Dissecting Characteristics Nonparametrically*

Joachim Freyberger† Andreas Neuhierl‡ Michael Weber§

This version: January 2019

Abstract

We propose a nonparametric method to study which characteristics provide incremental information for the cross-section of expected returns. We use the adaptive group LASSO to select characteristics and to estimate how they affect expected returns nonparametrically. Our method can handle a large number of characteristics, allows for a flexible functional form, and our implementation is insensitive to outliers. Many of the previously identified return predictors don't provide incremental information for expected returns, and nonlinearities are important. We study the properties of our method in simulations and find large improvements both in model selection and prediction compared to alternative selection methods.

JEL classification: C14, C52, C58, G12

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

*We thank Alessandro Beber, Jonathan Berk, Oleg Bondarenko, Svetlana Bryzgalova, John Campbell, Jason Chen, Josh Coval, Kent Daniel, Victor DeMiguel, Andres Donangelo, Gene Fama, Ken French, Erwin Hansen, Lars Hansen, Benjamin Holcblat, Andrew Karolyi, Bryan Kelly, Leonid Kogan, Shimon Kogan, Jon Lewellen, Yingying Li, Binying Liu, Bill McDonald, Stefan Nagel, Stavros Panageas, Ľuboš Pástor, Seth Pruitt, Alberto Rossi, Shrihari Santosh, Olivier Scaillet, Andrei Shleifer, George Skoulakis, Raman Uppal, Adrien Verdelhan, Yan Xu, Amir Yaron and conference and seminar participants at Chinese University of Hong Kong, City University of Hong Kong, Cornell, Dartmouth College, Finance UC 13th International Conference, FRA Conference 2016, HEC Montreal, HKUST, HKU, 2017 Imperial Hedge Fund Conference, London Business School, 2017 Luxembourg Asset Management Summit, McGill, New Methods for the Cross Section of Returns Conference, NUS, NTU, 2017 Revelstoke Finance Conference, Santiago Finance Workshop, Schroders, 2017 SFS Cavalcade, SMU, Stockholm School of Economics, TAU Finance Conference 2016, Tsinghua University PBCSF, Tsinghua University SEM, the 2017 Texas Finance Festival, the University of Chicago, Université de Genève, University of Frankfurt, the University of Illinois at Chicago, the University of Notre Dame, and the University of Washington for valuable comments. We thank Xiao Yin for providing excellent research assistance. Weber gratefully acknowledges financial support from the Fama-Miller Center and the Fama Research Fund at the University of Chicago Booth School of Business.

†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, the University of Chicago, Chicago, IL, USA and NBER. e-Mail: michael.weber@chicagobooth.edu.

I Introduction

In his presidential address, Cochrane (2011) argues the cross section of the expected return "is once again descending into chaos." Harvey et al. (2016) identify "hundreds of papers and factors" that have predictive power for the cross section of expected returns. Many economic models, such as the consumption CAPM of Lucas (1978), Breeden (1979), and Rubinstein (1976), instead predict that only a small number of state variables suffice to 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?" Portfolio sorts are subject to the curse of dimensionality when the number of characteristics is large, and linear regressions make strong functional-form assumptions and are sensitive to outliers.[1] In addition, in many empirical settings, most of the variation in characteristic values and returns are in the extremes of the characteristic distribution and the association between characteristics and returns appears nonlinear (see Fama and French (2008)). 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 method to determine which firm characteristics provide incremental 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 developed 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,

[1] We discuss these, and related concerns in Section A.2 and compare current methods with our proposed framework in Section A.3 of the Online Appendix.

given the other characteristics. Nonparametric estimation deals with estimating the effect of important characteristics on expected returns without imposing a strong functional form.

We show three applications of our proposed framework. First, we study which characteristics provide incremental information for the cross section of expected returns. We estimate our model on 62 characteristics including size, book-to-market, beta, and other prominent variables and anomalies on a sample period from July 1965 to June 2014. Only 13 variables, including size, total volatility, and past return-based predictors, have incremental explanatory power for expected returns for the full sample period and all stocks. A hedge portfolio going long stocks with high predicted returns and shorting stocks with low predicted returns has an in-sample Sharpe ratio of more than 3. Only 11 characteristics have predictive power for returns in the first half of our sample. In the second half, instead, we find 14 characteristics are associated with cross-sectional return premia. For stocks whose market capitalization is above the 20% NYSE size percentile, only nine characteristics, including changes in shares outstanding, past returns, and standardized unexplained volume, remain incremental return predictors. The in-sample Sharpe ratio is still 2.25 for large stocks.

Second, we compare the out-of-sample performance of the nonparametric model with a linear model. Estimating flexible functional forms raises the concern of in-sample overfitting. We estimate both a linear and the nonparametric model over a period until 1990 and select return predictors. We then use 10 years of data to estimate the models on the selected characteristics. In the first month after the end of our estimation period, we take the selected characteristics, predict one-month-ahead returns, and construct a hedge portfolio similar to our in-sample exercise. We roll the estimation and prediction period forward by one month and repeat the procedure until the end of the sample.

Specifically, we perform model selection once until December 1990 for both the linear model and the nonparametric model. Our first estimation period is from December of 1981 until November of 1990, and the first out-of-sample prediction is for January 1991

using characteristics from December 1990.[2] We then move the estimation and prediction period forward by one month. The nonparametric model generates an out-of-sample Sharpe ratio of 2.75 compared to 1.06 for the linear model.[3] When we adjust Sharpe ratios for transaction costs, we find the nonlinear model still compares favorably to the linear model with Sharpe ratios of 1.56 and 0.29, respectively. The characteristics we study are not a random sample, but have been associated with cross sectional return premia in the past. Therefore, we focus mainly on the comparison across models rather than emphasizing the overall magnitude of the Sharpe ratios.

The linear model selects 30 characteristics in-sample compared to only eleven for the nonparametric model, but performs worse out-of-sample and nonlinearities are important. We find an increase in out-of-sample Sharpe ratios relative to the Sharpe ratio of the linear model when we employ the nonparametric model for prediction but use the 30 characteristics the linear model selects. The linear model appears to overfit the data in-sample. We find an identical Sharpe ratio for the linear model when we use the 11 characteristics selected by the nonparametric model, as we do with the 30 characteristics selected by the linear model. These results underscore once more the importance of nonlinearities. With the same set of 11 characteristics the nonlinear model selects, we find the nonparametric model has a Sharpe ratio that is larger by a factor of 2.5 relative to the Sharpe ratio of the linear model using the same set of characteristics.

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. Size conditional on the

[2]We merge balance-sheet variables to returns following the Fama and French (1993) convention of requiring a lag of at least six months, and our results are therefore indeed out-of-sample.

[3]The linear model we estimate and the results for the linear model are similar to Lewellen (2015).

other selected return predictors, instead, has a significant predictive power for expected returns throughout our sample period similar to the findings in Asness, Frazzini, Israel, Moskowitz, and Pedersen (2017).

The method we propose has several "tuning" parameters and one might be concerned that our conclusions depends on some of the choices we have to make. We document in an extensive simulation study both aspects of our proposed method: model selection and and return prediction. Across a wide array of choices regarding the tuning parameters, we find the adaptive group LASSO performs well along both dimensions, that is, it has a high probability to select the "right" set of characteristics and performs well in predicting returns out of sample. We also compare the performance of the nonlinear adaptive group LASSO for model selection and return prediction to a linear LASSO and popular recent proposals like increased thresholds for t-statistics or p-value adjustments for false-discovery rates. We find along both dimensions that allowing for nonlinearities improves performance substantially.

The paper provides a new method in empirical asset pricing to understand which of the previously published firm characteristics provide information for expected returns conditional on other characteristics. We see this exercise as a natural first step in the "multidimensional challenge." Once we understand which characteristics indeed provide incremental information, we can aim to relate characteristics to factor exposures, estimate factors and stochastic discount factors directly, or relate characteristics and factors to economic models.

A Related Literature

The capital asset pricing model (CAPM) of Sharpe (1964), Lintner (1965), and Mossin (1966) predicts that an asset's beta with respect to the market portfolio is a sufficient statistic for the cross section of expected returns. Subsequently, researchers identified many variables that contain additional independent information for expected 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 factor, and a value factor can explain cross sections of stocks sorted on characteristics that appeared anomalous relative to the CAPM. In this sense, Fama and French (1993) 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 20 years following the publication of Fama and French (1992), many researchers joined a "fishing expedition" to identify characteristics and factor exposures that the three-factor model cannot explain. Harvey, Liu, and Zhu (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. 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 the predictive power of 15 characteristics and finds that only few are significant predictors for the cross section of expected returns. Green, Hand, and Zhang (2017) adjust Fama-MacBeth regressions to avoid overweighting microcaps, adjust p-values for a data snooping bias and find for a sample starting in 1980 that many return predictors do not provide independent information. Although Fama-MacBeth regressions carry a lot of intuition, they do not offer a formal model selection method. We build on Lewellen (2015) and provide a framework that allows for nonlinear associations between characteristics and returns, provide a formal framework to disentangle important from unimportant return predictors, and study many more characteristics.

We also build on a large literature in economics and statistics using penalized regressions. Horowitz (2016) gives a general overview of model selection in high-dimensional models, and Huang, Horowitz, and Wei (2010) discuss variable selection in a nonparametric additive model similar to the one we implement empirically. Recent applications of LASSO methods in finance are Huang and Shi (2016), who use an adaptive group LASSO in a linear framework and construct macro factors to test for determinants of bond risk premia. Chinco, Clark-Joseph, and Ye (2018) use a linear model for high-frequency return predictability using past returns of related stocks, and find their method increases predictability relative to OLS. Goto and Xu (2015) use a LASSO to obtain a sparse estimator of the inverse covariance matrix for mean variance portfolio optimization. And Chinco, Neuhierl, and Weber (Chinco et al.) invert the best-fit tuning parameter in penalized regressions to estimate the anomaly baserate.

Gagliardini, Ossola, and Scaillet (2016) develop a weighted two-pass cross-sectional regression method to estimate risk premia from an unbalanced panel of individual stocks. Giglio and Xiu (2016) instead propose a three-pass regression method that combines principal component analysis and a two-stage regression framework to estimate consistent factor risk premia in the presence of omitted factors when the cross section of test assets is large. DeMiguel, Martin-Utrera, Nogales, and Uppal (2016) extend the parametric portfolio approach of Brandt et al. (2009) to study which characteristics provide valuable information for portfolio optimization. Kelly, Pruitt, and Su (2017) generalize standard PCA to allow for time-varying loadings and extract common factors from the universe of individual stocks. Kim, Korajczyk, and Neuhierl (2018) develop a new method to estimate arbitrage portfolios by utilizing information contained in firm characteristics for both abnormal returns and factor loadings. Kozak, Nagel, and Santosh (2017) exploit economic restrictions relating expected returns to covariances to construct stochastic discount factors. Lettau and Pelger (2018) generalizes PCA by including a penalty on the pricing error in expected returns that allows them to identify "weak" factors with high Sharpe ratios.

We, instead, are mainly concerned with formal model selection, that is, which characteristics provide incremental information in the presence of other characteristics.

II Current Methods and Nonparametric Models

A Expected Returns and Current Methods

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

$$m_t(c_1, \ldots, c_S) = E[R_{it} \mid C_{1,it-1} = c_1, \ldots, C_{S,it-1} = c_S], \tag{1}$$

where $C_{1,it-1}, \ldots, C_{S,it-1}$ are S firm characteristics.

We often use portfolio sorts to approximate m_t for a single characteristic. 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. They suffer from the curse of dimensionality, they do not offer formal guidance to discriminate between characteristics, and they assume returns do not vary within portfolio.

An alternative to portfolio sorts is to *assume* linearity of m_t and run linear panel regressions of excess returns on characteristics, namely,

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

Linear regressions allow us to study the predictive power for expected returns of many characteristics jointly, but they also have potential pitfalls. Most importantly, no a priori reason exists why the conditional mean function should be linear.

We discuss many of these shortcomings in more detail in Section A.2 of the online appendix and how researchers typically address some of the shortcomings. Cochrane (2011) synthesizes many of the challenges that portfolio sorts and linear regressions face in the context of many return predictors, and suspects "we will have to use different methods."

B Nonparametric Estimation

Cochrane (2011) conjectures in his presidential address, "[P]ortfolio sorts are really the same thing as nonparametric cross-sectional regressions, using nonoverlapping histogram weights." We establish a formal equivalence result between portfolio sorts and regressions in the online appendix. Specifically, suppose we have a single characteristic $C_{1,it-1}$ and we sort stocks into L portfolios depending on the value of the characteristic. We show in the appendix a one-to-one relationship exists between the portfolio returns and regression coefficients in a regression of returns on L indicator functions, where indicator function l is equal to 1 if stock i is in portfolio l for $l = 1, \ldots, L$. Hence, portfolio sorts are equivalent to approximating the conditional mean function with a step function. The nonparametric econometrics literature also refers to these functions as constant splines. We discuss in the online appendix how our proposed framework is a natural generalization of portfolio sorts. Specifically, we use a smooth extension of this estimation strategy with many possible regressors.

Estimating the conditional mean function, m_t, fully nonparametrically with many regressors results in a slow rate of convergence and imprecise estimates in practice.[4] Specifically, the optimal rate of convergence decreases as the number of characteristics increases. Consequently, we get an estimator with poor finite sample properties if the number of characteristics is large.[5] Nevertheless, if we are interested in which characteristics provide incremental information for expected returns given other

[4]The literature refers to this phenomenon as the "curse of dimensionality" (see Stone (1982) for a formal treatment).

[5]Section A.4 of the online appendix contains some concrete examples.

characteristics, we cannot look at each characteristic in isolation. A natural solution in the nonparametric regression framework is to assume an additive model, that is,

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

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

An important restriction of any additive model, including multivariate linear models or Fama-MacBeth regressions, is

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

for all $s \neq s'$; therefore, the additive model does not allow for cross dependencies between characteristics. For example, the predictive power of the book-to-market ratio for expected returns does not vary with firm size (conditional on size). 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. Brandt et al. (2009) make a similar assumption, but also stress that we can always interpret characteristics as the cross product of a more basic set of characteristics. In our empirical application, we show results for all stocks and all-but micro caps, but also show results when we interact each characteristic with size.

Although the assumption of an additive model is somewhat restrictive, it provides desirable econometric advantages. In addition, this assumption is far less restrictive than assuming additivity and linearity, 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 subsection C

below. In addition, the additive structure allows us to extrapolate to regions with very few data points on the joint support of characteristics from regions with more data, because we can average over the marginal distributions of characteristics. Without any additional structure, we would get very imprecise estimates when the joint distribution of characteristics is sparse.

Before providing the formal model selection procedure, we describe a normalization of the characteristics that will allow us to map our nonparametric estimator directly to portfolio sorts and ensures our results are insensitive to outliers. For each characteristic s, let $\tilde{C}_{s,it-1}$ be the rank transformation of $C_{s,it-1}$, which maps the cross-sectional distribution of the characteristic to the unit interval; that is, $\tilde{C}_{s,it-1} \in [0,1]$. It is easy to show a function $\tilde{m}_t$ exists, such that

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

Hence, knowledge of the conditional mean function m_t is equivalent to knowing the transformed conditional mean function $\tilde{m}_t$, which is the function we estimate.[6] Similar to portfolio sorting, we are typically not interested in the actual value of a characteristic in isolation, but rather in 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 it is not. Our normalization considers the relative size in the cross section rather than the absolute size, similar to portfolio sorting.

C Adaptive Group LASSO

We use a group LASSO procedure developed 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

[6]We show in Section A.5 of the online appendix that the general econometric theory we discuss in subsection *C* (model selection, consistency, etc.) also applies to any other monotonic transformation or the non-transformed conditional mean function.

function of characteristics; that is,

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

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

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 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 not constant.[7]

We can interpret portfolio sorts as estimating $\tilde{m}_{ts}$ by a constant within each portfolio. We also partition the support of each characteristic into L intervals similar to portfolio sorts. The endpoints of the intervals are *knots* and we set them to the quantiles of the rank transformed characteristic distribution. We then approximate each function $\tilde{m}_{ts}$ by a quadratic function on each of the intervals, such that the whole function is continuously differentiable on $[0, 1]$, that is, we approximate $\tilde{m}_{ts}$ by *quadratic splines*. We use these splines for our baseline results because these are the lowest-order splines such that $\tilde{m}_{ts}$ is continuously differentiable. Thus, we can interpret our estimator as a smooth extension of portfolio sorts. Interestingly, we can then approximate $\tilde{m}_{ts}$ as a linear combination of $L + 2$ basis function, i.e.,

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

where $p_k(c)$ are known functions and β_{tsk} are parameters we estimate. We provide a formal definition of splines and the corresponding basis functions in Section A.2 of the online appendix. The number of intervals L is a user-specified smoothing parameter, analogous to the number of portfolios. As L increases, the precision of the approximation increases,

[7]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).

but so does the number of parameters we have to estimate and hence the variance. We discuss these and other choices we have to make and the robustness of our empirical results in an extensive simulation study in Section IV.

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

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

where $\tilde{\boldsymbol{\beta}}_t$ is an $(L+2) \times S$ vector of 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 group 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 incremental information for expected returns. Due to the penalty, the LASSO is applicable even when the number of characteristics is larger than the sample size. Yuan and Lin (2006) propose to choose λ_1 in a data-dependent way to minimize Bayesian Information Criterion (BIC) which we follow in our application.

However, the first step of the LASSO may select 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 introduces characteristic-specific weights in the LASSO group penalty function as a function of first-step estimates to address this problem. The online appendix discusses in Section A.3 the second step, the consistency conditions, and the efficiency properties of the resulting estimates in detail.

If the cross section is sufficiently large, we could perform model selection and estimation 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 over

time. McLean and Pontiff (2016) show that for 97 return predictors, predictability decreases by 58% post publication. However, if the conditional mean function was 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 III, we estimate our model over the whole sample in the baseline but also over subsamples and estimate rolling specifications to investigate the variation in the conditional mean function over time.

D 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 S constants such that $\sum_{s=1}^{S} \alpha_s = 0$. Then,

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

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.

We report estimates of the functions using the common normalization that the functions integrate to 0, which is identified.

Section A.6 of the online appendix discusses how we construct confidence bands for the figures which we report and how we select the number of interpolation points in the empirical application of Section III below.

III Empirical Application

We now discuss the universe of characteristics we use in our empirical application and study which of the 62 characteristics provide incremental information for expected returns,

using the adaptive group LASSO for selection and estimation.

A 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.

Balance-sheet data are from the Standard and Poor's Compustat database. We use balance-sheet data from the fiscal year ending in calendar year $t-1$ for estimation starting in June of year t until May of year $t+1$ predicting returns from July of year t until June of year $t+1$.

Table 1 provides an overview of the 62 characteristics we apply our method to. We group them into six categories: past return based predictors such as momentum ($\mathbf{r_{12-2}}$) and short-term reversal ($\mathbf{r_{2-1}}$), investment-related characteristics such as the annual percentage change in total assets (**Investment**) or the change in inventory over total assets (**IVC**), profitability-related characteristics such as gross profitability over the book-value of equity (**Prof**) or return on operating assets (**ROA**), intangibles such as operating accruals (**OA**) and tangibility (**Tan**), value-related characteristics such as the book-to-market ratio (**BEME**) and earnings-to-price (**E2P**), and trading frictions such as the average daily bid-ask spread (**Spread**) and standard unexplained volume (**SUV**). We follow Hou, Xue, and Zhang (2015) in the classification of characteristics.

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 1965 until June 2014. Table 2 reports summary statistics for various firm characteristics and return predictors. We calculate all statistics annually and then average over time. We have 1.6 million observations in our baseline analysis.

Section A.1 in the online appendix contains a detailed description of the characteristics, the construction, and the relevant references.

B Selected Characteristics and Their Influence

The purpose of this section is to show different applications of the adaptive group LASSO. We do not aim to exhaust all possible combinations of characteristics, sample periods, and firm sizes, or all possible applications but rather aim to provide some insights into the flexibility of the method in actual data. Section IV contains an extensive simulation to study the choices researchers have to make when implementing the method, such as the number of interpolation points, the order of the spline functions, or the information criterion for model selection. Another goal of the simulation is to compare in detail the performance to alternative (linear) models and model selection techniques such as the t-statistic adjustment of Harvey et al. (2016) or the false-discovery rate p-value adjustment of Green et al. (2017).

Table 3 reports average annualized returns with standard errors in parentheses of 10 equally-weighted portfolios sorted on the characteristics we study. Most of the 62 characteristics individually have predictive power for expected returns in our sample period and result in large and statistically significant hedge portfolio returns and alphas relative to the Fama and French three-factor model (Table 4). Thirty-one sorts have annualized hedge returns of more than 5%, and 13 characteristics are even associated with excess returns of more than 10%. Thirty-six characteristics have a t-statistic above 2. Correcting for exposure to the Fama-French three-factor model has little impact on these findings. The vast majority of economic models, that is, the ICAPM (Merton (1973)) or consumption-based 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 incremental information for expected returns.

To tackle the "multidimensional challenge," we now estimate the adaptive group LASSO with 10, 15, 20, and 25 knots. The number of knots corresponds to the smoothing parameter we discuss in Section II. Ten knots corresponds to 11 portfolios in sorts.

We first show in a series of figures a few characteristics which provide large cross sectional return premia univariately. However, some of the characteristics do not provide

incremental predictive power once we condition on other firm characteristics.

Figure 1 and Figure 2 plot estimates of the function $\tilde{m}(\tilde{C}_{it-1})$ for adjusted turnover (**DTO**), idiosyncratic volatility (**Idio vol**), the change in inventories (**IVC**), and net operating assets (**NOA**). The left panels report the unconditional mean functions, whereas the right panels plot the associations between the characteristics and expected returns conditional on all selected characteristics.[8]

Stocks with low change in inventories, low net operating assets but high turnover and high idiosyncratic volatility have higher expected returns than stocks with high change in inventories, net operating assets, and low turnover or idiosyncratic volatility unconditionally. These results are consistent with our findings for portfolio sorts in Table 3. Portfolio sorts result in average annualized hedge portfolio returns of around 13%, 3%, 8%, and 9% for sorts on turnover, idiosyncratic volatility, change in inventories, and net operating profits, respectively. Change in inventories, net operating assets, and turnover have t-stats relative to the Fama-French three-factor model substantially larger than the threshold Harvey et al. (2016) suggest (see Table 4).

These characteristics, however, are correlated with other firm characteristics. We now want to understand whether they have marginal predictive power for expected returns conditional on other firm characteristics. We see in the right panels that the association of these characteristics with expected returns vanishes once we condition on other stock characteristics. The estimated conditional mean functions are now close to constant and do not vary a lot with the value of the characteristics. The constant conditional mean functions imply turnover, idiosyncratic volatility, the change in inventories, and net operating assets have no marginal predictive power for expected returns once we condition on other firm characteristics.

The examples of turnover, idiosyncratic volatility, the change in inventories, and net operating assets show the importance of conditioning on other characteristics to infer the predictive power of characteristics for expected returns. We now study this question

[8]We estimate the plots over the full sample and all firms using 20 interpolations points, see column (1) of Table 5.

systematically for 62 firm characteristics using the adaptive group LASSO.

