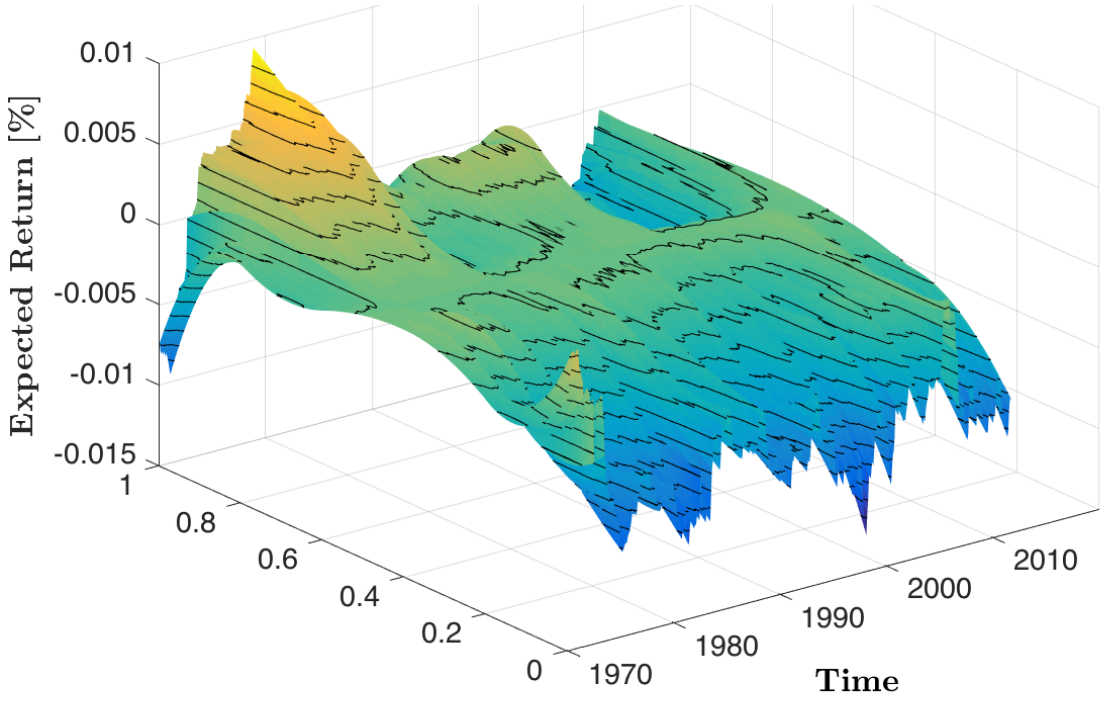
Effect of Herfindahl-Hirschman index on average returns (conditional on all other selected characteristics).

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



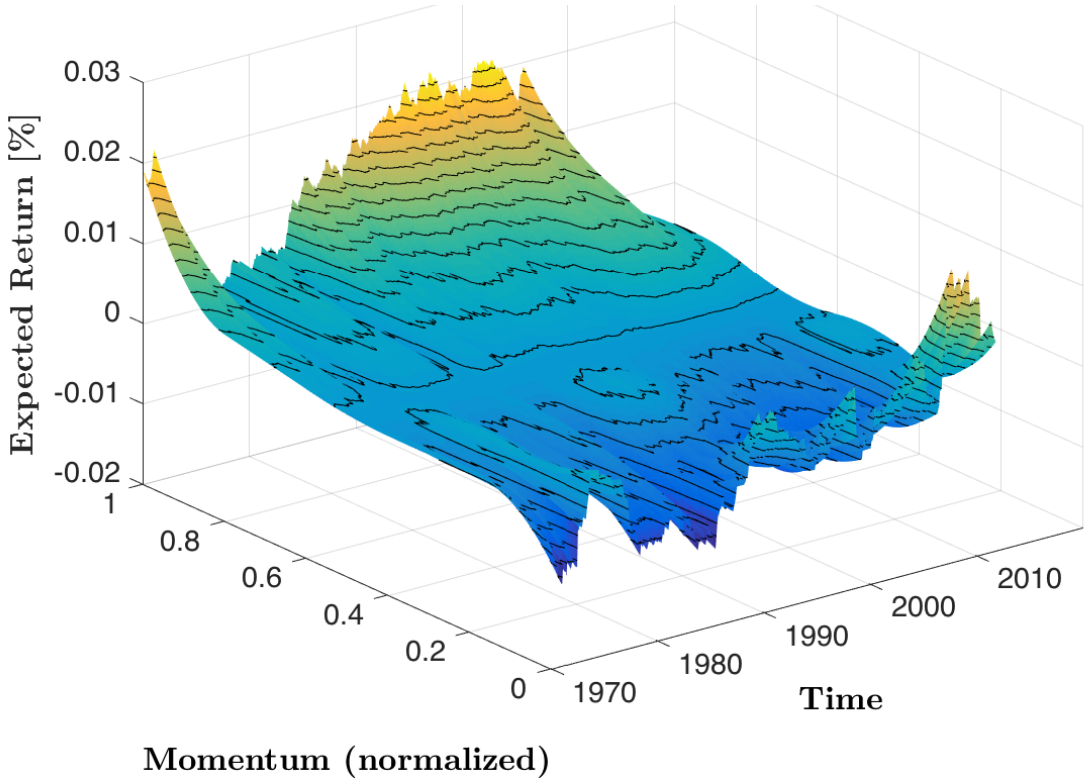
Effect of short-term reversal on average returns (conditional on all other selected characteristics).

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



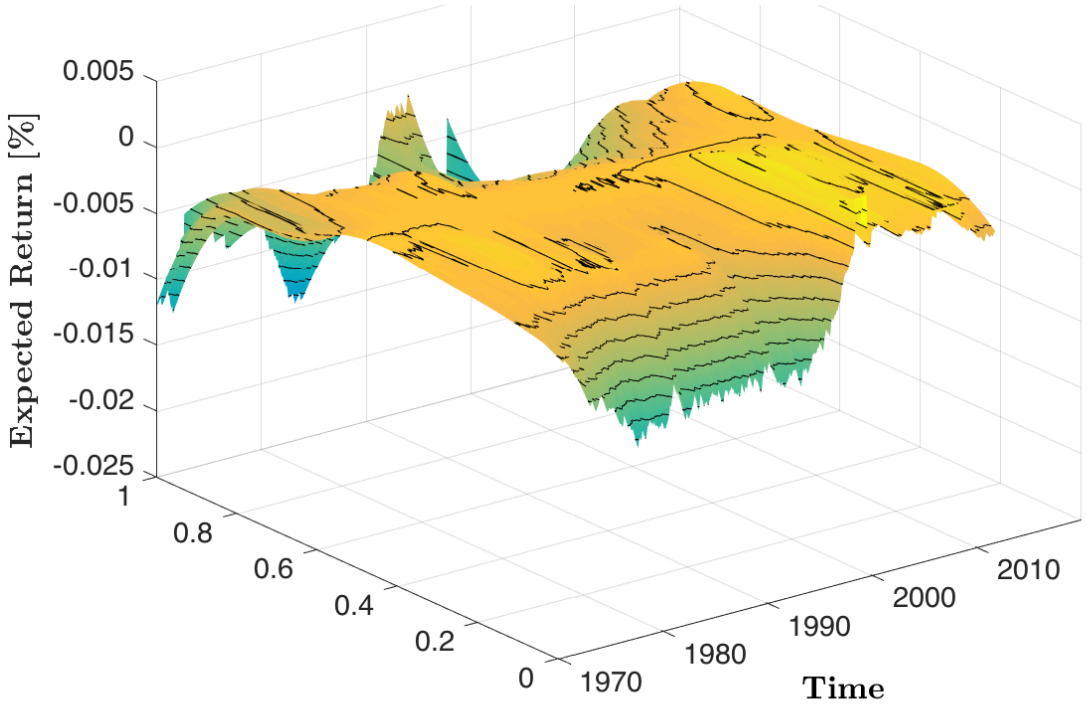
Effect of intermediate momentum on average returns (conditional on all other selected characteristics).