Table 5 reports the selected characteristics of the nonparametric model for different numbers of knots, sets of firms, and sample periods. Theory does not tell us what the right number of interpolation points is similar to the number of portfolios in sorts but only that we should use more interpolation points when the sample grows large. Allowing for more interpolation points allows for a better approximation of the conditional mean function but comes at the cost of having to estimate more parameters and, hence, higher estimation uncertainty. Previous research also documents that some firm characteristics have larger predictive power for smaller firms and that the predictive power of characteristics varies over time.

We see in column (1) that the baseline estimation for all stocks over the full sample period using 20 knots selects 13 out of the universe of 62 firm characteristics. The change in shares outstanding, investment, size, share turnover, the adjusted profit margin, short-term reversal, momentum, intermediate momentum, closeness to the 52 weeks high, the return on cash, standard unexplained volume, and total volatility all provide incremental information conditional on all other selected firm characteristics. When we allow for a wider grid in column (2) with only 15 knots, we also select the book to market ratio, net operating assets, and long-term reversal. We instead select the same characteristics when we impose a finer grid and estimate the group LASSO with 25 interpolation points (see column (3)).

Figure 5 shows how the number of characteristics we select varies with the number of interpolation points. We see the number of selected characteristics is stable around 20 interpolation points and varies between 16 when we use only 10 knots and 12 when we use 30 interpolation points. We consider the stability of the number and identity of selected characteristics a success documenting the method we propose is not sensitive to the choice of tuning parameters but we provide substantially more robustness checks in the controlled environment of a simulation below.

We estimate the nonparametric model only on large stocks above the 10%- and

20%-size quantile of NYSE stocks in columns (4) to (6), reducing the sample size from more than 1.6 million observations to around 760,000.

The change in shares outstanding, investment, short-term reversal, momentum, intermediate momentum, the return on cash, standard unexplained volume, and total volatility are significant return predictors both for a sample of firms above the 10%-size threshold and the sample of all stocks in column (1), whereas the sales to price ratio becomes a significant return predictor. For firms above the 20%-size threshold of NYSE firms, we also see momentum losing predictive power, but returns over the last six months becoming a significant return predictor. When we impose a coarser grid with only 10 knots for a sample of firms above the 20%-size threshold of NYSE firms in column (6), we see closeness to the 52 weeks high and long-term reversal regaining predictive power, whereas standard momentum driving out intermediate momentum.[9]

Columns (7) and (8) split our 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. Only 11 characteristics have predictive power for expected returns in the sample until 1990, whereas 14 characteristics provide incremental predictive power in the second half of the sample until 2014.

The change in shares outstanding, short-term reversal, momentum, the closeness to the previous 52-week high, the return on cash, standardized unexplained volume, and total volatility are the most consistent return predictors across different sample periods, number of interpolation points, and sets of firms. Some of these characteristics might proxy for risk exposures and their associations with returns could be a rational compensation for risk (see Kelly et al. (2017)). Instead, the predictive power of variables like the change in shares outstanding, past-return based predictors or the closeness to the 52-week high for returns is unlikely risk based and possibly reflects mispricing. Pontiff and Woodgate (2008) discuss several mispricing stories for why the change in shares outstanding predicts returns, such as market timing of managers. If market timing and mispricing partially

[9]The number of knots increases with the sample size. The penalty function instead increases in the number of knots, which is why we select fewer characteristics with more knots.

explain the predictive power of characteristics for returns, then we would also expect to find time variation in the importance of characteristics for return prediction and to find different characteristics to contain predictive power for returns across parts of the firm-size distribution. Moreover, academic research or data mining might also partially destroy return predictability, which would suggest variation in the predictive power of characteristics for returns (see McLean and Pontiff (2016) and Harvey et al. (2016)). Below, we indeed find time variation in the importance of characteristics for return predictions and Table 5 also shows variation in the importance of firm characteristics for small and big firms.

Figure 3 and Figure 4 plot the conditional and unconditional mean functions for short-term reversal, the closeness to the previous 52-week high, size, and standard unexplained volume. We see in Figure 3 both for reversal and closeness to the 52 weeks high a monotonic association between the characteristic distribution and expected returns both unconditionally and once we condition on other characteristics in the right panel. Size matters for returns for all firms in the right panel of Figure 4 and the conditional association is more pronounced than the unconditional relationship in the left panel. This finding is reminiscent of Asness, Frazzini, Israel, Moskowitz, and Pedersen (2017), who argue "size matters, if you control your junk." We see in the lower panels, standardized unexplained volume is both unconditionally and conditionally positively associated with expected returns.

This section shows that many of the univariately significant return predictors do not provide incremental predictive power for expected returns once we condition on other stock characteristics. In particular, out of the 62 firm characteristics we study, we never selected 41 of them! The other 21 characteristics were selected at least for some sample periods, cuts by firm size or number of interpolation points with three of them being selected for each single cut of the data.

C Interactions of Firm Characteristics and Selection in the Linear Model

We discuss in Section II the impact of estimating our model fully nonparametrically on the rate of convergence of the estimator, the so-called curse of dimensionality, and that imposing an additive structure on the conditional mean function offers a solution. The additive structure implies the effect of one characteristic on returns is independent of other characteristics once we condition on them, a form of conditional independence, just as in any multivariate regression. Creating pseudo characteristics, which are themselves interactions of firm characteristics, offers a possible solution to the additive structure and we now show a simple application. Specifically, we interact each of the 61 firm characteristics other than firm size with firm size for a total of 123 firm characteristics. For example, one of the new characteristics is $LME \times BEME$, firm size interacted with the book-to-market ratio.

Table 6 tabulates the results. Instead of selecting 13 characteristics as in the baseline (see column (1) of Table 5), we now select a total of 25 out of the 123 firm characteristics. The model selects 10 of the 13 characteristics it already selected in the baseline. Interestingly, return on cash, which is one of the most consistent return predictors in our baseline table across specifications, is no longer a significant return predictor once we allow for interactions with firm size. Contrary to our baseline, we also no longer select firm size in levels in the model with interactions. Among the 25 characteristics we select in the new model with interactions, almost half are interactions with firm size.

We see in columns (2) to (4) of Table 6 that interactions with firm size are mainly important among small stocks. Once we focus on stocks above the 10%- and 20%-size quantile of NYSE stocks only short-term reversal, momentum, and return over the previous six months interact with firms size and provide incremental information for expected returns.

These results are reassuring for previous research which relied on multivariate

regressions to dissect anomalies, especially for papers which tested models on different parts of the firm-size distribution.

Table 7 estimates a linear model with the adaptive LASSO to gain some intuition for the importance of nonlinearities. Specifically, we endow the linear model with the same two-step LASSO machinery we use for our nonparametric model and report how many and which characteristics the linear model selects in-sample. We also implement the false discovery rate p-value adjustment to benchmark our selection results to the influential findings in Green et al. (2017).

When we compare column (1) in Table 5 for the nonparametric model with column (1) in Table 7 for the linear model, we see the linear model selects nine more characteristics in-sample for a total of 24. Interestingly, the linear model selects eight of the 13 characteristics the nonparametric model selects but also selects the book-to-market ratio, the earnings-to-price ratio, or the average bid-ask spread over the previous month, among others.

So far, we used raw characteristics for the linear model, whereas we applied the rank transformation to characteristics in the nonparametric model. We now estimate a linear model with the adaptive LASSO to see whether the use of raw characteristics might explain the larger number of characteristics we select in the linear model. We see in column (2) of Table 7 that estimating a linear model on rank-transformed characteristics results in an even larger number of characteristics which seem to provide incremental information for expected returns.

Table 7 shows nonlinearities between characteristics and returns might result in a larger number of selected characteristics in a linear model, even when we endow it with the same two-step LASSO machinery that we use for the nonlinear model. Hence, allowing for nonlinearities between characteristics and returns is important from the perspective of data reduction. We explore these features more below in simulations. The selection of more characteristics for the linear model is something which we will see again below when we compare the out-of-sample performance of our nonparametric model with the linear

model.

Column (3) of Table 7 uses the false-discovery rate (FDR) p-value adjustment Green et al. (2017) suggest for model selection. Similar to the linear LASSO model, we find FDR selects many more characteristics compared to the nonlinear models. We will study in detail the differences between linear and nonlinear selection methods in a simulation study below (see Section IV).

D Time Variation in Return Predictors

McLean and Pontiff (2016) document substantial variation over time in the predictive power of many characteristics for expected returns. Figure 6 to Figure 9 show the conditional mean function for a subset of characteristics for our baseline nonparametric model for all stocks and ten knots over time. We perform model selection on the first 10 years of data. We then fix the selected characteristics and estimate the nonparametric model on a rolling basis using 10 years of data.

We see in the top panel of Figure 6 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 selected return predictors. Interestingly, the size effect seems largest during the end of our sample period, contrary to conventional wisdom (see Asness et al. (2017) for a related finding). The bottom panel shows that firms with higher profit margin relative to other firms within the same industry have higher expected returns conditional on other firm characteristics, contrary to the unconditional association (see Table 3).

We see in the top panel of Figure 7 that intermediate momentum has a significant conditional association with expected returns throughout the sample period. Interestingly, we do not observe a crash for intermediate momentum, because intermediate losers have always lower returns compared to intermediate winners. In the bottom panel, we see momentum conditional on other firm characteristics was a particular strong return predictor in the middle sample but lost part of the predictive power for expected returns

in the more recent period because of high returns of past losers, consistent with findings in Daniel and Moskowitz (2016).

Figure 8 shows the effect of short-term reversal on expected returns has been strongest in the early sample period because recent losers used to appreciate more than they currently do. The bottom panel shows the association of the change in shares outstanding and returns has been almost flat until the early 1990s and only afterwards did stocks with the highest level of issuances earn substantially higher returns than all other stocks conditional on other firm characteristics.

Figure 9 plots the conditional mean function for turnover and standard unexplained volume over time. Both high unexplained volume and turnover are associated with high returns but whereas the effect of unexplained volume conditional on other characteristics appears stronger early on, the predictive power of turnover seems stronger in the second part of the sample.

We see those figures as one application of our proposed method for the cross section of stock returns and do not want to put too much weight on the *eyeball econometrics* we performed in the previous section. Ultimately, we cannot tell causal stories and the results might change when we condition on additional firm characteristics. Nevertheless, we consider those three-dimensional surface plots for a given characteristic conditional on other characteristics useful for providing some insights into the time variation of and possible drivers for disappearing or (re-)appearing predictability of a given characteristic.

E Out-of-Sample Performance and Model Comparison

We argued above the nonparametric method we propose overcomes potential shortcomings of more traditional methods, and show potential advantages of the adaptive group LASSO in simulations below.

We now want to compare the performance of the nonparametric model with the linear model out-of-sample. The out-of-sample context ensures that in-sample overfit does not explain a potentially superior performance of the nonparametric model.

We estimate the nonparametric model for a period from 1965 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 over the same sample period, that is, we give both the nonparametric model and the linear model the same machinery and, hence, equal footing. We then use 10 years of data to estimate the model on the selected characteristics. In the next month, we take the selected characteristics and predict one-month-ahead returns and construct a hedge portfolio going long stocks with the highest predicted returns and shorting stocks with the lowest predicted returns. We then roll the estimation and prediction period forward by one month and repeat the procedure until the end of the sample.

Specifically, in our first out-of-sample predictions, we use return data from January 1981 until December 1990 and characteristics data from January 1981 until November 1990 to get estimates of $\boldsymbol{\beta}$.[10] We then take the estimated coefficients and characteristics data of December 1990 to predict returns for January 1991 and form two portfolios for each method. We buy the stocks with the highest predicted returns and sell the stocks with the lowest predicted returns. We then move our estimation sample forward by one month from February 1981 until January 1991, get new estimates $\hat{\boldsymbol{\beta}}$, and predict returns for February 1991.

Panel A of Table 8 reports the out-of-sample Sharpe ratios for both the nonparametric and linear models for different sample periods and firms when we go long the 10% of firms with highest predicted returns and short the 10% of firms with lowest predicted return. For a sample from 1991 to 2014 and ten knots, the nonparametric model generates an out-of-sample Sharpe ratio for an equally-weighted hedge portfolio of 2.75 compared to 1.06 for the linear model (compare columns (1) and (2)). The linear model selects 30 characteristics in-sample compared to only 11 for the nonparametric model, but performs worse out-of-sample.[11] Splitting the Sharpe ratio into a return part and a standard

[10] To be more precise, for returns until June 1981, many of the balance-sheet variables will be from the fiscal year ending in 1979.

[11] The linear model might be misspecified and therefore selects more variables (see discussion and simulation results below in Section IV).

deviation part, we see the nonlinear model generates hedge returns that are almost twice as large compared to the returns the linear mode generates but with substantially lower standard deviation. The nonlinear model has slightly higher positive skewness and similar kurtosis relative to the linear model. When we calculate average monthly turnover statistics over time, we find the nonlinear model has slightly larger turnover. Turnover1 follows Koijen, Moskowitz, Pedersen, and Vrugt (2018) and is defined as $turn_t = \frac{1}{4}\sum_i^{N_t}|(1+r_{it})w_{it-1} - w_{it}|$ where w_{it} is the portfolio weight of stock i at time t and N_t is the number of stocks and Turnover2 corresponds to $turn_t = \frac{1}{4}\frac{1}{N_t}\sum_i^{N_t}|\omega_{it-1} - \omega_{it}|$ where $\omega_{it} \in \{-1, 0, 1\}$ and hence corresponds to the fraction of stocks that change portfolios. When then follow Lewellen (2015) to study how accurate the individual models are in predicting returns. Specifically, we regress realized returns at the stock level on predicted returns month by month and report average slopes and R^2s over time. Ideally, we want to find slope coefficients close to 1 and high predictive power. Lewellen (2015) discusses slope coefficients below 1 indicate predictive models exaggerate expected return dispersion. The nonlinear adaptive group LASSO has a slope coefficient of 0.78 and a R^2 of almost 2%. The slope coefficient for the full sample is very similar to Lewellen (2015) but the predictive power is somewhat larger. The linear model instead has an average slope which is only half the size and the predictive power for realized returns is more than 30% lower. Panel B and C repeat the same statistics but for the long and the short leg of the hedge portfolio separately. In general, we find higher returns for the long leg and more negative skewness for the short leg with similar kurtosis.

Nonlinearities are important. We find a substantial increase in out-of-sample Sharpe ratios relative to the Sharpe ratio of the linear model when we employ the nonparametric model for prediction on the 30 characteristics the linear model selects (see column (3)).

The linear model appears to overfit the data in-sample. When we use the 11 characteristics we select with the nonparametric model, we find the Sharpe ratio for the linear model is identical to the one we find when we use the 30 characteristics the linear model selects (see column (4)). But even with the same set of 11 characteristics,

we find the Sharpe ratio for the linear model is still substantially smaller compared to the Sharpe ratio of the nonparametric model. In line with our findings above, it appears the linear model selects many characteristics in-sample that do not provide incremental information for return prediction, but also that nonlinearities are important.

Columns (5) and (6) focus on a longer out-of-sample period starting in 1973 to be comparable to results in the literature (see, e.g., Lewellen (2015)). Results are very similar to when we split the sample in half.

We see in columns (7) to (10) that Sharpe ratios drop substantially for both models when we exclude firms below the 10^{th} or 20^{th} percentile of NYSE stocks. Lewellen (2015) also finds Sharpe ratios for equally-weighted hedge portfolios that are lower by 50% when he excludes "all but tiny stocks." The Sharpe ratios are still around 1 for the nonparametric model for both sets of stocks, whereas Sharpe ratios are only around 0.10 for the linear model.

Panel A also reports Sharpe ratios adjusted for transaction costs. We follow DeMiguel et al. (2016) and model proportional transaction costs, κ_{it}, as a function of time and firm size

$$\kappa_{it} = y_t z_{it}, \tag{6}$$

where y_t starts at a value of 3.3 in January 1980 and linearly decreases to a value of 1 in January 2002 and remains constant afterwards and $z_{it} = 0.006 - 0.0025 \times me_{it}$, where me_{it} is the ranked-normalized market capitalization of firm i in period t. We then calculate the portfolio return after transaction costs, $r_{p,t}^{\text{after cost}}$ as

$$r_{p,t}^{\text{after cost}} = \sum_{i=1}^{N_t} \left[\underbrace{w_{i,t} r_t}_{\text{return contribution of i}} - \underbrace{\left| (w_{i,t} - (1 + r_{i,t}^*) w_{i,t-1}) \right| \times tc_{i,t}}_{\text{trading cost from changing holdings in i}} \right], \tag{7}$$

where $r_{i,t}$ is the return with dividends reinvested for stock i in period t, $r_{i,t}^*$ is the return without dividends reinvested, that is, the capital gain, and $w_{i,t}$ is the portfolio weight. We see in column (1) that the Sharpe ratio decreases from 2.75 without transaction costs

to 1.56 after transaction costs for the adaptive group LASSO. Average transaction costs are 1.71% per month. For the linear model, the Sharpe ratio decreases from 1.06 before transaction costs to only 0.29 after transaction costs with average transaction costs of 1.54 per month (see column (2)). Once we exclude small firms in columns (7) to (10), we see that Sharpe ratios adjusted for transaction costs become small and often negative with more negative values for the linear model. We want to stress the way we use to adjust Sharpe ratios for transaction costs is only ad-hoc in an ex-post sense. In practical applications, researchers would typically take transaction costs directly into account at the portfolio construction stage which would possibly lead to higher adjusted Sharpe ratios, both for the linear but also for the nonlinear model.

Results are similar when we perform rolling selection. So far, we performed model selection once, fixed the selected characteristics for the nonparametric and linear model, and performed rolling model estimation and return prediction. As a robustness check, we also perform annual model selection on a constant sample size of 26 years, fix the selected characteristics for 12 months and perform rolling monthly estimation and prediction. We then roll forward the selection period by one year. The first selection period is from January 1965 until December 1990 and the first out-of-sample return prediction is for January 1991.

Table 9 reports the results for the rolling selection with 10 knots. Overall, the results for the rolling selection are very similar to before. The nonlinear model selects 14 characteristics on average and has a similar out-of-sample Sharpe ratio, but slightly higher predictive power for future returns, and a higher R^2. Figure 10 plots the characteristics the nonlinear adaptive group LASSO selects over time and Figure 11 the corresponding figure for the linear adaptive LASSO. Selected characteristics are indicated in dark blue. The nonlinear model consistently selects a lower number of characteristics over time relative to the linear model throughout the period and the identity of characteristics the nonlinear model selects is surprisingly consistent over time suggesting that certain firm characteristics reliably provide information for return prediction.

IV Simulation

Section III shows the nonlinear adaptive group LASSO achieves a large data reduction relative to the linear model and increases out-of-sample predictability but so far, we do not know the assumptions on the data-generating process under which the nonlinear adaptive group LASSO performs well and what happens to model selection and out-of-sample prediction when we change assumptions. The aim of this section is to discuss some of the *tuning* parameters of the method we lay out in Section II such as the choice of the penalty parameter or the number of interpolation points and compare the adaptive group LASSO to alternative model selection methods.

Specifically, we want to simulate returns using our full set of return predictors and compare model selection techniques and the choices of penalty parameters, knots, and order of splines in the LASSO. We consider the following selection methods:

- Conventional t-statistic cutoff of 2
- t-statistic cutoff of 3 to account for multiple testing (Harvey et al. (2016))
- The false discovery rate (FDR) p-value adjustment of Green et al. (2017)
- Linear single-step LASSO
- Linear adaptive LASSO
- Nonlinear group LASSO
- Nonlinear adaptive group LASSO.

We also employ different LASSO methods in addition to the adaptive group LASSO. The single-step LASSO only estimates the first step of the method we outline in Section II. The adaptive LASSO consists of two stages. The group LASSO treats a given characteristic across the whole distribution as a joint return predictor.

Regarding the choice of penalty parameter, we consider:

- Akaike information criterion (AIC)
- Bayesian information criterion (BIC)
- BIC as in Yuan and Lin (2006)

- Ten fold cross validation.

All three information criteria trade off the costs of a larger number of parameters against the better fit. AIC and BIC differ in how they penalize additional parameters. For AIC, the penalty is twice the number of parameters, whereas for BIC, it is the number of parameters times the natural logarithm of the number of observations. Yuan and Lin (2006) develop an adjusted BIC for the case of grouped variables. In ten fold cross validation, we partition our data into ten subset, estimate the models on nine subsets and use the remaining one for out-of-sample return prediction, that is, model validation. We repeat the procedure nine times, using each sample exactly once for validation and then average across samples. Cross validation then chooses the penalty parameter which is associated with the lowest mean-squared prediction error. We also study the importance of the number of knots, the order of the polynomial, and the firm-size distribution, both for selection and out-of-sample prediction.

Our simulation then proceeds in the following steps:

1. Take the full data set of 62 characteristics, C_{it} from Section III
2. Focus on a sample from 1965 to 2012
3. Assume the 13 characteristics of column (1) of Table 5 are the "true" predictors
4. Transform all characteristics to be standard normal distributed
5. Fit a fifth-order polynomial on the true characteristics to estimate $g_s(C_{s,it-1})$ for each characteristic pooled over the entire sample
6. Generate returns according to: $r_{it} = \sum_{s=1}^{13} g_s(C_{s,it-1}) + \varepsilon_{it}$
7. ε_{it} is resampled with replacement from the empirical residuals preserving industry structure (details below)
8. Estimate nonparametric model on rank-transformed data with 20 knots

9. Estimate linear model on data from step 4

10. Redo steps 6 to 9 500 times

Regarding step 7, we save the unbalanced panel of residuals from step 5 and assign an industry label to each firm i in each time period t, following the 48 industry classification of Fama and French. To generate the residuals in a particular time period t, we then first draw a random time period, say time period s, from which we sample the residuals. For example, to generate the residuals for time period 1, we might use the residuals from time period 120. For each firm in time period t, we then draw a random residual from time period s, but use only residuals from firms with the same industry code. Since we have a different number of firms in different time periods, we sample the residuals within each industry with replacement. Moreover, to ensure we sample from distributions with means of 0, we re-center the original residuals by time and industry. Notice that this sampling process leads to both time and industry heterogeneity because for each time period and for each industry we sample from different distributions, which can have, for example, very different variances and skewness.[12]

A Model Selection

The advantage of this setup is that we directly take into account the cross sectional and time series correlation structure of the actual data in the simulation and do not have to make any assumption on whether the true model is linear or nonlinear. The aim of the simulation is then to see how the different methods for model selection perform, which in our context means: does a given model select on average the right number and identify of characteristics and does not select characteristics that do not provide information for returns according to the data generating process. For the selected characteristics, we then also study the out-of-sample predictive power using two years of data for 2013 to 2014.

Figure 12 graphically illustrates the results of the simulations for the different model selection methods. We indicate the different models on the x-axis and the characteristics

[12]We thank an anonymous referee for inspiring this re-sampling procedure.

on the left y-axis. The color scheme on the right y-axis indicates the frequency with which a given characteristic is selected. The darker the color, the more frequently a given selection method selects a characteristic. The darkest blue indicates a given characteristic is selected in 100% of the simulations and white indicates the characteristic was never selected. The red horizontal line below Total_vol represents the cut-off for return predictors. The 13 characteristic above the line indicate true return predictors, whereas the 49 other characteristics below the line do not predict returns.

Given the structure of the data, we want a model selection method which selects all relevant return predictors with high probability and does not select all irrelevant return predictors. Hence, ideally we want to have methods for selection that have dark blue shaded areas above the red line and white areas below the line. We see in column (1) the adaptive group LASSO, which corresponds to our baseline model in the empirical application with 20 interpolation points and second-order polynonial using the BIC of Yuan and Lin (2006) tends to select all 13 of the true return predictors, and it does not select the irrelevant return predictors. On average, across the 500 simulations, the nonlinear adaptive group LASSO selects 12.99 characteristics. Column (2) employs the same basic setup for model selection, but estimates only a single-step LASSO. We see the group LASSO tends to select all relevant return predictors, but also few irrelevant ones as we would expect from the irrepresentable condition of Meinshausen and Bühlmann (2006). On average, it selects 16.89 characteristics.

Columns (3) and (4) endow the linear model with the same LASSO methods we used for the nonlinear model. Similar to our empirical application, we see that the linear LASSO tends to select the relevant return predictors but also many characteristics that are not associated with returns. The adaptive LASSO tends to select 29.36 characteristics across simulation and the single-step LASSO even 47.72. The last three columns of the figure use the FDR p-value adjustment of Green et al. (2017), a t-statistic cutoff of 3 (t3) to account for multiple testing as Harvey et al. (2016) suggest, and the conventional t-statistics cutoff of 2 (t2). Across the three selection methods, we see a high probability

of selecting relevant return predictors, but also a high probability to select irrelevant return predictors. FDR selects 27.30 characteristics, t3 selects 26.40, and t2 selects 33.75 on average.

In the online appendix, we graphically illustrates the results of the simulations for different choices of tuning parameters. Figure A.4 shows the result for the adaptive group LASSO for different information criteria. Column (1) repeats our baseline choice. In column (2), we use an AIC to determine the penalty parameters. Using AIC tends to result in a high probability of selecting relevant returns predictors, but does also select a few irrelevant predictors for a total of 14.59 on average. When we use the standard BIC instead of the one proposed by Yuan and Lin (2006) for group LASSO applications, we find the standard BIC tends to perform very similar to the BIC of Yuan and Lin (2006), and selects 12.97 characteristics on average. The last column uses cross validation to determine the penalty parameters. We see that for the context of return prediction when using the actual characteristics data, cross validation does not result in a desirable model selection. It tends to select all characteristics with high probability for an average of 56.34.

Figure A.5 studies the effect of choosing a different number of interpolation points – ranging from 10 to 25 – on the number and identity of selected characteristics. Across columns, we see a high probability of selecting relevant return predictors and not selecting irrelevant return predictors. On average, we select 13.03 for 10 knots, 13.01 for 15 knots, 12.99 for 20 knots, and 12.94 for 25 knots. In Figure A.6 instead, we study how the choice of the order of the splines affects the selection results for our baseline adaptive group LASSO with 20 knots where order 0 corresponds to a step function, order 1 to a piecewise linear function, etc. Across orders, splines tend to perform well in selecting relevant return predictors. The average number of selected characteristics across simulations are 13.08 (order 0), 15.55 (order 1), 12.99 (order 2), 16.63 (order 3) and 16.61 (order 4). Selection results for large stocks are similar to results for all stocks (see Figure A.7 in the online appendix).

The simulation study using the true underlying data and functional relationship between characteristics and returns so far shows that: (i) t-statistics based selection methods have little power; (ii) nonlinearities are important for selection; and (iii) the second-stage of the LASSO matters, that is, the irrepresentable condition does not hold in our data; (iv) the adjusted BIC performs best in selecting relevant return predictors and not selecting irrelevant return predictors.

Instead of using the approximated true functional relationship between characteristics and returns with a fifth-order polynomial on the true characteristics, we can also assume the true data generating process is linear and simulate returns under this assumption. Unfortunately, the actual relationship in the data is nonlinear and we do not know the "true" number and identity of characteristics for a linear model. To ensure the simulation setup is comparable to the true data-generating process we simulate above, we do the following: (i) we assume also in the linear model 13 characteristics predict returns; (ii) we choose the 13 characteristics by "walking along the LASSO path", that is, we vary the penalty parameter until the adaptive LASSO in the linear model selects 13 characteristics; and (iii) we estimate the linear association between these 13 characteristics and returns.

Figure 13 plots the selection results. Again, the 13 characteristics above the red horizontal line represent the "true" return predictors. In column (1), we see that even when we assume the data-generating process is linear, allowing for nonlinearities with the nonlinear adaptive group LASSO does no harm in the model selection stage. The model selects 12 out of 13 return predictors with high probability and does not the select irrelevant return predictors. In particular, we also see that the nonlinear adaptive group LASSO performs similar to the linear adaptive LASSO, FDR, or t3 on a dataset which by construction favors linear models. Both single-step LASSO procedures and t2 tend to select too many characteristics that do not provide information for return prediction. On average, across 500 simulation, the nonlinear adaptive group LASSO selects 10.98 characteristics, the group LASSO 15.58, the linear adaptive LASSO 12.60, the linear LASSO 16.77, FDR 10.45, t3 10.64 and t2 14.15 characteristics.

Hence, when nonlinearities matter as in the actual data, the nonlinear adaptive group LASSO performs best in model selection compared to linear methods but when we force the data-generating process to be linear, the nonlinear adaptive group LASSO does just as well. Hence, it seems natural to at least allow for nonlinearities.

B Out-of-Sample Prediction

Overall, we saw the nonlinear adaptive group LASSO does a good job in selecting relevant return predictors and not selecting irrelevant return predictors across different assumptions on the underlying data generating process and tuning parameters. The good performance in model selection, however, does not necessarily mean the nonlinear adaptive group LASSO performs well in out-of-sample return predictions. To study the latter, we now predict returns out-of-sample for all model selection methods, tuning parameters, and assumptions regarding the data generating process for a sample from 2013 to 2014. We simulate returns again for 500 times, perform model selection and estimate the model using the sample form 1965 until 2012 and predict returns. To study how well the models predict returns, we regress realized returns on predicted returns and report R^2s but also report root mean squared prediction errors (RMSPE).

Panel A of Table 10 reports the results. The first line first reports results for the true parametric model underlying the simulation. When we regress realized returns that include sampling uncertainty on predicted returns, we find a R^2 of 1.6% and a RMSPE of 0.12. In the following, we directly report R^2s and RMSPEs for the different model selection methods relative to these "true" numbers. The second line reports results for the "true" nonparametric model, that is, we endow the nonlinear model with the knowledge on the actual 13 return predictors but estimate the nonlinear functions from the data before predicting returns. The true nonparametric model without selection uncertainty achieves a relative R^2 of 88.84% and a RMSPE that is larger by 0.09% relative to the true model. Line three now reports results for the nonlinear adaptive group LASSO. We see the model achieves a relative R^2 of 88.61% and a relative RMSPE of 0.092% which documents the

high model selection accuracy of the method. In case we are purely interested in predicting returns out-of-sample, then we see a group LASSO performs almost equally well. The following lines instead show all of the linear models do substantially worse predicting returns out-of-sample when we follow the true data-generating process. Independent of whether we use LASSO-based methods for the linear model, t-statistics based methods, or the FDR p-value adjustment of Green et al. (2017), the relative R^2 is never larger than 58%, thirty percentage points less than for the nonlinear LASSO methods and the RMSPE is larger by a factor of 3 relative to the nonlinear LASSO: 0.1% versus 0.3%.

Panel B of Table 10 reports the results for the linear data-generating process. We see the true parametric model now achieves a R^2 of slightly below 1% and the true nonparametric model, that is, the nonlinear model endowed with the true 13 characteristics, achieves a relative R^2 of 94%. Both the nonlinear adaptive group and group LASSO achieve a relative R^2 which is similar. Hence, even when we counterfactually assume that the data-generating process is linear, we still find a good out-of-sample return prediction for the nonlinear model. In the following lines, we see the linear model selection methods have relative out-of-sample R^2s between 97% and 99%. Interestingly, from a pure out-of-sample prediction perspective, a t-statistics threshold of 2 has a higher out-of-sample predictive power than the FDR p-value adjustment of Green et al. (2017) or a t-statistics threshold of 3 similar to out-of-sample prediction results in Green et al. (2017). Both linear and nonlinear models achieve low relative RMSPE. The linear selection methods achieve a relative RMSPE of around 0.01%, whereas the nonlinear methods achieve a relative RMSPE of around 0.03%.

When we simulate the true, nonlinear data-generating process, we find large increases in out-of-sample R^2s for the nonlinear models relative to the linear models. When we instead assume that the data-generating process is linear, we find out-of-sample R^2 for the nonlinear models which are almost identical to the linear models. Hence, it appears natural to us to at least allow for nonlinearities ex-ante in situations in which it is not clear whether nonlinearities matter.

Table A.1 and Table A.2 show robustness tests for different information criteria, number of knots, order of splines or only firms above the 20^{th} NYSE size percentile. Out-of-sample prediction results mirror the model selection conclusions: the BIC of Yuan and Lin (2006) performs better in out-of-sample prediction relative to a standard BIC, results are not very sensitive to the number of knots initially but start to deteriorate with 25 knots, order 0 and order 1 spline perform worse than our baseline model but higher-order splines even improve the out-of-sample forecasting performance, and results for large firms are similar in that the nonlinear models outperform substantially linear models in out-of-sample predictions.

V Conclusion

We propose a nonparametric method to tackle the challenge posed by Cochrane (2011) in his presidential address, namely, which firm characteristics provide incremental information for expected returns. We use the adaptive group LASSO to select important 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 62 characteristics, only nine to 16 provide incremental 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 in-sample and has a Sharpe ratio that is larger by a factor of 2.5 out-of-sample.

In a simulation study, we document the nonlinear adaptive group LASSO performs well in model selection, that is, identifying true return predictors with high probability

and not selecting irrelevant return predictors. Linear model selection methods including t-statistic based cutoffs or false-discovery rate p-value adjustments result in large over-selection, that is, they also classify as return predictors characteristics that do not predict returns. We also show the nonlinear models outperform linear models in out-of-sample return prediction and show our conclusions are robust to variations in the *tuning parameters* our method has.

We see our paper as a starting point only and pose the following questions for future research. 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 dimensions of expected returns with $K << N$ similar to Fama and French (1993) and Fama and French (1996)?

References

Abarbanell, J. S. and B. J. Bushee (1997). Fundamental analysis, future earnings, and stock prices. *Journal of Accounting Research 35*(1), 1–24.

Anderson, A.-M. and E. A. Dyl (2005). Market structure and trading volume. *Journal of Financial Research 28*(1), 115–131.

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 (2017). Size matters, if you control your junk. *Journal of Financial Economics (forthcoming)*.

Asness, C. S., R. B. Porter, and R. L. Stevens (2000). Predicting stock returns using industry-relative firm characteristics. *Unpublished Manuscript, AQR*.

Balakrishnan, K., E. Bartov, and L. Faurel (2010). Post loss/profit announcement drift. *Journal of Accounting and Economics 50*(1), 20–41.

Bali, T. G., N. Cakici, and R. F. Whitelaw (2011). Maxing out: Stocks as lotteries and the cross-section of expected returns. *Journal of Financial Economics 99*(2), 427 – 446.

Ball, R., J. Gerakos, J. T. Linnainmaa, and V. V. Nikolaev (2015). Deflating profitability. *Journal of Financial Economics 117*(2), 225–248.

Bandyopadhyay, S. P., A. G. Huang, and T. S. Wirjanto (2010). The accrual volatility anomaly. *Unpublished Manuscript, University of Waterloo*.

Basu, S. (1977). Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis. *The Journal of Finance 32*(3), 663–682.

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.

Boudoukh, J., R. Michaely, M. Richardson, and M. R. Roberts (2007). On the importance of measuring payout yield: Implications for empirical asset pricing. *The Journal of Finance 62*(2), 877–915.

Brandt, M. W., P. Santa-Clara, and R. Valkanov (2009). Parametric portfolio policies: Exploiting characteristics in the cross-section of equity returns. *Review of Financial Studies 22*(9), 3411–3447.

Breeden, D. T. (1979). An intertemporal asset pricing model with stochastic consumption and investment opportunities. *Journal of Financial Economics 7*(3), 265–296.

Brown, D. P. and B. Rowe (2007). The productivity premium in equity returns. *Unpublished Manuscript, University of Wisconsin*.

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and E. Schaumburg (2016). Characteristic-sorted portfolios: Estimation and inference. *Unpublished Manuscript, University of*

Chicago.

Chandrashekar, S. and R. K. Rao (2009). The productivity of corporate cash holdings and the cross-section of expected stock returns. *Unpublished Manuscript, UT Austin*.

Chinco, A., A. D. Clark-Joseph, and M. Ye (2018). Sparse signals in the cross-section of returns. *Journal of Finance (forthcoming)*.

Chinco, A., A. Neuhierl, and M. Weber. Estimating the anomaly baserate. *Unpublished Manuscript*.

Chordia, T., A. Subrahmanyam, and V. R. Anshuman (2001). Trading activity and expected stock returns. *Journal of Financial Economics 59*(1), 3–32.

Chung, K. H. and H. Zhang (2014). A simple approximation of intraday spreads using daily data. *Journal of Financial Markets 17*, 94–120.

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.

Cooper, M. J., H. Gulen, and M. J. Schill (2008). Asset growth and the cross-section of stock returns. *The Journal of Finance 63*(4), 1609–1651.

D'Acunto, F., R. Liu, C. E. Pflueger, and M. Weber (2017). Flexible prices and leverage. *Journal of Financial Economics (forthcoming)*.

Daniel, K. and T. J. Moskowitz (2016). Momentum crashes. *Journal of Financial Economics 122*(2), 221–247.

Datar, V. T., N. Y. Naik, and R. Radcliffe (1998). Liquidity and stock returns: An alternative test. *Journal of Financial Markets 1*(2), 203–219.

Davis, J. L., E. F. Fama, and K. R. French (2000). Characteristics, covariances, and average returns: 1929 to 1997. *The Journal of Finance 55*(1), 389–406.

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

DeMiguel, V., A. Martin-Utrera, F. Nogales, and R. Uppal (2016). A portfolio perspective on the multitude of firm characteristics. *Unpublished Manuscript, London Business School*.

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.

Gagliardini, P., E. Ossola, and O. Scaillet (2016). Time-varying risk premium in large cross-sectional equity data sets. *Econometrica 84*(3), 985–1046.

Gandhi, P. and H. Lustig (2015). Size anomalies in US bank stock returns. *The Journal of Finance 70*(2), 733–768.

Garfinkel, J. A. (2009). Measuring investors' opinion divergence. *Journal of Accounting Research 47*(5), 1317–1348.

George, T. J. and C.-Y. Hwang (2004). The 52-week high and momentum investing. *The Journal of Finance 59*(5), 2145–2176.

Gibbons, M. R., S. A. Ross, and J. Shanken (1989). A test of the efficiency of a given portfolio. *Econometrica 57*(5), 1121–1152.

Giglio, S. W. and D. Xiu (2016). Inference on risk premia in the presence of omitted factors. *Unpublished Manuscript, University of Chicago*.

Gorodnichenko, Y. and M. Weber (2016). Are sticky prices costly? Evidence from the stock market. *The American Economic Review 106*(1), 165–199.

Goto, S. and Y. Xu (2015). Improving mean variance optimization through sparse hedging restrictions. *Journal of Financial and Quantitative Analysis 50*(06), 1415–1441.

Green, J., J. R. Hand, and X. F. Zhang (2017). The characteristics that provide independent information about average us monthly stock returns. *The Review of Financial Studies 30*(12), 4389–4436.

Hahn, J. and H. Lee (2009). Financial constraints, debt capacity, and the cross-section of stock returns. *The Journal of Finance 64*(2), 891–921.

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*(3), 401–439.

Hirshleifer, D., K. Hou, S. H. Teoh, and Y. Zhang (2004). Do investors overvalue firms with bloated balance sheets? *Journal of Accounting and Economics 38*, 297–331.

Horowitz, J., J. Klemelä, and E. Mammen (2006). Optimal estimation in additive regression models. *Bernoulli 12*(2), 271–298.

Horowitz, J. L. (2016). Variable selection and estimation in high-dimensional models. *Canadian Journal of Economics 48*(2), 389–407.

Hou, K., G. A. Karolyi, and B.-C. Kho (2011). What factors drive global stock returns? *Review of Financial Studies 24*(8), 2527–2574.

Hou, K., C. Xue, and L. Zhang (2015). Digesting anomalies: An investment approach. *Review of Financial Studies 28*(3), 660–705.

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

Huang, J.-Z. and Z. Shi (2016). Determinants of bond risk premia. *Unpublished Manuscript, Penn State University*.

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

Jegadeesh, N. and S. Titman (1993). Returns to buying winners and selling losers:

Implications for stock market efficiency. *Journal of Finance 48*, 65–91.

Kelly, B. T., S. Pruitt, and Y. Su (2017). Some characteristics are risk exposures, and the rest are irrelevant. *Unpublished Manuscript, University of Chicago*.

Kim, S., R. A. Korajczyk, and A. Neuhierl (2018). Arbitrage portfolios in large panels. *Unpublished Manuscript 80*, 713–754.

Koijen, R. S., T. J. Moskowitz, L. H. Pedersen, and E. B. Vrugt (2018). Carry. *Journal of Financial Economic (forthcoming)*.

Kozak, S., S. Nagel, and S. Santosh (2017). Shrinking the cross section. *Unpublished Manuscript, University of Chicago*.

Lakonishok, J., A. Shleifer, and R. W. Vishny (1994). Contrarian investment, extrapolation, and risk. *The Journal of Finance 49*(5), 1541–1578.

Lettau, M. and M. Pelger (2018). Factors that fit the time series and cross-section of stock returns. Technical report, National Bureau of Economic Research.

Lewellen, J. (2015). The cross section of expected stock returns. *Critical Finance Review 4*(1), 1–44.

Lewellen, J. and S. Nagel (2006). The conditional CAPM does not explain asset-pricing anomalies. *Journal of Financial Economics 82*(2), 289–314.

Lintner, J. (1965). The valuation of risky assets and the selection of risky investments in stock portfolios and capital budgets. *The Review of Economics and Statistics 47*(1), 13–37.

Litzenberger, R. H. and K. Ramaswamy (1979). The effect of personal taxes and dividends on capital asset prices: Theory and empirical evidence. *Journal of Financial Economics 7*(2), 163–195.

Lucas, R. E. (1978). Asset prices in an exchange economy. *Econometrica 46*(6), 1429–1445.

Lyandres, E., L. Sun, and L. Zhang (2008). The new issues puzzle: Testing the investment-based explanation. *Review of Financial Studies 21*(6), 2825–2855.

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.

Novy-Marx, R. (2011). Operating leverage. *Review of Finance 15*(1), 103–134.

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

Ou, J. A. and S. H. Penman (1989). Financial statement analysis and the prediction of stock returns. *Journal of Accounting and Economics 11*(4), 295–329.

Palazzo, B. (2012). Cash holdings, risk, and expected returns. *Journal of Financial Economics 104*(1), 162–185.

Pontiff, J. and A. Woodgate (2008). Share issuance and cross-sectional returns. *The Journal of Finance 63*(2), 921–945.

Richardson, S. A., R. G. Sloan, M. T. Soliman, and I. Tuna (2005). Accrual reliability, earnings persistence and stock prices. *Journal of Accounting and Economics 39*(3), 437–485.

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.

Soliman, M. T. (2008). The use of DuPont analysis by market participants. *The Accounting Review 83*(3), 823–853.

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.

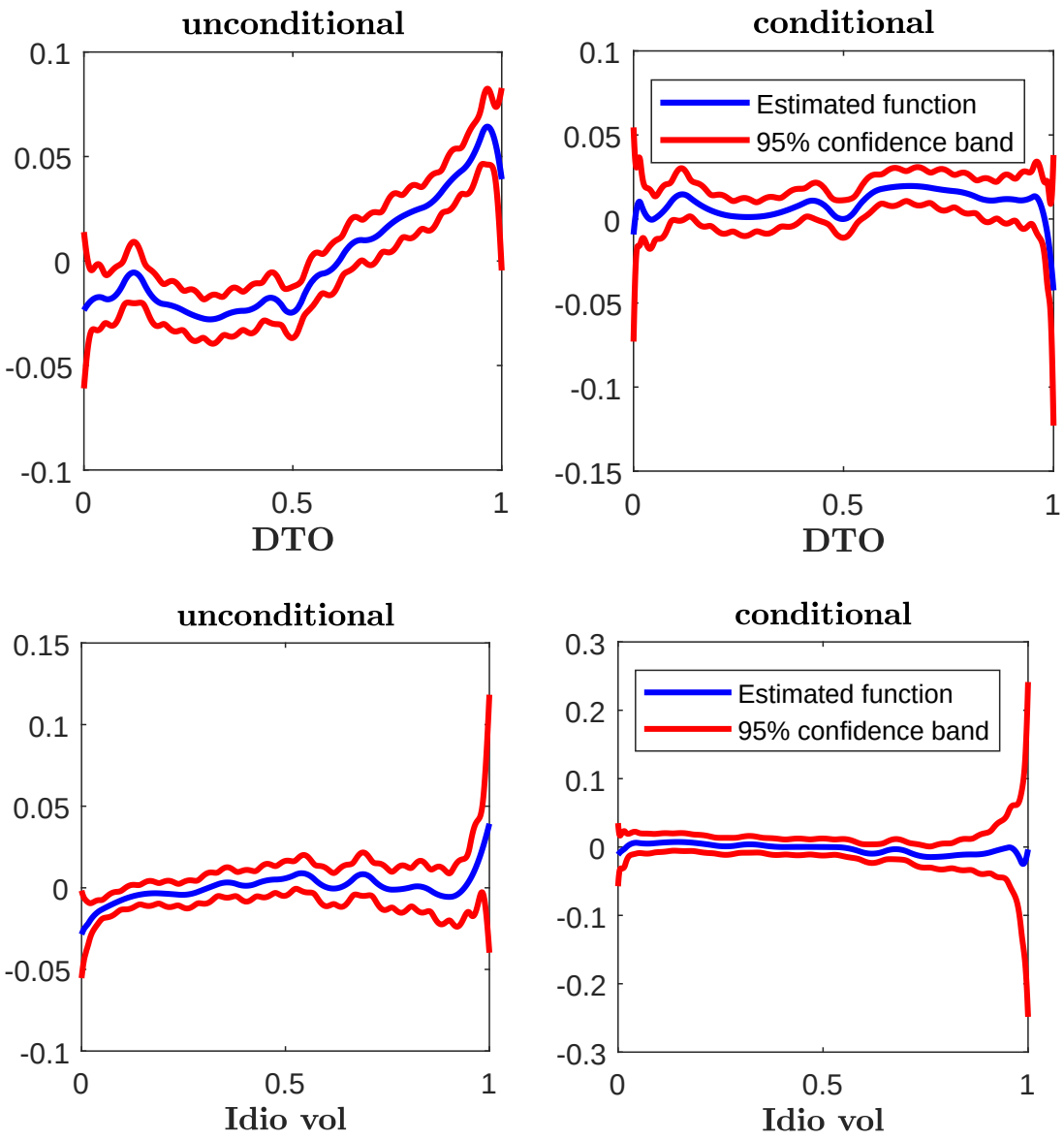
Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *The Annals of Statistics 14*(2), 590–606.