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



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

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



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

Table 1: **Descriptive Statistcs 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 ten 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	16.89	17.41	17.01	16.40	16.44	17.17	15.09	15.54	14.07	-2.88
	(14.61)	(16.34)	(17.25)	(18.12)	(19.11)	(20.70)	(22.89)	(24.93)	(28.49)	(34.93)	(25.65)
ME_{t-1}	33.06	16.34	14.95	14.61	14.74	15.47	14.70	14.01	13.30	11.73	-21.33
	(29.82)	(24.63)	(23.23)	(23.23)	(22.95)	(22.07)	(21.44)	(20.43)	(18.86)	(16.46)	(25.03)
Book-to-Market	9.20	12.09	13.20	14.84	15.80	17.03	18.06	18.83	20.09	23.81	14.61
	(24.15)	(22.41)	(21.15)	(21.07)	(20.26)	(20.39)	(20.40)	(20.85)	(21.46)	(24.20)	(16.02)
Profitability	14.99	14.22	15.39	15.43	16.17	16.21	16.56	17.43	18.33	18.22	3.23
	(25.34)	(20.54)	(20.18)	(20.20)	(20.61)	(20.50)	(20.76)	(21.45)	(21.64)	(23.01)	(12.53)
Investment	22.44	20.78	19.02	17.73	16.52	16.19	15.49	14.17	12.62	7.98	-14.46
	(28.79)	(22.57)	(20.21)	(18.66)	(18.52)	(18.77)	(19.37)	(20.35)	(22.38)	(25.22)	(12.85)
Idio Vol	12.85	14.96	15.92	16.85	17.75	17.46	18.12	16.37	15.88	16.80	3.95
	(13.61)	(16.12)	(17.51)	(19.05)	(20.36)	(22.13)	(23.90)	(25.30)	(28.21)	(32.85)	(25.89)
r_{12-2}	14.77	12.70	13.24	13.54	14.62	15.14	16.57	18.71	20.12	23.52	8.75
	(33.49)	(24.91)	(22.23)	(19.98)	(19.29)	(18.44)	(18.48)	(19.01)	(20.78)	(25.04)	(25.14)
r_{12-7}	13.30	12.90	13.39	15.32	16.22	16.03	16.80	17.81	19.71	21.45	8.15
	(29.42)	(23.32)	(21.05)	(19.88)	(19.10)	(18.99)	(19.06)	(19.77)	(21.35)	(24.95)	(17.19)
r_{1-0}	36.62	21.40	18.78	16.84	15.69	14.70	12.86	11.72	9.97	4.07	-32.55
	(32.12)	(24.44)	(21.95)	(20.61)	(19.47)	(18.70)	(18.44)	(18.78)	(20.03)	(23.86)	(21.69)
r_{36-13}	23.75	19.54	18.84	15.96	16.04	15.44	15.13	13.92	13.53	10.78	-12.97
	(31.28)	(24.48)	(21.97)	(19.76)	(19.19)	(18.44)	(18.68)	(19.12)	(20.69)	(24.68)	(20.44)
Cash	15.52	15.13	15.33	15.41	16.70	16.47	17.33	17.16	16.30	17.60	2.07
	(20.59)	(20.64)	(20.01)	(20.58)	(21.14)	(21.11)	(21.64)	(22.07)	(23.34)	(24.78)	(14.44)
Total Assets	21.62	19.25	17.65	16.09	15.75	15.74	15.70	14.51	13.95	12.67	-8.95
	(27.43)	(25.15)	(23.77)	(22.73)	(21.60)	(21.09)	(20.60)	(20.40)	(19.61)	(17.57)	(21.57)
D&A	12.77	12.68	15.09	15.61	16.65	17.03	16.67	18.65	18.62	19.17	6.41
	(22.73)	(21.43)	(20.88)	(21.17)	(20.72)	(21.04)	(20.29)	(20.44)	(21.19)	(23.81)	(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	15.34	15.28	16.90	14.59	15.66	15.94	17.32	18.55	17.25	1.16
	(19.95)	(19.92)	(20.43)	(20.53)	(20.00)	(20.50)	(20.84)	(22.15)	(24.40)	(29.11)	(19.51)
Free CF	15.18	15.95	16.82	16.48	16.19	16.85	15.91	16.12	16.53	16.91	1.73
	(29.79)	(25.06)	(22.27)	(21.12)	(20.31)	(19.41)	(18.89)	(18.60)	(18.92)	(20.19)	(15.28)
Capex	18.85	18.31	17.14	16.90	17.09	16.75	16.39	14.74	13.94	12.85	-6.00
	(23.01)	(21.88)	(21.20)	(20.88)	(20.84)	(20.75)	(20.74)	(20.71)	(21.27)	(23.07)	(12.43)
$Turnover_{t-1}$	12.48	14.89	15.66	16.35	17.95	17.74	17.71	17.08	18.06	15.04	2.56
	(15.69)	(17.96)	(19.16)	(19.86)	(20.84)	(21.45)	(22.35)	(24.15)	(26.34)	(29.89)	(21.95)
Oper Accruals	17.15	18.53	18.21	18.13	17.39	17.09	16.14	15.62	14.42	10.28	-6.87
	(25.65)	(21.47)	(20.07)	(19.61)	(19.70)	(19.77)	(20.07)	(20.79)	(21.30)	(24.36)	(8.59)
Oper Leverage	12.51	13.44	15.25	16.33	15.84	17.18	16.93	18.29	18.24	18.96	6.46
	(20.42)	(21.50)	(21.42)	(21.72)	(21.56)	(21.46)	(21.38)	(21.99)	(22.15)	(21.36)	(12.10)
P-C-M	18.14	15.84	16.00	16.17	16.27	15.97	15.86	15.84	16.25	16.62	-1.52
	(25.74)	(21.59)	(20.73)	(21.36)	(20.96)	(20.52)	(20.76)	(20.58)	(20.85)	(21.45)	(10.51)
Tobin's Q	23.20	19.63	18.62	17.81	16.86	16.28	14.94	14.34	12.11	9.15	-14.05
	(22.46)	(21.47)	(21.06)	(21.03)	(21.14)	(20.86)	(21.07)	(21.53)	(21.99)	(24.14)	(14.99)
ROE	17.76	19.08	18.42	17.69	16.59	15.31	15.15	15.23	14.54	13.16	-4.60
	(32.18)	(26.37)	(22.08)	(20.30)	(18.93)	(18.50)	(18.33)	(19.03)	(19.74)	(21.83)	(18.76)
bid-ask Spread	13.89	15.31	15.41	15.87	15.79	16.24	16.03	15.50	16.36	22.55	8.66
	(18.50)	(18.52)	(19.05)	(20.14)	(21.01)	(22.28)	(22.99)	(23.98)	(25.28)	(29.14)	(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	All Stocks Full Sample 10 Knots	Large Stocks Full Sample 10 Knots	All Stocks Early Sample 5 Knots	All Stocks Late Sample 5 Knots
	(1)	(2)	(3)	(4)	(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	$Idiosyncratic\ Vol$				