Thomas, J. K. and H. Zhang (2002). Inventory changes and future returns. *Review of Accounting Studies 7*(2), 163–187.

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.

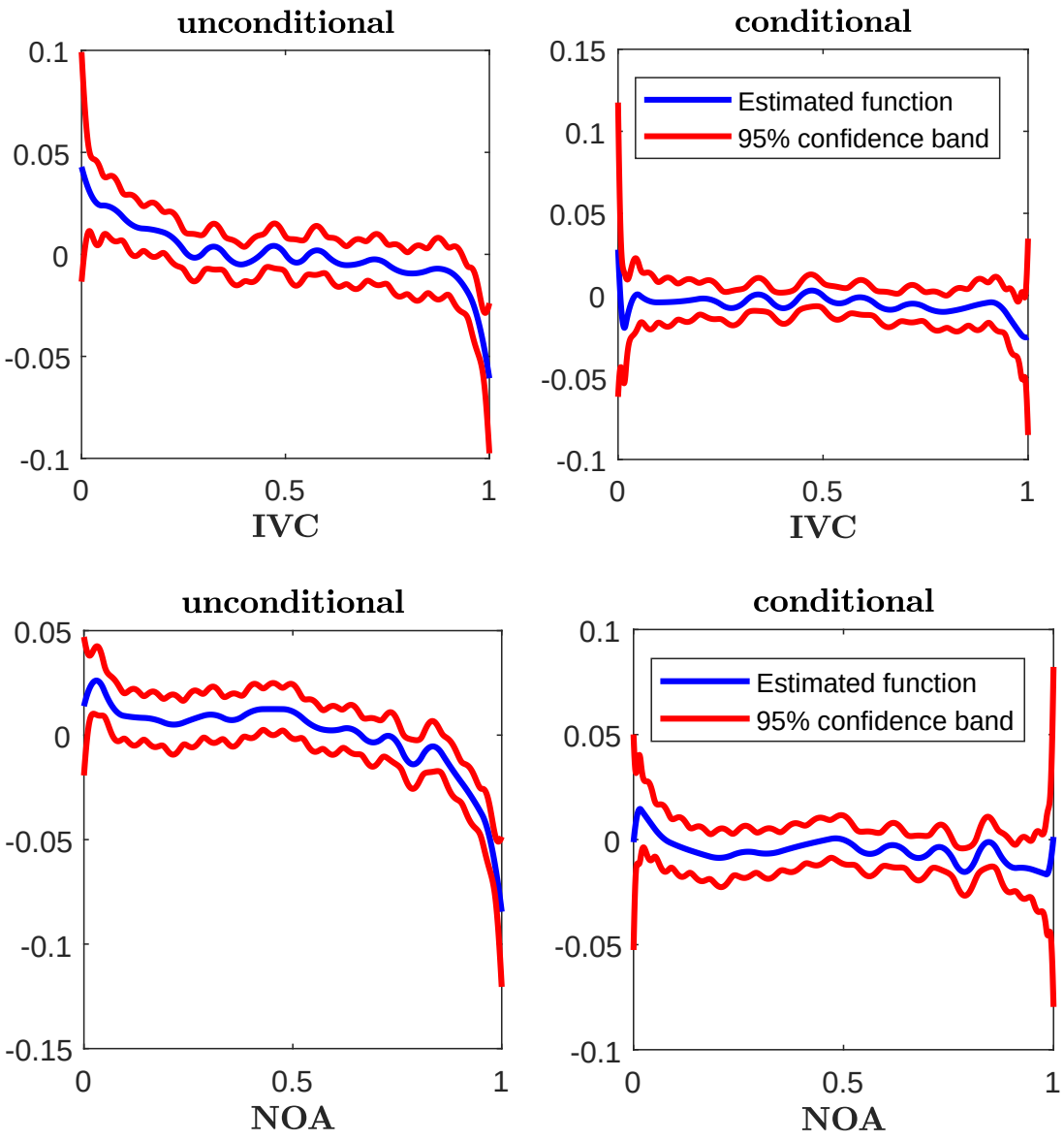
Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.

Figure 1: **Unconditional and Conditional Mean Function: Adjusted Turnover (DTO) and Idiosyncratic Volatility (Idio vol)**



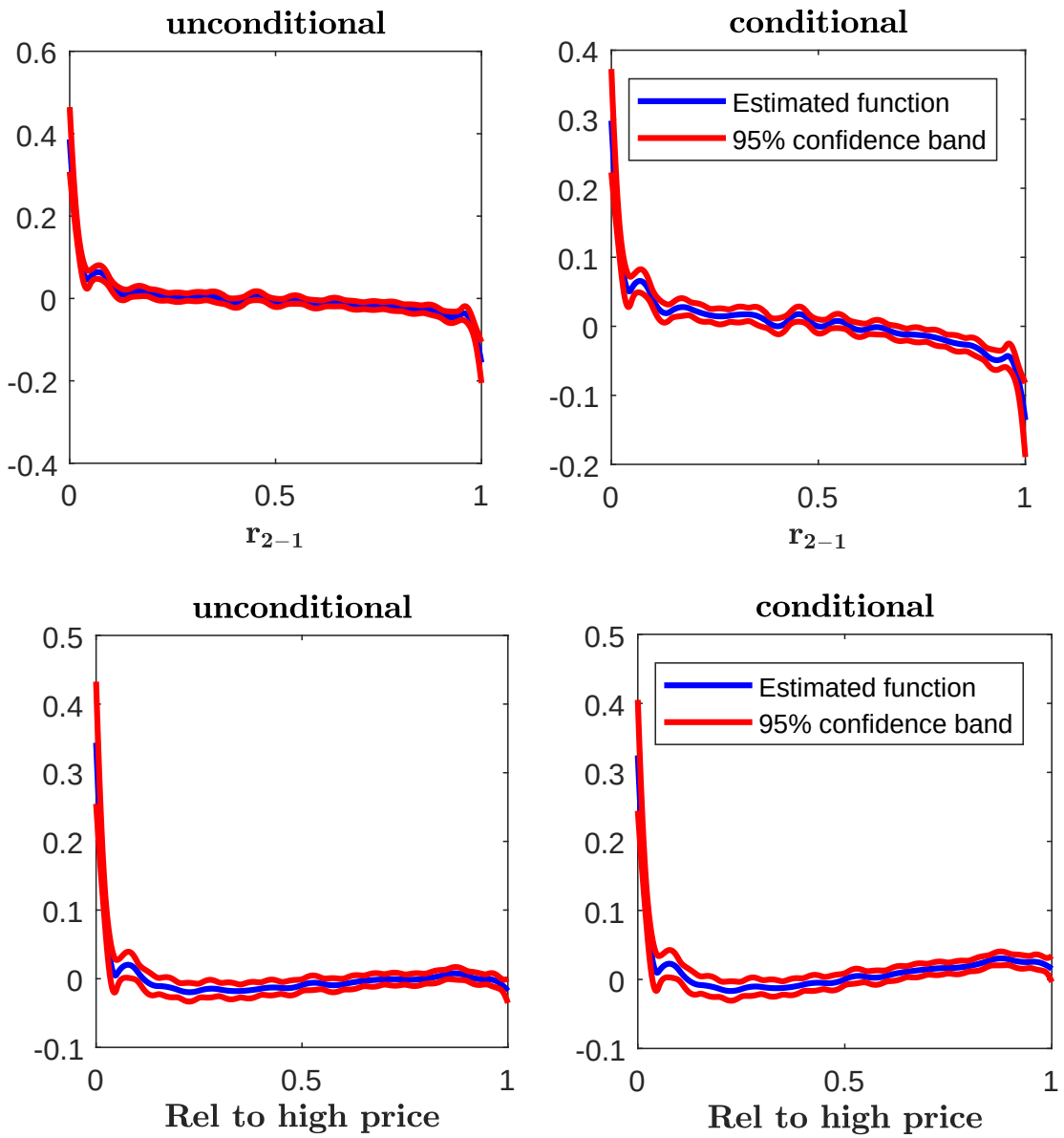
*Effect of normalized adjusted turnover (**DTO**) and idiosyncratic volatility (**Idio vol**) on average returns (see equation (3)). The left panels report unconditional associations between a characteristic and returns, and the right panels report associations conditional on all other selected characteristics. The sample period is January 1965 to June 2014. See Section A.1 in the online appendix for variable definitions.*

Figure 2: **Unconditional and Conditional Mean Function: Change in Inventories (IVC) and Net Operating Assets (NOA)**



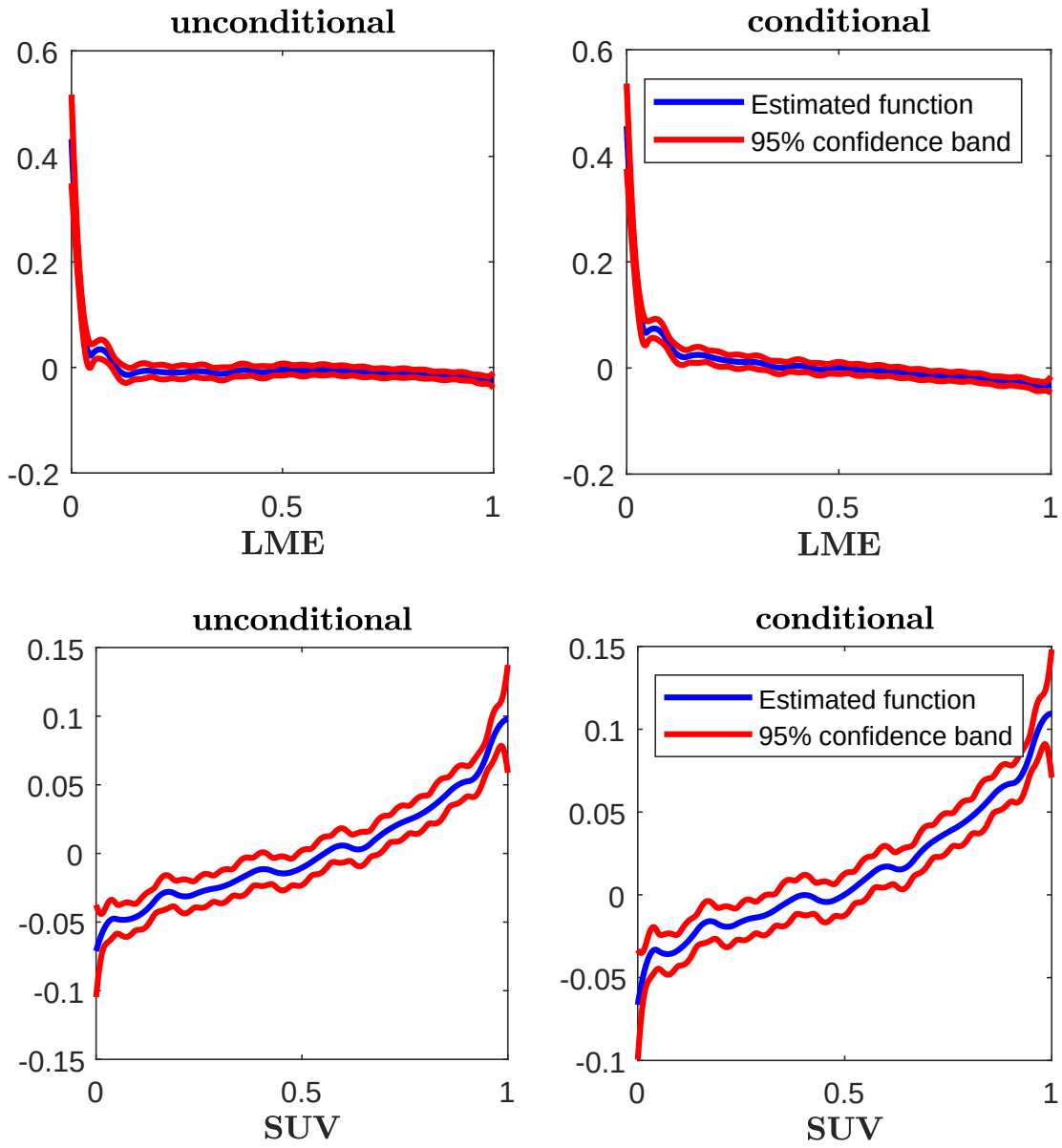
*Effect of normalized change in inventories (**IVC**) and net operating assets (**NOA**) on average returns (see equation (3)). The left panels report unconditional associations between a characteristic and returns and the right panels report associations conditional on all other selected characteristics. The sample period is January 1965 to June 2014. See Section A.1 in the online appendix for variable definitions.*

Figure 3: **Unconditional and Conditional Mean Function: Short-Term Reversal (r_{2-1}) and Closeness to 52 week's High (Rel to high price)**



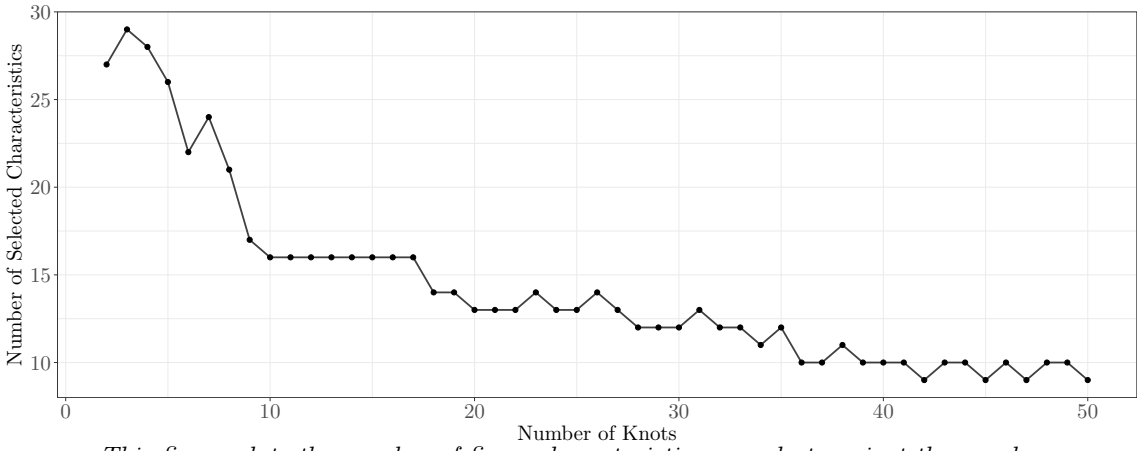
*Effect of normalized short-term reversal ($\mathbf{r_{2-1}}$) and closeness to 52 week's high (**Rel to high price**) on average returns (see equation (3)). The left panels report unconditional associations between a characteristic and returns, and the right panels report associations conditional on all other selected characteristics. The sample period is January 1965 to June 2014. See Section A.1 in the online appendix for variable definitions.*

Figure 4: **Unconditional and Conditional Mean Function: Size (LME) and Standard Unexplained Volume (SUV)**



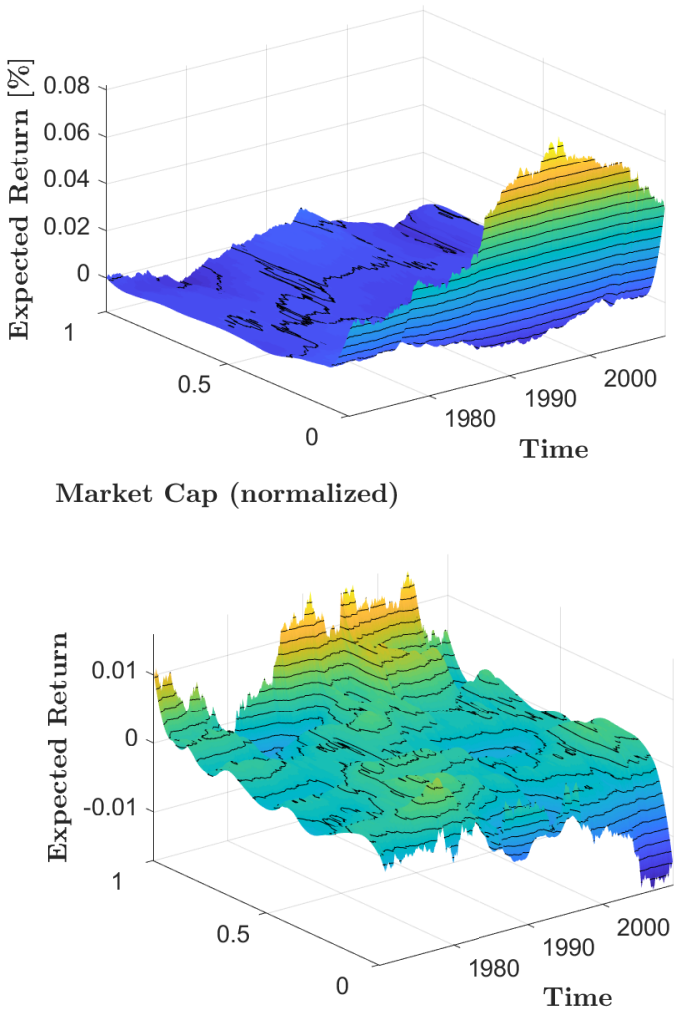
*Effect of normalized size (**LME**) and standard unexplained volume (**SUV**) on average returns (see equation (3)). The left panels report unconditional associations between a characteristic and returns, and the right panels report associations conditional on all other selected characteristics. The sample period is January 1965 to June 2014. See Section A.1 in the online appendix for variable definitions.*

Figure 5: **Number of Selected Characteristics versus Number of Interpolation Points**



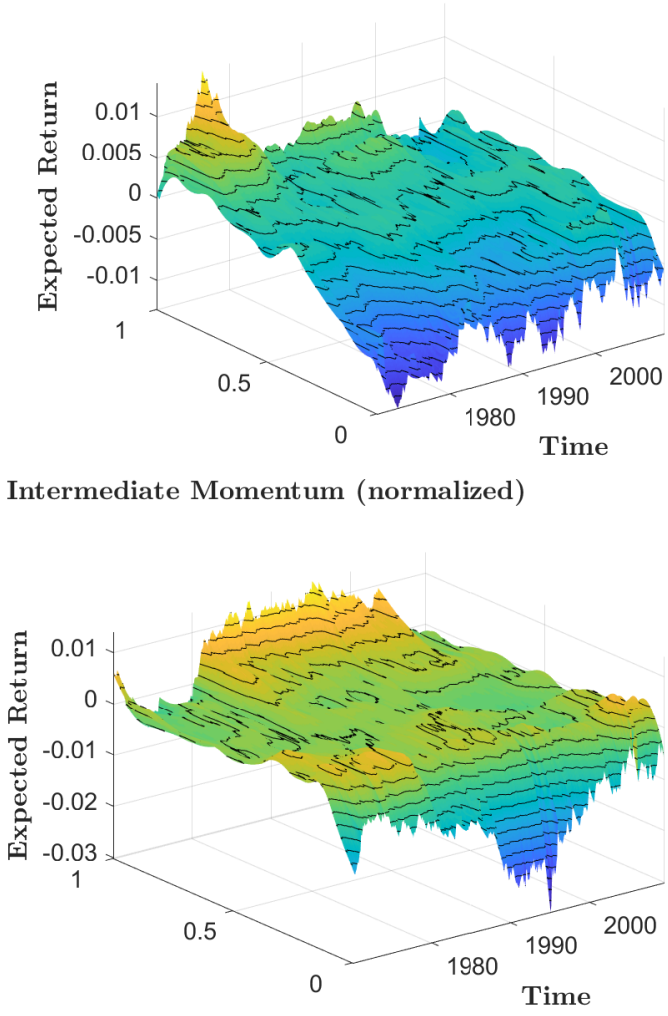
This figure plots the number of firms characteristics we select against the number of interpolation points in our baseline analysis. We use the adaptive group LASSO to select significant return predictors out of a universe of 63 characteristics during a sample period from 1965 to 2014. We detail the method in Section A.3.

Figure 6: **Time-varying Conditional Mean Function: Size (LME) and adjusted Profit Margin (PM_adj)**



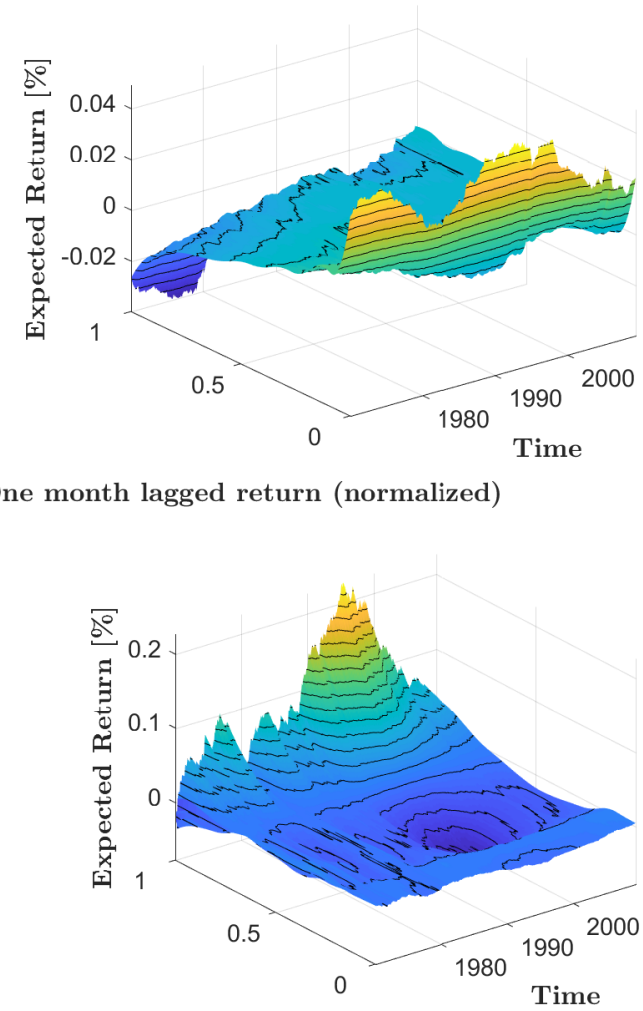
*Effect of normalized size (**LME**) and adjusted profit margin (**PM_adj**) on average returns over time (see equation (3)) conditional on all other selected characteristics. The sample period is January 1965 to June 2014. See Section A.1 in the online appendix for variable definitions.*

Figure 7: **Time-varying Conditional Mean Function: Intermediate Momentum (r_{12-7}) and Standard Momentum (r_{12-2})**



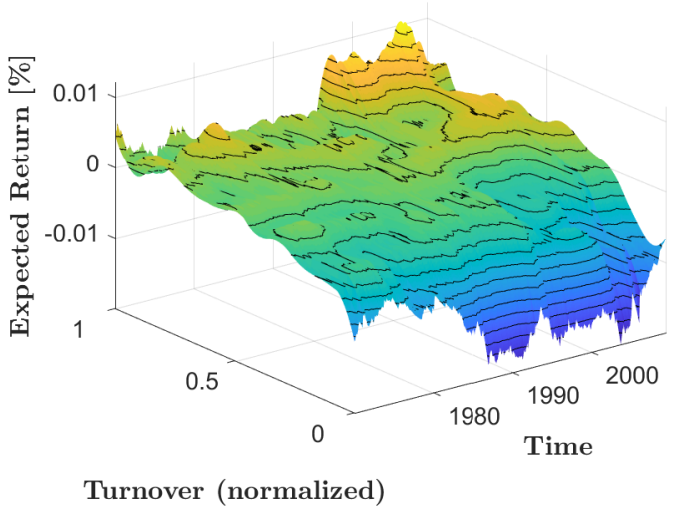
Effect of normalized intermediate momentum ($\mathbf{r_{12-7}}$) and standard momentum ($\mathbf{r_{12-2}}$) on average returns over time (see equation (3)) conditional on all other selected characteristics. The sample period is January 1965 to June 2014. See Section A.1 in the online appendix for variable definitions.

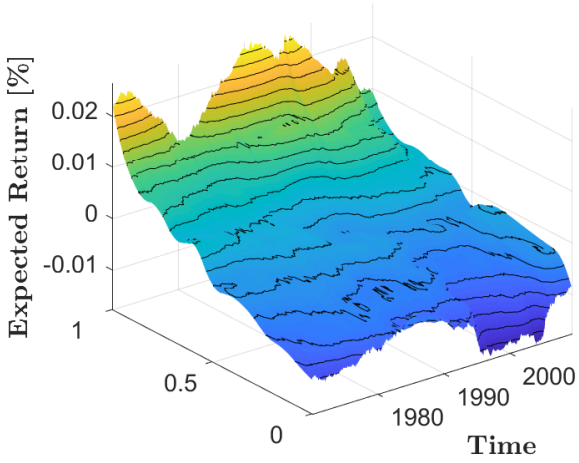
Figure 8: **Time-varying Conditional Mean Function: Short-Term Reversal (r_{2-1}) and Change in Shares Outstanding ($\Delta Shrout$)**



*Effect of normalized short-term reversal (**r36−13**) and the percentage change in shares outstanding (**ΔShrout**) on average returns over time (see equation (3)) conditional on all other selected characteristics. The sample period is January 1965 to June 2014. See Section A.1 in the online appendix for variable definitions.*

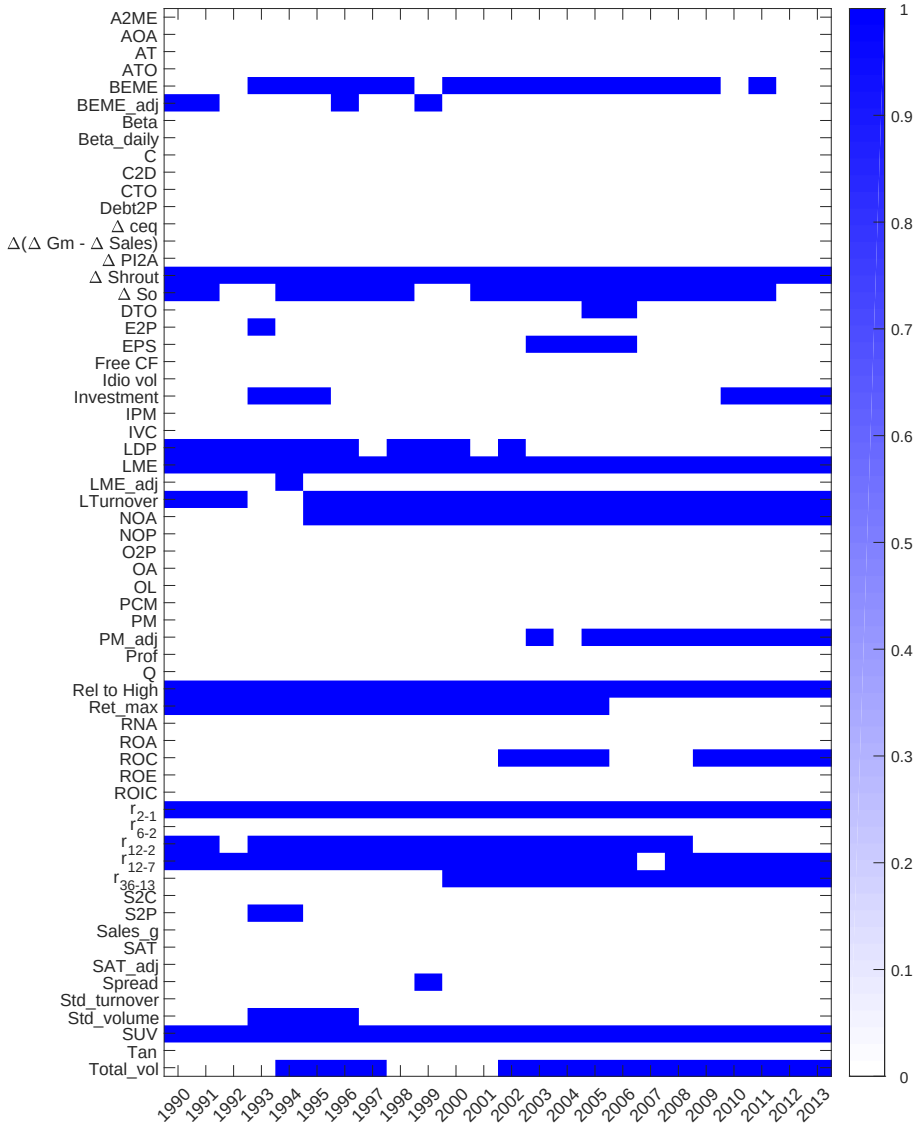
Figure 9: **Time-varying Conditional Mean Function: Turnover (Lturnover) and Standard Unexplained Volume (SUV)**





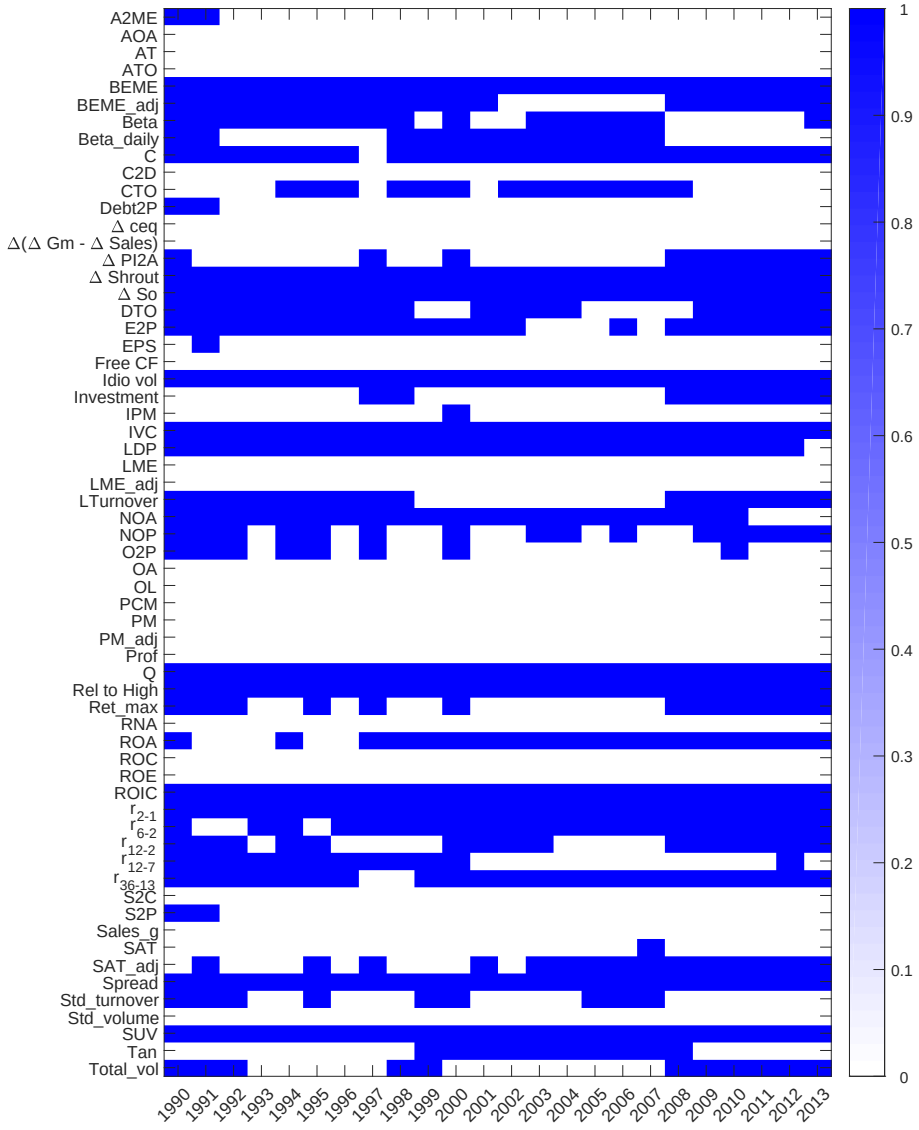
*Effect of normalized turnover (**Lturnover**) and standard unexplained volume (**SUV**) on average returns over time (see equation (3)) conditional on all other selected characteristics. The sample period is January 1965 to June 2014. See Section A.1 in the online appendix for variable definitions.*

Figure 10: **Selected Characteristics in Rolling Selection: Nonlinear Model**



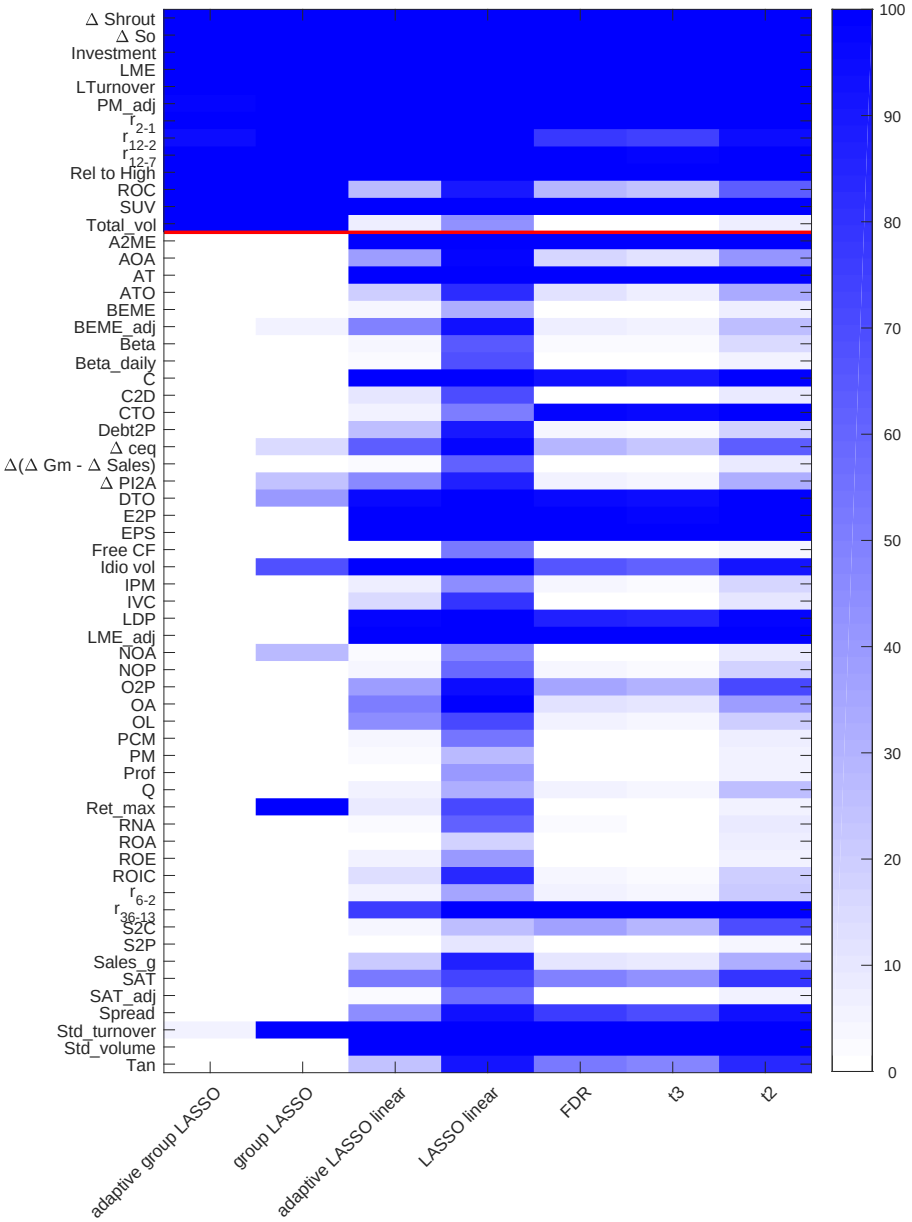
The figure graphically shows over time which characteristics from the universe of 62 firm characteristics we discuss in Section A.1 of the online appendix are selected by the nonlinear adaptive group LASSO. The first selection period is from January 1965 until December 1990. Subsequently, we roll forward the selection period by one year keeping the selection window constant. Blue indicates the characteristic is selected. The average number of selected characteristics is 14.13. The sample period is January 1965 to June 2014.

Figure 11: **Selected Characteristics in Rolling Selection: Linear Model**



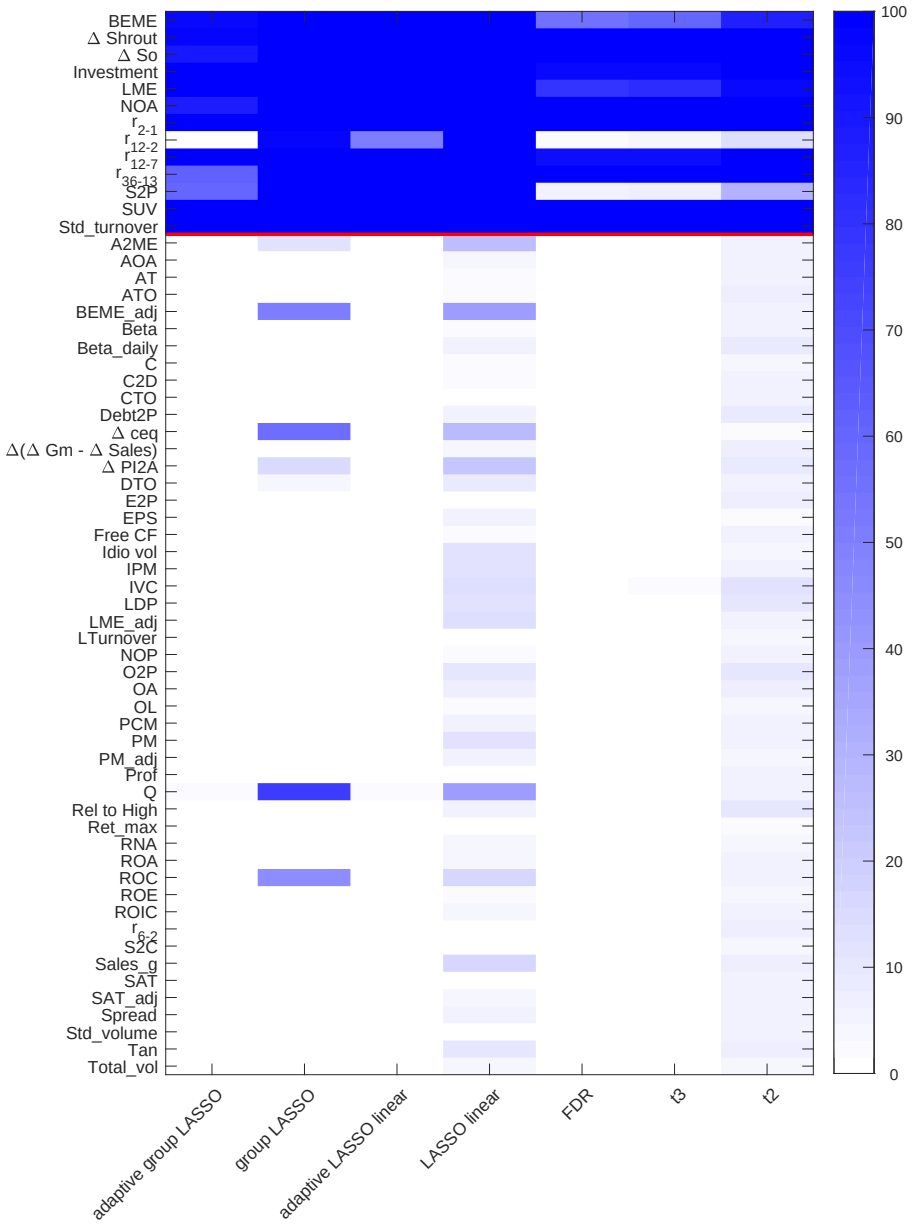
The figure graphically shows over time which characteristics from the universe of 62 firm characteristics we discuss in Section A.1 of the online appendix are selected by the linear adaptive LASSO. The first selection period is from January 1965 until December 1990. Subsequently, we roll forward the selection period by one year keeping the selection window constant. Blue indicates the characteristic is selected. The average number of selected characteristics is 26.58. The sample period is January 1965 to June 2014.

Figure 12: **Selected Characteristics in Simulations: Empirical Data-Generating Process**



The figure graphically shows for different model selection methods the frequency with which characteristics from the universe of 62 firm characteristics we discuss in Section A.1 of the online appendix are selected by each method. The darker the color, the more frequently a given selection method selects a given characteristic. The true model is nonlinear and consists of the 13 characteristics above the red vertical line. The average number of selected characteristics for the different methods across 500 simulations are: adaptive group LASSO: 12.99; group LASSO: 16.89; adaptive LASSO linear model: 29.36; LASSO linear model: 47.72; FDR: 27.30; t3: 26.40; t2: 33.75. The sample period is January 1965 to June 2014.

Figure 13: **Selected Characteristics in Simulations: Linear Data-Generating Process**



The figure graphically shows for different model selection methods the frequency with which characteristics from the universe of 62 firm characteristics we discuss in Section A.1 of the online appendix are selected by each method. The darker the color, the more frequently a given selection method selects a given characteristic. The true model is linear and consists of the 13 characteristics above the red vertical line. The average number of selected characteristics for the different methods across 500 simulations are: adaptive group LASSO: 10.98; group LASSO: 15.58; adaptive LASSO linear model: 12.60; LASSO linear model: 16.77; FDR: 10.45; t3: 10.64; t2: 14.15. The sample period is January 1965 to June 2014.

Table 1: **Firm Characteristics by Category**

This table lists the characteristics we consider in our empirical analysis by category. We report detailed variable definitions in Section A.1 of the online appendix. The sample period is January 1965 to June 2014.

	Past-returns:			**Value:**	
(1)	r_{2-1}	Return 1 month before prediction	(33)	A2ME	Total assets to Size
(2)	r_{6-2}	Return from 6 to 2 months before prediction	(34)	BEME	Book to market ratio
(3)	r_{12-2}	Return from 12 to 2 months before prediction	(35)	BEME_{adj}	BEME - mean BEME in Fama-French 48 industry
(4)	r_{12-7}	Return from 12 to 7 months before prediction	(36)	C	Cash to AT
(5)	r_{36-13}	Return from 36 to 13 months before prediction	(37)	C2D	Cash flow to total liabilities
			(38)	ΔSO	Log change in split-adjusted shares outstanding
	Investment:		(39)	Debt2P	Total debt to Size
(6)	Investment	% change in AT	(40)	E2P	Income before extraordinary items to Size
(7)	ΔCEQ	% change in BE	(41)	Free CF	Free cash flow to BE
(8)	ΔPI2A	Change in PP&E and inventory over lagged AT	(42)	LDP	Trailing 12-months dividends to price
(9)	ΔShrout	% change in shares outstanding	(43)	NOP	Net payouts to Size
(10)	IVC	Change in inventory over average AT	(44)	O2P	Operating payouts to market cap
(11)	NOA	Net-operating assets over lagged AT	(45)	Q	Tobin's Q
			(46)	S2P	Sales to price
	Profitability:		(47)	Sales_g	Sales growth
(12)	ATO	Sales to lagged net operating assets			
(13)	CTO	Sales to lagged total assets		**Trading frictions:**	
(14)	Δ(ΔGM-ΔSales)	Δ(% change in gross margin and % change in sales)	(48)	AT	Total assets
(15)	EPS	Earnings per share	(49)	Beta	Correlation × ratio of vols
(16)	IPM	Pre-tax income over sales	(50)	Beta daily	CAPM beta using daily returns
(17)	PCM	Sales minus costs of goods sold to sales	(51)	DTO	De-trended Turnover - market Turnover
(18)	PM	OI after depreciation over sales	(52)	Idio vol	Idio vol of Fama-French 3 factor model
(19)	PM_adj	Profit margin - mean PM in Fama-French 48 industry	(53)	LME	Price times shares outstanding
(20)	Prof	Gross profitability over BE	(54)	LME_adj	Size - mean size in Fama-French 48 industry
(21)	RNA	OI after depreciation to lagged net operating assets	(55)	Lturnover	Last month's volume to shares outstanding
(22)	ROA	Income before extraordinary items to lagged AT	(56)	Rel_to_high_price	Price to 52 week high price
(23)	ROC	Size + longterm debt - total assets to cash	(57)	Ret_max	Maximum daily return
(24)	ROE	Income before extraordinary items to lagged BE	(58)	Spread	Average daily bid-ask spread
(25)	ROIC	Return on invested capital	(59)	Std turnover	Standard deviation of daily turnover
(26)	S2C	Sales to cash	(60)	Std volume	Standard deviation of daily volume
(27)	SAT	Sales to total assets	(61)	SUV	Standard unexplained volume
(28)	SAT_adj	SAT - mean SAT in Fama-French 48 industry	(62)	Total vol	Standard deviation of daily returns
	Intangibles:				
(29)	AOA	Absolute value of operating accruals			
(30)	OL	Costs of goods solds + SG&A to total assets			
(31)	Tan	Tangibility			
(32)	OA	Operating accruals			

Table 2: **Descriptive Statistics for Firm Characteristics**

This table reports average returns, medians, and time series standard deviations for the firm characteristics discussed in Section A.1 of the online appendix. Frequency is the frequency at which the firm characteristics varies. m is monthly and y is yearly. The sample period is January 1965 to June 2014.

	Mean	Median	Std	Freq		Mean	Median	Std	Freq
Past-returns:					**Value:**				
r_{2-1}	0.01	0.00	(0.13)	m	A2ME	3.04	1.62	(5.75)	y
r_{6-2}	0.06	0.03	(0.31)	m	BEME	0.94	0.77	(0.80)	y
r_{12-2}	0.14	0.07	(0.51)	m	BEME_{adj}	0.01	−0.13	(0.77)	m
r_{12-7}	0.08	0.04	(0.34)	m	C	0.13	0.07	(0.15)	y
r_{36-13}	0.35	0.17	(0.96)	m	C2D	0.17	0.17	(1.26)	y
					ΔSO	0.03	0.00	(0.12)	y
Investment:					Debt2P	0.86	0.34	(2.37)	y
Investment	0.14	0.08	(0.44)	y	E2P	0.01	0.07	(0.36)	y
ΔCEQ	0.18	0.06	(2.00)	y	Free CF	−0.23	0.05	(9.70)	y
ΔPI2A	0.09	0.06	(0.22)	y	LDP	0.02	0.01	(0.05)	m
ΔShrout	0.01	0.00	(0.10)	m	NOP	0.01	0.01	(0.12)	y
IVC	0.02	0.01	(0.06)	y	O2P	0.03	0.02	(0.13)	y
NOA	0.67	0.67	(0.38)	y	Q	1.63	1.20	(1.47)	y
					S2P	2.75	1.60	(4.38)	y
Profitability:					Sales_g	0.37	0.09	(9.81)	y
ATO	2.52	1.94	(21.51)	y					
CTO	1.35	1.18	(1.11)	y	**Trading frictions:**				
Δ(ΔGM-ΔSales)	−0.29	0.00	(17.42)	y	AT	2,906.94	243.22	(19,820.90)	y
EPS	1.76	1.19	(21.66)	y	Beta	1.05	0.99	(0.55)	m
IPM	−1.01	0.07	(35.76)	m	Beta daily	0.89	0.81	(1.52)	m
PCM	−0.60	0.32	(34.01)	y	DTO	0.00	0.00	(0.01)	m
PM	−0.99	0.08	(35.90)	y	Idio vol	0.03	0.02	(0.02)	m
PM_adj	0.39	0.09	(35.79)	m	LME	1,562.03	166.44	(7,046.08)	m
Prof	1.01	0.64	(11.50)	y	LME_adj	287.02	−683.49	(6,947.60)	m
RNA	0.21	0.14	(6.79)	y	Lturnover	0.08	0.05	(0.12)	m
ROA	0.03	0.04	(0.15)	y	Rel_to_high_price	0.75	0.79	(0.18)	m
ROC	−6.86	−1.44	(332.86)	m	Ret max	0.07	0.05	(0.07)	m
ROE	0.06	0.10	(1.42)	y	Spread	0.03	0.02	(0.04)	m
ROIC	0.06	0.07	(0.12)	y	Std turnover	0.31	0.16	(0.68)	m
S2C	84.77	15.32	(970.18)	y	Std volume	162.84	33.51	(583.97)	m
SAT	1.21	1.08	(0.93)	y	SUV	0.22	−0.15	(2.39)	m
SAT_adj	0.02	−0.06	(0.74)	m	Total vol	0.03	0.02	(0.02)	m
Intangibles:									
AOA	5.23	0.07	(285.41)	y					
OL	1.10	0.95	(0.91)	y					
Tan	0.54	0.55	(0.12)	y					
OA	−0.47	−0.03	(78.52)	y					

Table 3: **Returns of 10 Portfolios Sorted on Characteristics**

This table reports equally-weighted returns with standard errors in parentheses for ten portfolios sorted on firm characteristics discussed in Section A.1 of the online appendix. The sample period is July 1965 to June 2014.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P10-P1
Past-return based:											
r_{2-1}	34.36	20.48	17.55	16.76	15.92	15.07	13.21	12.21	10.34	4.34	−13.43
	(3.22)	(4.38)	(2.62)	(2.80)	(2.52)	(3.36)	(2.71)	(2.56)	(2.97)	(2.44)	(2.97)
r_{6-2}	15.13	14.70	14.62	14.99	15.74	15.54	15.41	16.31	17.47	20.47	5.34
	(3.33)	(4.46)	(2.61)	(3.46)	(2.56)	(2.96)	(2.79)	(2.73)	(2.51)	(2.47)	(3.33)
r_{12-2}	13.90	12.57	13.02	14.02	14.28	15.14	16.49	18.38	19.73	22.82	8.92
	(3.45)	(4.49)	(2.55)	(3.04)	(3.44)	(2.51)	(2.57)	(2.69)	(2.48)	(2.83)	(3.46)
r_{12-7}	12.37	12.22	13.58	14.80	15.96	15.86	16.72	17.97	19.67	21.28	8.92
	(3.44)	(3.96)	(2.57)	(2.57)	(2.66)	(2.94)	(2.53)	(2.86)	(2.69)	(3.20)	(2.43)
r_{36-13}	23.45	19.32	17.56	15.91	15.26	14.97	15.23	13.69	13.54	11.47	−11.99
	(3.37)	(4.26)	(2.54)	(2.99)	(2.70)	(3.40)	(2.56)	(2.48)	(2.49)	(2.81)	(2.86)
Investment:											
Investment	22.50	20.08	18.59	17.38	15.74	15.49	15.09	14.20	12.85	8.64	−13.87
	(3.45)	(3.88)	(2.48)	(2.78)	(3.07)	(2.49)	(3.08)	(2.72)	(2.62)	(2.49)	(1.88)
ΔCEQ	19.98	19.73	17.47	16.34	17.62	15.49	15.26	15.38	13.76	9.55	−10.43
	(3.46)	(3.88)	(2.44)	(2.62)	(3.07)	(2.80)	(2.64)	(2.52)	(3.14)	(2.51)	(1.89)
ΔPI2A	20.93	18.65	17.95	16.50	16.31	16.18	15.37	14.99	13.37	10.28	−10.65
	(3.30)	(3.46)	(2.68)	(2.73)	(2.71)	(3.02)	(2.94)	(2.66)	(2.83)	(2.61)	(1.60)
ΔShrout	16.27	15.86	15.54	15.70	14.78	15.51	15.15	15.04	16.76	19.78	3.51
	(2.71)	(3.01)	(2.85)	(2.83)	(2.90)	(2.89)	(2.86)	(2.89)	(2.86)	(2.89)	(1.22)
IVC	20.84	18.43	16.24	16.69	15.18	15.76	15.36	15.33	14.04	12.59	−8.25
	(3.33)	(3.44)	(2.69)	(2.69)	(3.10)	(2.84)	(2.70)	(2.91)	(2.55)	(2.61)	(1.35)
NOA	18.36	17.88	17.62	17.31	18.30	17.09	16.04	14.64	13.94	9.47	−8.89
	(3.16)	(2.94)	(2.83)	(2.85)	(2.95)	(2.73)	(2.94)	(2.73)	(2.76)	(2.95)	(1.49)

continued on next page

Table 3: **Continued from Previous Page**

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P10-P1
Profitability:											
ATO	14.78	15.14	15.06	16.48	17.14	16.93	17.04	16.44	15.90	15.66	0.88
	(3.05)	(2.46)	(2.85)	(2.73)	(3.02)	(3.15)	(2.99)	(2.95)	(2.89)	(2.89)	(1.70)
CTO	14.72	14.10	15.99	16.89	16.44	16.65	17.58	16.36	16.16	15.69	0.97
	(3.00)	(2.57)	(2.95)	(2.90)	(3.06)	(2.69)	(3.01)	(2.98)	(3.00)	(2.93)	(1.65)
Δ(ΔGm-ΔSales)	14.31	16.04	16.21	16.58	15.82	16.47	15.78	16.79	15.98	16.58	2.28
	(3.26)	(3.45)	(2.59)	(2.71)	(2.62)	(2.79)	(2.64)	(2.83)	(2.89)	(3.03)	(1.17)
EPS	18.60	17.77	19.27	17.03	15.04	14.60	14.02	14.85	14.72	14.57	−4.03
	(2.27)	(4.13)	(2.85)	(2.39)	(2.56)	(3.57)	(3.09)	(2.71)	(2.29)	(3.94)	(2.99)
IPM	17.72	18.78	18.39	17.84	16.75	15.15	15.00	14.02	13.57	13.32	−4.40
	(2.36)	(4.34)	(2.79)	(2.73)	(2.63)	(3.19)	(3.68)	(2.97)	(2.49)	(2.36)	(3.08)
PCM	17.88	15.88	15.68	15.46	16.27	15.85	15.81	15.52	15.52	16.53	−1.35
	(2.87)	(3.38)	(2.90)	(2.80)	(2.84)	(2.85)	(2.71)	(2.94)	(2.92)	(2.71)	(1.52)
PM	17.24	19.35	18.03	18.04	16.30	15.69	14.88	13.79	13.81	13.44	−3.79
	(2.27)	(4.29)	(2.84)	(2.72)	(2.57)	(3.18)	(2.77)	(3.65)	(2.37)	(2.97)	(3.27)
PM_adj	16.98	17.27	17.06	15.64	15.51	14.17	15.89	15.28	16.37	16.31	−0.67
	(3.09)	(3.49)	(2.64)	(2.67)	(2.91)	(2.77)	(3.16)	(2.70)	(3.16)	(3.00)	(1.91)
Prof	14.82	13.64	15.03	15.31	15.43	16.44	16.39	16.93	18.35	18.14	3.32
	(3.21)	(3.18)	(2.75)	(2.73)	(2.89)	(2.49)	(2.94)	(2.66)	(3.03)	(3.09)	(1.67)
RNA	17.31	18.47	16.60	16.43	17.08	16.47	15.53	15.20	13.70	13.73	−3.58
	(2.93)	(3.87)	(2.67)	(2.92)	(2.69)	(2.74)	(2.68)	(2.94)	(3.14)	(2.65)	(2.03)
ROA	18.26	19.82	17.42	15.11	15.83	15.61	16.85	15.10	13.91	12.64	−5.62
	(2.87)	(4.42)	(2.51)	(2.52)	(2.54)	(2.67)	(3.68)	(2.65)	(2.97)	(2.71)	(2.72)
ROC	16.90	16.11	18.15	19.70	19.18	18.40	17.15	13.85	11.65	9.58	−7.32
	(2.85)	(2.69)	(2.93)	(3.20)	(2.71)	(3.10)	(2.99)	(2.84)	(3.11)	(2.95)	(1.71)
ROE	17.96	19.21	17.29	16.92	16.10	15.19	15.07	15.27	14.42	13.10	−4.86
	(3.03)	(4.42)	(2.45)	(2.67)	(2.44)	(2.95)	(2.47)	(2.63)	(3.59)	(2.77)	(2.69)
ROIC	18.59	17.06	16.88	15.27	16.09	16.15	16.00	15.58	14.49	14.35	−4.24
	(4.24)	(2.63)	(2.68)	(2.91)	(2.65)	(2.66)	(2.71)	(2.74)	(3.46)	(2.75)	(2.81)
S2C	14.96	16.06	15.39	16.15	17.13	15.96	16.05	15.82	16.48	16.51	1.55
	(2.78)	(3.10)	(2.92)	(2.86)	(2.84)	(2.93)	(2.89)	(2.86)	(2.91)	(2.86)	(1.68)
SAT	13.68	13.44	14.16	15.92	15.67	16.78	16.94	17.52	17.64	18.76	5.07
	(2.92)	(2.56)	(2.99)	(2.85)	(3.04)	(2.71)	(3.02)	(3.00)	(3.03)	(3.01)	(1.60)
SAT_adj	13.84	15.02	15.20	14.49	16.08	15.21	15.75	17.48	18.61	18.82	4.98
	(2.94)	(3.15)	(2.60)	(2.76)	(2.93)	(2.97)	(2.79)	(3.15)	(2.94)	(2.67)	(1.09)

continued on next page

Table 3: **Continued from Previous Page**

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P10-P1
Intangibles:											
AOA	15.73	15.38	16.42	16.03	17.42	16.85	17.24	16.42	16.24	12.89	−2.84
	(2.74)	(3.55)	(2.70)	(2.61)	(2.99)	(2.59)	(2.83)	(2.66)	(2.74)	(3.29)	(1.42)
OL	14.05	13.33	13.81	15.13	15.74	16.00	17.53	17.75	18.06	19.07	5.02
	(3.01)	(2.42)	(3.08)	(2.82)	(2.62)	(3.15)	(3.03)	(3.09)	(2.97)	(3.10)	(1.90)
Tan	16.05	14.63	15.27	15.17	16.36	16.21	16.56	16.85	16.19	17.16	1.11
	(3.38)	(3.03)	(2.73)	(2.82)	(2.93)	(2.68)	(2.87)	(2.81)	(2.64)	(3.12)	(1.91)
OA	17.36	17.69	17.66	17.24	16.67	16.46	15.90	16.09	14.65	10.95	−6.41
	(3.27)	(3.46)	(2.57)	(2.63)	(2.96)	(2.73)	(2.83)	(2.76)	(2.63)	(2.90)	(1.31)
Value:											
A2ME	9.22	13.04	14.19	15.57	16.80	17.34	18.24	18.48	18.72	19.06	9.84
	(3.31)	(3.22)	(2.80)	(2.80)	(2.92)	(2.78)	(2.87)	(2.92)	(3.06)	(2.85)	(2.69)
BEME	9.20	12.47	13.32	14.51	15.66	16.97	16.77	18.18	20.21	23.24	14.04
	(3.30)	(3.32)	(2.72)	(2.95)	(2.70)	(2.74)	(2.71)	(2.91)	(3.10)	(2.85)	(2.27)
BEME_adj	11.19	12.22	13.42	13.57	14.84	16.12	16.67	18.00	20.99	23.42	12.22
	(2.86)	(3.37)	(2.75)	(2.76)	(2.94)	(2.84)	(2.75)	(2.85)	(2.98)	(2.89)	(1.81)
C	15.28	14.45	15.37	15.21	16.60	16.72	16.65	16.72	16.29	17.20	1.92
	(3.40)	(2.61)	(2.82)	(2.78)	(2.82)	(3.22)	(2.74)	(2.95)	(3.03)	(2.89)	(2.13)
C2D	18.12	17.78	14.59	15.06	16.15	16.10	16.79	16.12	15.28	14.49	−3.63
	(2.63)	(4.41)	(2.70)	(2.76)	(2.69)	(2.75)	(3.50)	(2.72)	(2.72)	(2.66)	(2.64)
ΔSO	19.56	17.68	17.37	17.17	16.77	16.59	16.38	15.98	13.07	10.02	−9.53
	(3.39)	(2.64)	(2.77)	(3.06)	(3.00)	(3.22)	(2.62)	(2.85)	(2.76)	(2.70)	(1.74)
Debt2P	16.17	14.39	13.50	14.71	16.45	16.57	17.04	15.97	16.71	18.86	2.68
	(2.93)	(3.35)	(2.78)	(2.74)	(2.80)	(3.18)	(2.78)	(2.90)	(3.00)	(2.85)	(2.01)
E2P	20.17	13.80	12.39	14.17	14.44	15.60	15.30	16.09	17.98	20.24	0.06
	(2.91)	(4.25)	(2.68)	(2.95)	(3.20)	(2.59)	(3.47)	(2.57)	(2.49)	(2.49)	(2.47)
Free CF	14.96	15.64	16.44	15.62	16.29	16.05	15.62	16.58	16.61	16.70	1.74
	(2.80)	(4.04)	(2.72)	(2.78)	(2.63)	(2.58)	(3.38)	(2.96)	(2.62)	(2.57)	(2.25)
LDP	18.76	17.92	16.13	14.32	15.06	15.74	14.56	14.67	16.15	17.06	−1.70
	(2.18)	(3.68)	(3.09)	(2.28)	(3.02)	(3.54)	(2.62)	(2.40)	(3.21)	(3.70)	(2.48)
NOP	13.11	15.84	16.80	15.44	16.51	15.89	15.51	16.31	16.30	18.80	5.69
	(2.49)	(3.71)	(2.86)	(2.36)	(2.29)	(3.18)	(2.65)	(2.46)	(3.84)	(3.56)	(2.13)
O2P	15.82	18.75	15.44	14.64	14.43	14.92	15.51	16.35	16.32	18.29	2.47
	(2.58)	(3.51)	(2.87)	(2.39)	(2.26)	(3.26)	(2.63)	(2.46)	(3.78)	(3.49)	(1.71)
Q	22.62	19.76	18.04	17.29	16.36	15.91	14.63	14.26	12.49	9.15	−13.47
	(3.32)	(3.08)	(2.75)	(2.77)	(2.83)	(2.98)	(2.73)	(3.08)	(3.00)	(2.93)	(2.14)
S2P	10.10	11.35	12.86	14.44	15.85	16.36	17.92	19.19	20.14	22.31	12.21
	(3.42)	(3.43)	(2.66)	(2.61)	(2.65)	(3.22)	(2.97)	(2.92)	(2.76)	(2.76)	(2.45)
Sales_g	19.80	18.47	17.09	16.91	16.35	16.04	16.56	15.25	13.77	10.34	−9.47
	(3.69)	(3.44)	(2.47)	(2.54)	(2.87)	(2.54)	(2.70)	(3.13)	(2.92)	(2.67)	(1.56)

Table 3: **Continued from Previous Page**

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P10-P1
Trading frictions:											
AT	21.00	19.48	17.44	15.59	16.33	15.52	14.89	14.05	13.50	12.62	−8.38
	(2.31)	(3.79)	(2.95)	(2.71)	(3.31)	(2.76)	(2.54)	(2.85)	(3.57)	(3.14)	(3.13)
Beta	15.85	16.45	17.03	16.82	16.58	16.42	15.99	15.56	15.13	14.74	−1.11
	(1.91)	(4.77)	(2.58)	(3.43)	(2.85)	(3.11)	(2.46)	(2.10)	(3.91)	(2.29)	(3.66)
Beta daily	18.26	15.99	15.68	15.61	15.43	15.80	16.60	16.64	16.12	14.37	−3.89
	(3.17)	(4.25)	(2.49)	(2.47)	(3.13)	(2.84)	(2.34)	(2.39)	(3.51)	(2.64)	(2.34)
DTO	12.47	13.49	11.50	11.50	12.16	13.66	17.59	19.83	22.33	25.80	13.33
	(3.58)	(3.37)	(2.48)	(2.53)	(3.03)	(2.78)	(2.87)	(2.73)	(3.16)	(2.60)	(1.53)
Idio vol	12.54	14.76	15.35	16.67	18.19	17.34	17.04	16.69	16.49	15.59	3.06
	(4.48)	(1.77)	(2.76)	(3.00)	(2.12)	(2.35)	(3.54)	(3.26)	(3.97)	(2.54)	(3.67)
LME	31.91	16.09	14.97	14.77	14.50	15.36	14.83	13.81	12.66	11.18	−20.73
	(2.21)	(4.05)	(3.08)	(2.70)	(3.44)	(2.48)	(2.88)	(3.21)	(3.01)	(3.22)	(3.48)
LME_adj	19.83	16.76	16.56	18.08	17.49	16.91	16.01	14.77	12.68	11.42	−8.41
	(2.24)	(3.30)	(3.31)	(2.98)	(2.64)	(3.03)	(3.20)	(3.17)	(2.84)	(3.11)	(2.28)
Lturnover	12.04	14.31	15.72	16.17	17.04	17.74	17.53	17.37	17.62	15.16	3.13
	(2.11)	(4.10)	(2.79)	(3.28)	(2.56)	(2.88)	(3.67)	(3.04)	(2.64)	(2.40)	(3.08)
Rel_to_high_price	26.19	15.27	13.47	13.78	14.78	15.26	15.74	15.30	16.21	14.19	−12.00
	(2.11)	(4.85)	(2.85)	(2.50)	(3.78)	(3.06)	(2.66)	(3.37)	(2.26)	(2.39)	(4.01)
Ret max	15.10	16.45	17.01	16.66	17.74	17.04	17.28	16.93	15.02	11.50	−3.60
	(4.23)	(1.93)	(2.74)	(3.21)	(2.37)	(3.52)	(2.96)	(2.18)	(3.82)	(2.58)	(3.25)
Spread	13.32	14.61	15.52	15.54	16.42	16.30	15.96	15.40	16.01	21.33	8.01
	(2.40)	(3.97)	(2.85)	(3.32)	(2.68)	(3.16)	(3.54)	(2.52)	(2.42)	(3.04)	(2.99)
Std turnover	10.69	13.10	14.85	16.06	16.88	17.56	18.14	17.46	18.34	17.60	6.91
	(2.01)	(3.86)	(2.80)	(3.16)	(2.33)	(3.34)	(3.00)	(3.59)	(2.51)	(2.65)	(2.76)
Std volume	15.88	17.20	17.06	17.75	16.74	16.46	16.64	14.94	15.07	12.89	−2.99
	(2.35)	(3.05)	(3.04)	(3.04)	(3.02)	(2.72)	(3.06)	(3.14)	(3.11)	(2.91)	(2.26)
SUV	6.55	9.26	10.94	12.87	13.47	15.99	17.69	20.27	23.73	29.62	23.07
	(3.22)	(2.84)	(2.72)	(2.78)	(2.83)	(2.83)	(2.90)	(2.78)	(3.11)	(2.97)	(1.87)
Total vol	12.85	14.45	15.60	16.72	18.03	17.77	16.98	17.52	15.43	15.31	2.46
	(4.57)	(1.71)	(2.73)	(2.06)	(2.50)	(3.64)	(3.28)	(2.28)	(2.98)	(4.04)	(3.75)

Table 4: **Fama & French Three-Factor alphas for Characteristic-sorted Portfolios**

This table reports Fama&French three-factor alphas of long-short portfolios sorted on the characteristics we describe in Section A.1 of the online appendix with standard errors in parentheses and t-statistics. The sample period is July 1965 to June 2015.

	α_{FF3}	SE	t-stat		α_{FF3}	SE	t-stat
Past-return based:				**Value:**			
r_{2-1}	−26.48	(2.94)	-8.99	A2ME	2.75	(1.93)	1.42
r_{6-2}	9.13	(3.34)	2.73	BEME	7.80	(1.56)	5.00
r_{12-2}	12.73	(3.48)	3.66	BEME_adj	8.60	(1.56)	5.50
r_{12-7}	10.79	(2.46)	4.39	C	4.48	(1.69)	2.65
r_{36-13}	−6.68	(2.59)	-2.58	C2D	0.23	(2.23)	0.10
				ΔSO	−9.14	(1.54)	-5.95
Investment:				Debt2P	−3.21	(1.63)	-1.97
Investment	−11.85	(1.76)	-6.74	E2P	0.56	(2.19)	0.25
ΔCEQ	−8.02	(1.78)	-4.50	Free CF	3.29	(2.04)	1.61
ΔPI2A	−9.32	(1.53)	-6.07	LDP	0.99	(1.68)	0.59
ΔShrout	3.57	(1.17)	3.04	NOP	5.74	(1.61)	3.56
IVC	−7.30	(1.34)	-5.45	O2P	2.74	(1.34)	2.04
NOA	−9.56	(1.50)	-6.36	Q	−8.09	(1.50)	-5.40
				S2P	5.44	(1.91)	2.85
Profitability:				Sales_g	−7.52	(1.50)	-5.02
ATO	0.85	(1.45)	0.59				
CTO	0.57	(1.52)	0.37	**Trading frictions:**			
Δ(ΔGm-ΔSales	3.15	(1.18)	2.66	AT	−7.01	(2.27)	-3.09
EPS	1.01	(2.14)	0.47	Beta	−7.83	(2.38)	-3.29
IPM	−0.41	(2.47)	-0.17	Beta daily	−6.39	(1.97)	-3.24
PCM	1.87	(1.36)	1.37	DTO	13.08	(1.56)	8.37
PM	−0.88	(2.59)	-0.34	Idio vol	−2.92	(2.74)	-1.06
PM_adj	3.96	(1.70)	2.32	LME	−15.30	(2.76)	-5.53
Prof	1.73	(1.69)	1.03	LME_adj	−4.76	(1.51)	-3.14
RNA	−0.37	(1.81)	-0.21	Lturnover	0.44	(2.03)	0.22
ROA	−1.70	(2.36)	-0.72	Rel_to_high_price	−5.46	(3.54)	-1.54
ROC	−4.07	(1.37)	-2.96	Ret max	−8.41	(2.40)	-3.51
ROE	−1.89	(2.39)	-0.79	Spread	3.06	(2.74)	1.12
ROIC	−1.75	(2.49)	-0.70	Std turnover	4.03	(1.79)	2.26
S2C	−0.45	(1.56)	-0.29	Std volume	−3.55	(1.85)	-1.92
SAT	4.43	(1.52)	2.91	SUV	21.88	(1.89)	11.59
SAT_adj	5.36	(1.10)	4.88	Total vol	−3.94	(2.74)	-1.43
Intangibles:							
AOA	−4.34	(1.24)	-3.48				
OL	4.01	(1.67)	2.39				
Tan	4.29	(1.67)	2.58				
OA	−5.92	(1.34)	-4.41				

Table 5: **Selected Characteristics in Nonparametric Model**

This table reports the selected characteristics from the universe of 62 firm characteristics we discuss in Section A.1 of the online appendix for different numbers of knots, the sample size, and in-sample Sharpe ratios of an equally-weighted hedge portfolio going long the 10% of stocks with highest predicted returns and shorting the 10% of stocks with lowest predicted returns. q indicates the size percentile of NYSE firms. The sample period is January 1965 to June 2014 unless otherwise specified.

Firms		All	All	All	$Size > q_{10}$	$Size > q_{20}$	$Size > q_{20}$	All	All
Sample		Full	Full	Full	Full	Full	Full	1965-1990	1991-2014
Knots		20	15	25	15	15	10	15	15
Sample Size		1,629,155	1,629,155	1,629,155	959,757	763,850	763,850	603,658	1,025,497
# Selected		13	16	13	10	9	11	11	14
Sharpe Ratio		3.15	3.05	3.16	2.53	2.25	2.37	3.99	2.66
Characteristics	# Selected	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
BEME	2		BEME					BEME	
ΔShrout	8	ΔShrout	ΔShrout	ΔShrout	ΔShrout	ΔShrout	ΔShrout	ΔShrout	ΔShrout
ΔSO	7	ΔSO	ΔSO	ΔSO	ΔSO	ΔSO	ΔSO		ΔSO
Investment	5	Investment	Investment	Investment	Investment				Investment
LDP	1							LDP	
LME	5	LME	LME	LME				LME	LME
Lturnover	4	Lturnover	Lturnover	Lturnover					Lturnover
NOA	2		NOA						NOA
NOP	1							NOP	
PM_adj	4	PM_adj	PM_adj	PM_adj					PM_adj
r_{2-1}	8	r_{2-1}	r_{2-1}	r_{2-1}	r_{2-1}	r_{2-1}	r_{2-1}	r_{2-1}	r_{2-1}
r_{6-2}	2					r_{6-2}	r_{6-2}		
r_{12-2}	6	r_{12-2}	r_{12-2}	r_{12-2}	r_{12-2}		r_{12-2}	r_{12-2}	
r_{12-7}	7	r_{12-7}	r_{12-7}	r_{12-7}	r_{12-7}	r_{12-7}		r_{12-7}	r_{12-7}
r_{36-13}	3		r_{36-13}				r_{36-13}		r_{36-13}
Rel_to_high_price	6	Rel_to_high	Rel_to_high	Rel_to_high			Rel_to_high	Rel_to_high	Rel_to_high
Ret max	1							Ret max	
ROC	7	ROC	ROC	ROC	ROC	ROC	ROC		ROC
S2P	3				S2P	S2P	S2P		
SUV	8	SUV	SUV	SUV	SUV	SUV	SUV	SUV	SUV
Total vol	7	Total vol	Total vol	Total vol	Total vol	Total vol	Total vol		Total vol

Never selected: A2ME, AOA, AT, ATO, BEME_adj, Beta, Beta daily, C, C2D, CTO, Δ CEQ, Δ(ΔGm-Δ Sales), Δ PI2A, Debt2P, DTO, E2P, EPS, Free CF, Idio vol, IPM, IVC, LME_adj O2P, OL, OA, PCM, PM, Prof, Q, RNA, ROA, ROE, ROIC, S2C, Sales_g, SAT, SAT_adj, Spread, Std turnover, Std volume, Tan

Table 6: **Selected Characteristics in Nonparametric Model: Size Interactions**

This table reports the selected characteristics from the universe of 62 firm characteristics we discuss in Section A.1 of the online appendix for different numbers of knots, the sample size, and in-sample Sharpe ratios of an equally-weighted hedge portfolio going long the 10% of stocks with highest predicted returns and shorting the 10% of stocks with lowest predicted returns. We interact each firm characteristic with the previous month's market capitalization. q indicates the size percentile of NYSE firms. The sample period is January 1965 to June 2014.

Firms		All	$Size > q_{10}$	$Size > q_{20}$	$Size > q_{20}$
Sample		Full	Full	Full	Full
Knots		20	15	15	10
Sample Size		1,629,155	959,757	763,850	763,850
# Selected		25	15	9	13
Sharpe Ratio		3.33	3.13	2.48	2.72
Characteristics	# Selected	(1)	(2)	(3)	(4)
BEME	1	BEME			
ΔShrout	4	ΔShrout	ΔShrout	ΔShrout	ΔShrout
ΔSO	4	ΔSO	ΔSO	ΔSO	ΔSO
DTO	1	DTO			
Investment	1	Investment			
Lturnover	2	Lturnover	Lturnover		
NOA	1		NOA		
PM_adj	1	PM_adj			
r_{2-1}	1	r_{2-1}			
r_{6-2}	2		r_{6-2}		r_{6-2}
r_{12-2}	1		r_{12-2}		
r_{12-7}	4	r_{12-7}	r_{12-7}	r_{12-7}	r_{12-7}
r_{36-13}	3	r_{36-13}	r_{36-13}		r_{36-13}
Rel_to_high_price	2	Rel_to_high_price	Rel_to_high_price		Rel_to_high_price
S2P	3		S2P	S2P	S2P
SUV	4	SUV	SUV	SUV	SUV
Total vol	4	Total vol	Total vol	Total vol	Total vol
Characteristics × Size					
A2ME	1	A2ME			
BEME_adj	1	BEME_adj			
DTO	1	DTO			
EPS	1	EPS			
NOA	1	NOA			
r_{2-1}	4	r_{2-1}	r_{2-1}	r_{2-1}	r_{2-1}
r_{6-2}	4	r_{6-2}	r_{6-2}	r_{6-2}	r_{6-2}
r_{12-2}	4	r_{12-2}	r_{12-2}	r_{12-2}	r_{12-2}
Rel_to_high_price	1	Rel_to_high_price			
Ret max	1	Ret max			
ROC	1				ROC
ROE	1	ROE			
SUV	1	SUV			

Table 7: **Selected Characteristics in Linear Model**

This table reports the selected characteristics from the universe of 62 firm characteristics we discuss in Section A.1 of the online appendix for a linear model and raw characteristics in column (1), a linear model and ranked-transformed characteristics in column (2), and the false discovery rate adjusted p value selection model of Green et al. (2017) in (3) and in-sample Sharpe ratios of an equally-weighted hedge portfolio going long the 10% of stocks with highest predicted returns and shorting the 10% of stocks with lowest predicted returns. The sample period is January 1965 to June 2014.

Firms		All	All	All
Model		Linear Model	Linear Model: rank normalized	Linear Model: FDR
Sample		Full	Full	Full
Sample Size		1,629,155	1,629,155	1,629,155
# Selected		24	35	32
Sharpe Ratio		1.47	2.52	1.64
Characteristics	# Selected	(1)	(2)	(3)
A2ME	1			A2ME
AOA	1		AOA	
AT	1		AT	
BEME	2	BEME	BEME	BEME
BEME_adj	1	BEME_adj		BEME_adj
Beta	1	Beta		Beta
C	2	C	C	C
CTO	1		CTO	
ΔCEQ	1		ΔCEQ	
ΔPI2A	1			ΔPI2A
ΔShrout	2	ΔShrout	ΔShrout	ΔShrout
ΔSO	1	ΔSO		ΔSO
Debt2P	1		Debt2P	
DTO	2	DTO	DTO	DTO
E2P	2	E2P	E2P	E2P
EPS	1		EPS	
Idio vol	2	Idio vol	Idio vol	Idio vol
Investment	2	Investment	Investment	Investment
IPM	1			IPM
IVC	1			IVC
LDP	2	LDP	LDP	LDP
LME	1		LME	
Lturnover	2	Lturnover	Lturnover	
NOA	1			NOA
OA	1		OA	
OL	1		OL	OL
PCM	1		PCM	
PM	1		PM	
PM_adj	1		PM_adj	PM_adj
Prof	1		Prof	
Q	1	Q		Q
r_{2-1}	2	r_{2-1}	r_{2-1}	r_{2-1}
r_{6-2}	1	r_{6-2}		r_{6-2}
r_{12-2}	1		r_{12-2}	
r_{12-7}	2	r_{12-7}	r_{12-7}	r_{12-7}
r_{36-13}	2	r_{36-13}	r_{36-13}	r_{36-13}
Rel_to_high_price	2	Rel_to_high_price	Rel_to_high_price	Rel_to_high_price
Ret max	1	Ret max		Ret max
ROA	1	ROA		
ROE	1		ROE	
ROIC	2	ROIC	ROIC	
S2C	1		S2C	
S2P	1		S2P	
SAT	1			SAT
SAT_adj	2	SAT_adj	SAT_adj	SAT_adj
Spread	2	Spread	Spread	Spread
Std turnover	1			Std turnover
Std volume	1		Std volume	
SUV	2	SUV	SUV	SUV
Tan	1			Tan
Total vol	1			Total vol

Table 8: **Out-of-Sample Return Prediction**

This table reports out-of-sample Sharpe ratios of hedge portfolios going long the 10% of stocks with highest predicted returns and shorting the 10% of stocks with lowest predicted returns for different sets of firms, out-of-sample periods, number of interpolation points, for the nonparametric and linear models. The Table also reports mean returns, standard deviations, higher-order moments, turnover, and predictive slopes and R^2s for the hedge portfolios in Panel A, and separately for the long legs in Panel B and the short legs in Panel C. q indicates the size percentile of NYSE firms. We perform model selection from January 1965 until the months before start of the out-of-sample prediction.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Firms	All	All	All	All	All	All	$Size > q_{10}$	$Size > q_{10}$	$Size > q_{20}$	$Size > q_{20}$
oos period	1991-2014	1991-2014	1991-2014	1991-2014	1973-2014	1973-2014	1991-2014	1991-2014	1991-2014	1991-2014
Knots	10		10		10		10		10	
Sample Size	1,025,497	1,025,497	1,025,497	1,025,497	1,541,922	1,541,922	959,757	959,757	763,850	763,850
Model	NP	Linear	NP	Linear	NP	Linear	NP	Linear	NP	Linear
# Selected	11	30	30	11	12	30	9	24	9	24
Model for Selection	NP	Linear	Linear	NP	NP	Linear	NP	Linear	NP	Linear
Sharpe Ratio	2.75	1.06	2.61	1.09	3.11	1.41	1.22	0.13	0.89	0.06
Panel A: Long-Short Portfolio										
Mean Return (monthly)	3.82	1.95	3.59	2.09	4.36	2.17	1.55	0.19	1.20	0.09
Standard Deviation (monthly)	4.81	6.37	4.75	6.63	4.85	5.31	4.40	4.92	4.64	5.22
Sharpe Ratio	2.75	1.06	2.61	1.09	3.11	1.41	1.22	0.13	0.89	0.06
Sharpe Ratio_*adj*	1.56	0.29	1.50	0.33	1.05	0.05	0.01	-0.70	-0.20	-0.63
Transaction Costs	1.71	1.54	1.58	1.36	2.87	2.09	1.54	1.18	1.47	1.04
Skewness	2.77	2.27	1.54	3.14	3.59	2.12	0.54	1.18	0.74	-0.51
Kurtosis	19.56	19.21	7.69	29.84	34.07	22.34	8.45	20.36	10.21	16.92
Turnover1	69.26	55.24	65.04	62.17	73.46	55.47	74.29	55.57	73.77	50.68
Turnover2	33.11	25.72	31.07	29.48	35.51	25.96	36.17	26.32	35.94	23.85
β	0.78	0.38	0.56	0.45	0.88	0.39	0.51	0.10	0.44	0.03
R^2	1.95%	1.37%	1.78%	1.19%	2.78%	1.60%	2.12%	1.64%	2.38%	2.27%
Panel B: Long Leg										
Mean Return (monthly)	8.61	9.02	8.27	9.17	8.55	8.37	5.91	7.02	6.11	6.78
Standard Deviation (monthly)	0.46	0.33	0.46	0.32	0.45	0.33	0.34	0.20	0.29	0.20
Sharpe Ratio	1.60	1.15	1.60	1.11	1.57	1.13	1.17	0.68	1.01	0.69
Skewness	2.77	2.27	1.54	3.14	3.59	2.12	0.54	1.18	0.74	-0.51
Kurtosis	12.23	13.16	6.36	11.04	14.80	12.45	5.43	7.69	6.18	7.30
β	1.58	0.46	1.53	0.23	1.72	0.40	0.54	0.05	0.63	0.22
R^2	2.39%	0.98%	2.19%	0.68%	1.63%	1.00%	0.79%	1.01%	1.17%	1.56%
Panel C: Short Leg										
Mean Return (monthly)	6.51	7.00	6.37	6.08	6.22	7.21	6.63	7.17	6.38	7.15
Standard Deviation (monthly)	0.02	0.15	0.04	0.14	-0.08	0.08	0.07	0.17	0.09	0.18
Sharpe Ratio	0.08	0.51	0.12	0.48	-0.27	0.27	0.23	0.58	0.31	0.62
Skewness	-0.08	0.03	-0.28	0.15	-0.25	-0.04	-0.13	-0.10	-0.15	0.26
Kurtosis	4.73	5.42	5.44	5.29	5.42	5.67	4.43	4.21	4.73	6.04
β	0.87	0.38	0.45	0.49	1.07	0.31	0.29	0.22	0.39	0.07
R^2	0.69%	1.35%	0.78%	0.86%	0.89%	1.17%	1.01%	1.41%	1.18%	2.06%

Table 9: **Out-of-Sample Predictions (Rolling Selection)**

This table reports out-of-sample Sharpe ratios of hedge portfolios going long the 10% of stocks with highest predicted returns and shorting the 10% of stocks with lowest predicted returns for different sets of firms, out-of-sample periods, number of interpolation points, for the nonparametric and linear models. The Table also reports mean returns, standard deviations, higher-order moments, turnover, and predictive slopes and R^2s for the hedge portfolios in Panel A, and separately for the long legs in Panel B and the short legs in Panel C. q indicates the size percentile of NYSE firms. We perform the first model selection from January 1965 until the months before start of the out-of-sample prediction and then perform model selection once a year keeping the selection window constant.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Firms	All	All	All	All	$Size > q_{10}$	$Size > q_{10}$	$Size > q_{20}$	$Size > q_{20}$
oos period	1991-2014	1991-2014	1991-2014	1991-2014	1991-2014	1991-2014	1991-2014	1991-2014
Knots	10		10		10		10	
Sample Size	1,025,497	1,025,497	1,025,497	1,025,497	959,757	959,757	763,850	763,850
Model	NP	Linear	NP	Linear	NP	Linear	NP	Linear
Average # Selected	14.13	26.58	30	11	10.75	27.21	10.75	28.92
Model for Selection	NP	Linear	Linear	NP	NP	Linear	NP	Linear
Panel A: Long-Short Portfolio								
Mean Return (monthly)	4.26	2.37	3.72	2.45	1.76	0.63	1.29	0.58
Standard Deviation (monthly)	5.65	5.75	5.35	6.06	3.83	4.37	4.30	4.37
Sharpe Ratio	2.61	1.43	2.41	1.40	1.60	0.50	1.04	0.46
Skewness	3.53	2.61	5.02	2.28	0.30	0.14	-0.50	-0.48
Kurtosis	28.74	21.57	52.79	19.61	7.80	10.39	13.06	10.99
Turnover1	69.70	56.05	71.27	62.24	72.22	49.58	74.48	47.62
Turnover2	33.37	26.13	34.32	29.48	35.09	23.19	36.29	22.25
β	0.87	0.44	0.72	0.52	0.55	0.18	0.45	0.14
R^2	2.16%	1.26%	1.86%	1.18%	2.00%	1.63%	2.37%	2.07%
Panel B: Long Leg								
Mean Return (monthly)	4.17	3.21	3.79	3.25	2.20	1.59	1.95	1.50
Standard Deviation (monthly)	8.98	8.45	8.56	9.04	6.08	6.50	6.03	6.47
Sharpe Ratio	1.61	1.32	1.53	1.25	1.25	0.85	1.12	0.80
Skewness	3.53	2.61	5.02	2.28	0.30	0.14	-0.50	-0.48
Kurtosis	13.81	13.45	14.40	10.47	4.40	7.12	4.78	5.85
β	1.75	0.53	1.69	0.49	0.54	0.00	0.68	0.05
R^2	2.50%	0.98%	2.32%	0.75%	0.76%	0.82%	1.01%	1.13%
Panel C: Short Leg								
Mean Return (monthly)	-0.09	0.84	0.07	0.80	0.44	0.96	0.66	0.92
Standard Deviation (monthly)	6.31	7.37	6.33	6.41	6.80	7.49	6.77	7.28
Sharpe Ratio	-0.05	0.39	0.04	0.43	0.22	0.44	0.34	0.44
Skewness	-0.18	0.10	-0.17	0.06	0.07	-0.21	0.32	-0.22
Kurtosis	5.47	5.19	5.51	5.78	5.46	4.44	7.10	4.40
β	0.97	0.37	0.75	0.47	0.56	0.35	0.29	0.27
R^2	0.77%	0.98%	0.70%	0.95%	1.33%	1.63%	1.66%	2.09%

Table 10: **Out-of-Sample Predictability in Simulation**

This table reports results from an out-of-sample prediction exercise for different model selection methods and data generating processes. Column (1) reports first the out-of-sample R^2 of regressing ex-post realized returns on ex-ante predicted returns for the true model and then the out-of-sample R^2 for the different model selection techniques relative to the true out-of-sample R^2. Column (2) reports the root mean squared prediction error (RMSPE) of the true model and the % differences between the RMSPEs of the true model and the different specifications. The sample period is January 1965 to June 2012 for model selection and 2013 to 2014 for out-of-sample prediction. We simulate each model 500 times. Panel A reports results for the nonparametric data generating process and Panel B reports results for the linear data generating process.

	(Relative) R^2 (1)	(Relative) RMSPE (2)
	Panel A: Nonlinear Data-Generating Process	
True parametric model	0.0160	0.1204
True nonparametric model	88.64%	0.091%
Adaptive group LASSO	88.61%	0.092%
Group LASSO	87.08%	0.106%
Adaptive LASSO linear	57.42%	0.328%
LASSO linear	57.61%	0.327%
FDR	57.65%	0.326%
t3	57.54%	0.327%
t2	58.03%	0.323%
	Panel B: Linear Data-Generating Process	
True parametric model	0.0088	0.1213
True nonparametric model	94.09%	0.028%
Adaptive group LASSO	93.64%	0.030%
Group LASSO	92.76%	0.035%
Adaptive LASSO linear	99.92%	0.000%
LASSO linear	99.74%	0.001%
FDR	97.23%	0.012%
t3	97.52%	0.011%
t2	99.09%	0.004%

Online Appendix: Dissecting Characteristics Nonparametrically

Not for Publication

A.1 Data

This section details the construction of variables we use in the main body of the paper with CRSP and Compustat variable names in parentheses and the relevant references. Unless otherwise specified, we use balance-sheet data from the fiscal year ending in year $t-1$ for returns from July of year t to June of year $t+1$ following the Fama and French (1993) timing convention.

A2ME: We follow Bhandari (1988) and define assets-to-market cap as total assets (AT) over market capitalization as of December t-1. Market capitalization is the product of shares outstanding (SHROUT) and price (PRC).

AOA: We follow Bandyopadhyay et al. (2010) and define AOA as absolute value of operation accruals (OA) which we define below.

AT Total assets (AT) as in Gandhi and Lustig (2015).

ATO: Net sales over lagged net operating assets as in Soliman (2008). Net operating assets are the difference between operating assets and operating liabilities. Operating assets are total assets (AT) minus cash and short-term investments (CHE), minus investment and other advances (IVAO). Operating liabilities are total assets (AT), minus debt in current liabilities (DLC), minus long-term debt (DLTT), minus minority interest (MIB), minus preferred stock (PSTK), minus common equity (CEQ).

BEME: Ratio of book value of equity to market value of equity. Book equity is shareholder equity (SH) plus deferred taxes and investment tax credit (TXDITC), minus preferred stock (PS). SH is shareholders' equity (SEQ). If missing, SH is the sum of common equity (CEQ) and preferred stock (PS). If missing, SH is the difference between total assets (AT) and total liabilities (LT). Depending on availability, we use the redemption (item PSTKRV), liquidating (item PSTKL), or par value (item PSTK) for PS. The market value of equity is as of December t-1. The market value of equity is the product of shares outstanding (SHROUT) and price (PRC). See Rosenberg, Reid, and Lanstein (1985) and Davis, Fama, and French (2000).

BEME_adj: Ratio of book value of equity to market value of equity minus the average industry ratio of book value of equity to market value of equity at the Fama-French 48 industry level as in Asness et al. (2000). Book equity is shareholder equity (SH) plus deferred taxes and investment tax credit (TXDITC), minus preferred stock (PS). SH is shareholders' equity (SEQ). If missing, SH is the sum of common equity

(CEQ) and preferred stock (PS). If missing, SH is the difference between total assets (AT) and total liabilities (LT). Depending on availability, we use the redemption (item PSTKRV), liquidating (item PSTKL), or par value (item PSTK) for PS. The market value of equity is as of December t-1. The market value of equity is the product of shares outstanding (SHROUT) and price (PRC).

Beta: We follow Frazzini and Pedersen (2014) and define the CAPM beta as product of correlations between the excess return of stock i and the market excess return and the ratio of volatilities. We calculate volatilities from the standard deviations of daily log excess returns over a one-year horizon requiring at least 120 observations. We estimate correlations using overlapping three-day log excess returns over a five-year period requiring at least 750 non-missing observations.

Beta_daily: Beta_daily is the sum of the regression coefficients of daily excess returns on the market excess return and one lag of the market excess return as in Lewellen and Nagel (2006).

C: Ratio of cash and short-term investments (CHE) to total assets (AT) as in Palazzo (2012).

C2D: Cash flow to price is the ratio of income and extraordinary items (IB) and depreciation and amortization (dp) to total liabilities (LT).

CTO: We follow Haugen and Baker (1996) and define capital turnover as ratio of net sales (SALE) to lagged total assets (AT).

Debt2P: Debt to price is the ratio of long-term debt (DLTT) and debt in current liabilities (DLC) to the market capitalization as of December t-1 as in Litzenberger and Ramaswamy (1979). Market capitalization is the product of shares outstanding (SHROUT) and price (PRC).

Δceq: We follow Richardson et al. (2005) in the definition of the percentage change in the book value of equity (CEQ).

Δ(ΔGm-ΔSales): We follow Abarbanell and Bushee (1997) in the definition of the difference in the percentage change in gross margin and the percentage change in sales (SALE). We define gross margin as the difference in sales (SALE) and costs of goods sold (COGS).

ΔSo: Log change in the split adjusted shares outstanding as in Fama and French (2008). Split adjusted shares outstanding are the product of Compustat shares outstanding (CSHO) and the adjustment factor (AJEX).

Δshrout: We follow Pontiff and Woodgate (2008) in the definition of the percentage change in shares outstanding (SHROUT).

ΔPI2A: We define the change in property, plants, and equipment following Lyandres, Sun, and Zhang (2008) as changes in property, plants, and equipment (PPEGT) and inventory (INVT) over lagged total assets (TA).

DTO: We follow Garfinkel (2009) and define turnover as ratio of daily volume (VOL) to shares outstanding (SHROUT) minus the daily market turnover and de-trend it by its 180 trading day median. We follow Anderson and Dyl (2005) and scale down the volume of NASDAQ securities by 38% after 1997 and by 50% before that to address the issue of double-counting of volume for NASDAQ securities.

E2P: We follow Basu (1983) and define earnings to price as the ratio of income before extraordinary items (IB) to the market capitalization as of December t-1. Market capitalization is the product of shares outstanding (SHROUT) and price (PRC).

EPS: We follow Basu (1977) and define earnings per share as the ratio of income before extraordinary items (IB) to shares outstanding (SHROUT) as of December t-1..

Free CF: Cash flow to book value of equity is the ratio of net income (NI), depreciation and amortization (DP), less change in working capital (WCAPCH), and capital expenditure (CAPX) over the book-value of equity defined as in the construction of BEME (see Hou et al. (2011)).

Idio vol: Idiosyncratic volatility is the standard deviation of the residuals from a regression of excess returns on the Fama and French (1993) three-factor model as in Ang, Hodrick, Xing, and Zhang (2006). We use one month of daily data and require at least fifteen non-missing observations.

Investment: We define investment as the percentage year-on-year growth rate in total assets (AT) following Cooper, Gulen, and Schill (2008).

IPM: We define pre-tax profit margin as ratio of pre-tax income (PI) to sales (SALE).

IVC: We define IVC as change in inventories (INVT) between $t-2$ and $t-1$ over the average total assets (AT) of years $t-2$ and $t-1$ following Thomas and Zhang (2002).

Lev: Leverage is the ratio of long-term debt (DLTT) and debt in current liabilities (DLC) to the sum of long-term debt, debt in current liabilities, and stockholders' equity (SEQ) following Lewellen (2015).

LDP: We follow Litzenberger and Ramaswamy (1979) and define the dividend-price ratio as annual dividends over last months price (PRC). We measure annual dividends as the sum of monthly dividends over the last 12 months. Monthly dividends are the scaled difference between returns including dividends (RET) and returns excluding dividends (RETX).

LME: Size is the total market capitalization of the previous month defined as price (PRC) times shares outstanding (SHROUT) as in Fama and French (1992).

LME_adj: Industry-adjusted-size is the total market capitalization of the previous month defined as price (PRC) times shares outstanding (SHROUT) minus the average industry market capitalization at the Fama-French 48 industry level as in Asness et al. (2000).

LTurnover: Turnover is last month's volume (VOL) over shares outstanding (SHROUT) (Datar, Naik, and Radcliffe (1998)).

NOA: Net operating assets are the difference between operating assets minus operating liabilities scaled by lagged total assets as in Hirshleifer, Hou, Teoh, and Zhang (2004). Operating assets are total assets (AT) minus cash and short-term investments (CHE), minus investment and other advances (IVAO). Operating liabilities are total assets (AT), minus debt in current liabilities (DLC), minus long-term debt (DLTT), minus minority interest (MIB), minus preferred stock (PSTK), minus common equity (CEQ).

NOP: Net payout ratio is common dividends (DVC) plus purchase of common and preferred stock (PRSTKC) minus the sale of common and preferred stock (SSTK) over the market capitalization as of December as in Boudoukh, Michaely, Richardson, and Roberts (2007).

O2P: payout ratio is common dividends (DVC) plus purchase of common and preferred stock (PRSTKC) minus the change in value of the net number of preferred stocks outstanding (PSTKRV) over the market capitalization as of December as in Boudoukh, Michaely, Richardson, and Roberts (2007).

OA: We follow Sloan (1996) and define operating accruals as changes in non-cash working capital minus depreciation (DP) scaled by lagged total assets (TA). Non-cash working capital is the difference between non-cash current assets and current liabilities (LCT), debt in current liabilities (DLC) and income taxes payable (TXP). Non-cash current assets are current assets (ACT) minus cash and short-term investments (CHE).

OL: Operating leverage is the sum of cost of goods sold (COGS) and selling, general, and administrative expenses (XSGA) over total assets as in Novy-Marx (2011).

PCM: The price-to-cost margin is the difference between net sales (SALE) and costs of goods sold (COGS) divided by net sales (SALE) as in Gorodnichenko and Weber (2016) and D'Acunto, Liu, Pflueger, and Weber (2017).

PM: The profit margin is operating income after depreciation (OIADP) over sales (SALE) as in Soliman (2008).

PM_adj: The adjusted profit margin is operating income after depreciation (OIADP) over net sales (SALE) minus the average profit margin at the Fama-French 48 industry level as in Soliman (2008).

Prof: We follow Ball, Gerakos, Linnainmaa, and Nikolaev (2015) and define profitability as gross profitability (GP) divided by the book value of equity as defined above.

Q: Tobin's Q is total assets (AT), the market value of equity (SHROUT times PRC) minus cash and short-term investments (CEQ), minus deferred taxes (TXDB) scaled by total assets (AT).

Rel to High: Closeness to 52-week high is the ratio of stock price (PRC) at the end of the previous calendar month and the previous 52 week high price defined as in George and Hwang (2004).

Ret_max: Maximum daily return in the previous month following Bali, Cakici, and Whitelaw (2011).

RNA: The return on net operating assets is the ratio of operating income after depreciation to lagged net operating assets (Soliman (2008)). Net operating assets are the difference between operating assets minus operating liabilities. Operating assets are total assets (AT) minus cash and short-term investments (CHE), minus investment and other advances (IVAO). Operating liabilities are total assets (AT), minus debt in current liabilities (DLC), minus long-term debt (DLTT), minus minority interest (MIB), minus preferred stock (PSTK), minus common equity (CEQ).

ROA: Return-on-assets is income before extraordinary items (IB) to lagged total assets (AT) following Balakrishnan, Bartov, and Faurel (2010).

ROC: ROC is the ratio of market value of equity (ME) plus long-term debt (DLTT) minus total assets to Cash and Short-Term Investments (CHE) as in Chandrashekar and Rao (2009).

ROE: Return-on-equity is income before extraordinary items (IB) to lagged book-value of equity as in Haugen and Baker (1996).

ROIC: Return on invested capital is the ratio of earnings before interest and taxes (EBIT) less nonoperating income (NOPI) to the sum of common equity (CEQ), total liabilities (LT), and Cash and Short-Term Investments (CHE) as in Brown and Rowe (2007).

$\mathbf{r_{12-2}}$: We define momentum as cumulative return from 12 months before the return prediction to two months before as in Fama and French (1996).

r_{12-7} : We define intermediate momentum as cumulative return from 12 months before the return prediction to seven months before as in Novy-Marx (2012).

r_{6-2} : We define r_{6-2} as cumulative return from 6 months before the return prediction to two months before as in Jegadeesh and Titman (1993).

r_{2-1} : We define short-term reversal as lagged one-month return as in Jegadeesh (1990).

r_{36-13} : Long-term reversal is the cumulative return from 36 months before the return prediction to 13 months before as in De Bondt and Thaler (1985).

S2C: Sales-to-cash is the ratio of net sales (SALE) to Cash and Short-Term Investments (CHE) following Ou and Penman (1989).

S2P: Sales-to-price is the ratio of net sales (SALE) to the market capitalization as of December following Lewellen (2015).

Sales_g: Sales growth is the percentage growth rate in annual sales (SALE) following Lakonishok, Shleifer, and Vishny (1994).

SAT: We follow Soliman (2008) and define asset turnover as the ratio of sales (SALE) to total assets (AT).

SAT_adj: We follow Soliman (2008) and define adjusted asset turnover as the ratio of sales (SALE) to total assets (AT) minus the average asset turnover at the Fama-French 48 industry level.

SGA2S: SG&A to sales is the ratio of selling, general and administrative expenses (XSGA) to net sales (SALE).

Spread: The bid-ask spread is the average daily bid-ask spread in the previous months as in Chung and Zhang (2014).

Std_turnover: Std_turnover is the standard deviation of the residuals from a regression of daily turnover on a constant as in Chordia, Subrahmanyam, and Anshuman (2001). Turnover is the ratio of volume (VOL) times shares outstanding (SHROUT) We use one month of daily data and require at least fifteen non-missing observations.

Std_volume: Std_volume is the standard deviation of the residuals from a regression of daily volume on a constant as in Chordia, Subrahmanyam, and Anshuman (2001). We use one month of daily data and require at least fifteen non-missing observations.

SUV: Standard unexplained volume is difference between actual volume and

predicted volume in the previous month. Predicted volume comes from a regression of daily volume on a constant and the absolute values of positive and negative returns. Unexplained volume is standardized by the standard deviation of the residuals from the regression as in Garfinkel (2009).

Tan: We follow Hahn and Lee (2009) and define tangibility as (0.715 $\times$ total receivables (RECT) + 0.547 $\times$ inventories (INVT) + 0.535 $\times$ property, plant and equipment (PPENT) + cash and short-term investments (CHE)) / total assets (AT).

Total_vol: Total volatility is the standard deviation of the residuals from a regression of excess returns on a constant as in Ang, Hodrick, Xing, and Zhang (2006). We use one month of daily data and require at least fifteen non-missing observations.

A.2 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 in the following period, R_{it}. Formally, we try to describe the conditional mean function,

$$E[R_{it} \mid C_{1,it-1}, \ldots, C_{S,it-1}]. \tag{A.1}$$

We often use portfolio sorts to approximate equation (1) for a single characteristic. 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.[1] Second, portfolio sorts offer little formal guidance to discriminate between characteristics. Consider the case of sorting stocks into five portfolios based on size, and within these, into five portfolios based on the book-to-market ratio. If we now find the book-to-market ratio only leads to a spread in returns for the smallest stocks, do we conclude it does not matter for expected returns? 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."[2] Nonetheless, portfolio sorts are by far the most commonly used technique to analyze which characteristics have predictive power for expected returns.

[1] 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 the factor construction for their five-factor model. They also argue not-well-diversified portfolios have little power in asset-pricing tests.

[2] Portfolio 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.

Instead of (conditional) double sorts, we could sort stocks into portfolios and perform spanning tests, that is, we regress long-short portfolios on a set of risk factors. Take 10 portfolios sorted on profitability and regress the hedge return on the three Fama and French (1993) factors. A significant time-series intercept would correspond to an increase in Sharpe ratios for a mean-variance investor relative to the investment set the three Fama and French (1993) factors span (see Gibbons, Ross, and Shanken (1989)). The order in which we test characteristics matters, and spanning tests cannot solve the selection problem of which characteristics provide incremental information for the cross section of expected returns.

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

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

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 exists why the conditional mean function should be linear.[3] 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 and extreme observations of the characteristics might drive point estimates. Researchers often use ad hoc techniques to mitigate these concerns, 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 that portfolio sorts and linear regressions face in the context of many return predictors, and suspects "we will have to

[3]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.

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.[4]

Suppose we observe excess returns R_{it} and a single characteristic C_{it-1} for stocks $i = 1, \ldots, N_t$ and time periods $t = 1, \ldots, T$. We sort stocks into L portfolios depending on the value of the lagged characteristic, C_{it-1}.[5] 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 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_{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

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

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,

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

[4]Cattaneo et al. (2016) develop inference methods for a portfolio-sorting estimator and also show the equivalence between portfolio sorting and nonparametric estimation.

[5]We only consider univariate portfolio sorts in this example to gain intuition.

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, $\hat{\beta} = (X'X)^{-1}X'\mathcal{R}$. It is easy to show that

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

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

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

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

$$\frac{1}{\sum_{t=1}^{T} N_{tl}}\sum_{t=1}^{T} N_{tl}P_{tl} - \frac{1}{\sum_{t=1}^{T} N_{tl'}}\sum_{t=1}^{T} N_{tl'}P_{tl'},$$

which equals $\hat{\beta}_l - \hat{\beta}_{l'}$.

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

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

A.3 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 incremental information for expected returns within that framework.

Suppose we knew the conditional mean function $m_t(c) \equiv E[R_{it} \mid C_{it-1} = c]$.[6] Then,

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

where $f_{C_{it-1}|C_{it-1}\in I_{tl}}$ 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, $\mathbf{1}(C_{it-1} \in I_{tl})$,

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

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

In nonparametric estimation, we call indicator functions of the form $\mathbf{1}(C_{it-1} \in I_{tl})$ 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. 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.[7]

Figures A.1–A.3 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 A.1 that most of the dispersion in book-to-market ratios and returns is in the extreme portfolios. Little variation in returns occurs 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 A.2 documents that a nonparametric estimator of the conditional mean function provides a good approximation for the relationship between book-to-market ratios and returns for intermediate values of the characteristic, but also in the extremes of the distribution.

Finally, we see in Figure A.3 that portfolio means provide a better fit in the tails of the distribution once we allow for more portfolios. Portfolio mean returns become more comparable to the predictions from the nonparametric estimator the larger the number of portfolios.

[7] We formally define our estimator in Section A.3. C below.

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: $(SV), (SG), (BV),$ and (BG). However, portfolio sorts quickly become infeasible as the number of characteristics increases. For example, if we have four characteristics and partition each characteristic into five portfolios, we end up with $5^4 = 625$ portfolios. Analyzing 625 portfolio returns would, of course, be impractical, but would also result in poorly diversified portfolios.

In nonparametric regressions, an analogous problem arises. Estimating the conditional mean function, m_t, fully nonparametrically with many regressors results in a slow rate of convergence and imprecise estimates in practice.[8] Specifically, with S characteristics and N_t observations, assuming technical regularity conditions, 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.[9] 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 comparable finite sample properties as in the case with only one characteristic. Consider a sample size, N_t, of 1,000. Then, we

[8] The literature refers to this phenomenon as the "curse of dimensionality" (see Stone (1982) for a formal treatment).

[9] We assume the conditional mean function, m_t, is twice continuously differentiable.

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 is to assume an additive model,

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

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

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

for all $s \neq s'$. For example, the predictive power of the book-to-market ratio for expected returns does not vary with firm size (conditional on size). 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. Brandt et al. (2009) make a similar assumption, but also stress that we can always interpret characteristics c as the cross product of a more basic set of characteristics. In our empirical application, we show results for all stocks and all-but micro caps, but also show results when we interact each characteristic with size.

Although the assumption of an additive model is somewhat restrictive, it provides desirable econometric advantages. In addition, we always make this assumption when we estimate multivariate regressions and in our context this assumption 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 C.

B 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

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

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,

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

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 in 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 it 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, $\text{rank}(\min_{i=1\ldots,N_t} C_{s,it-1}) = 1$ and $\text{rank}(\max_{i=1\ldots,N_t} C_{s,it-1}) = N_t$. 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.[10]

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

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

In 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.[11]

[10] The general econometric theory we discuss in Section C (model selection, consistency, etc.) also applies to any other monotonic transformation or the non-transformed conditional mean function.

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

C Adaptive Group LASSO

We use a group LASSO procedure developed 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,

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

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

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 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 not constant.[12]

In portfolio sorts, we approximate $\tilde{m}_{ts}$ by a constant within each portfolio. We instead propose to estimate quadratic functions over parts of the normalized characteristic distribution. Let $0 = t_0 < t_1 < \cdots < t_{L-1} < t_L = 1$ be a sequence of increasing numbers between 0 and 1 similar to portfolio breakpoints, and let $\tilde{I}_l$ for $l = 1, \ldots, L$ be a partition of the unit interval, that is, $\tilde{I}_l = [t_{l-1}, t_l)$ for $l = 1, \ldots, L-1$ and $\tilde{I}_L = [t_{L-1}, t_L]$. We refer to $t_0, \ldots, t_{L-1}$ as knots and choose $t_l = l/L$ for all $l = 0, \ldots, L-1$ in our empirical application. Because we apply the rank transformation to the characteristics, the knots correspond to quantiles of the characteristic distribution and we can think of $\tilde{I}_l$ as the l^{th} portfolio.

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 $[0, 1]$. We can approximate each $\tilde{m}_{ts}$ by a series expansion with

[12]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).

these properties, i.e.,

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

where $p_k(c)$ are known basis functions.[13]

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 A.1 – A.3).

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

$$\tilde{\boldsymbol{\beta}}_t = \underset{b_{sk}:s=1,\ldots,S;k=1,\ldots,L+2}{\arg\min} \sum_{i=1}^{N} \left(R_{it} - \sum_{s=1}^{S} \sum_{k=1}^{L+2} b_{sk} p_k(\tilde{C}_{s,it-1}) \right)^2 + \lambda_1 \sum_{s=1}^{S} \left(\sum_{k=1}^{L+2} b_{sk}^2 \right)^{\frac{1}{2}}, \tag{A.5}$$

where $\tilde{\boldsymbol{\beta}}_t$ is an $(L+2) \times S$ vector of 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 incremental information for expected returns. Due to the penalty, the LASSO is applicable even when the number of characteristics is larger than the sample size. Yuan and Lin (2006) propose to choose λ_1 in a data-dependent way to minimize Bayesian Information Criterion (BIC) which we follow in our application.

However, as in a linear model, the first step of the LASSO selects too many characteristics unless restrictive conditions on the design matrix hold. Informally

[13] In particular, $p_1(c) = 1$, $p_2(c) = c$, $p_3(c) = c^2$, and $p_k(c) = \max\{c - t_{k-3}, 0\}^2$ for $k = 4, \ldots, L+2$. See Chen (2007) for an overview of series estimation.

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:

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

Intuitively, these weights guarantee we do not select any characteristic in the second step that we did not select in the first step.

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

$$\breve{\boldsymbol{\beta}}_t = \underset{b_{sk}:s=1,\ldots,S;k=1,\ldots,L+2}{\arg\min} \sum_{i=1}^{N} \left(R_{it} - \sum_{s=1}^{S} \sum_{k=1}^{L+2} b_{sk} p_k(\tilde{C}_{s,it-1}) \right)^2 + \lambda_2 \sum_{s=1}^{S} \left(w_{ts} \sum_{k=1}^{L+2} b_{sk}^2 \right)^{\frac{1}{2}}. \tag{A.7}$$

We again follow Yuan and Lin (2006) and choose λ_2 to minimize BIC.

Huang et al. (2010) provide conditions under which $\breve{\boldsymbol{\beta}}_t$ is model-selection consistent; that is, it correctly selects the non-constant functions with probability approaching 1 as the sample size grows large.

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

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

If the cross section is sufficiently large, model selection and estimation could 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 over time. McLean and Pontiff (2016) show that for 97 return predictors, predictability decreases by 58% post publication. However, if the conditional mean function was 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

III, we estimate our model over subsamples and also estimate rolling specifications to investigate the variation in the conditional mean function over time.

D 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 S constants such that $\sum_{s=1}^{S} \alpha_s = 0$. Then,

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

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.

We report estimates of the functions using the common normalization that the functions integrate to 0, which is identified.

Section A.6 of the online appendix discusses how we construct confidence bands for the figures which we report and how we select the number of interpolation points in the empirical application of Section III.

A.4 Additive Conditional Mean Function

Estimating the conditional mean function, m_t, fully nonparametrically with many regressors results in a slow rate of convergence and imprecise estimates in practice.[14] Specifically, with S characteristics and N_t observations, assuming technical regularity conditions, 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

[14]The literature refers to this phenomenon as the "curse of dimensionality" (see Stone (1982) for a formal treatment).

of convergence decreases as S increases.[15] 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 comparable 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 is to assume an additive model,

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

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)).

[15] We assume the conditional mean function, m_t, is twice continuously differentiable.

An important restriction of the additive model is

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

for all $s \neq s'$; therefore, the additive model does not allow for cross dependencies between characteristics. For example, the predictive power of the book-to-market ratio for expected returns does not vary with firm size (conditional on size). 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. Brandt et al. (2009) make a similar assumption, but also stress that we can always interpret characteristics c as the cross product of a more basic set of characteristics. In our empirical application, we show results for all stocks and all-but micro caps, but also show results when we interact each characteristic with size.

Although the assumption of an additive model is somewhat restrictive, it provides desirable econometric advantages. In addition, we always make this assumption when we estimate multivariate regressions and in our context this assumption 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 C.

A.5 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

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

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,

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

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 in 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 it 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, $\text{rank}(\min_{i=1\ldots,N_t} C_{s,it-1}) = 1$ and $\text{rank}(\max_{i=1\ldots,N_t} C_{s,it-1}) = N_t$. 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.[16]

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

[16]The general econometric theory we discuss in subsection C below (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,

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

In 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.[17]

A.6 Confidence Bands

We also report uniform confidence bands for the estimated functions in the plots later to gain some intuition for estimation uncertainty. Note that the set of characteristics the LASSO selects does not rely on these confidence bands. As explained above, we assume that

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

In a linear model, we could report confidence intervals for the individual slope coefficients. Analogously, because we are mainly interested in the slopes of the functions $\tilde{m}_{ts}$ and because the levels of the functions are not separately identified, we report estimates and confidence bands for the functions $\tilde{m}_{ts}(\tilde{c}_s) - \int \tilde{m}_{ts}(\tilde{c}_s) d\tilde{c}_s$. That is, we normalize the functions such that they are 0 on average. By inspecting the confidence bands we can then test hypotheses that do not depend on the levels of the functions, such as whether a constant function or a linear function is consistent with the data. However, the bands are not informative about the levels of the estimated functions similar to confidence intervals

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

for slope coefficients in a linear model.

Recall that we approximate $\tilde{m}_{ts}(\tilde{c}_s)$ by $\sum_{k=1}^{L+2} \beta_{tsk} p_k(\tilde{c}_s)$ and estimate it by $\sum_{k=1}^{L+2} \hat{\beta}_{tsk} p_k(\tilde{c}_s)$. Let $\tilde{p}_k(\tilde{c}_s) = p_k(\tilde{c}_s) - \int p_k(\tilde{c}_s) d\tilde{c}_s$ be the normalized basis functions and let $\tilde{p}(\tilde{c}_s) = (\tilde{p}_1(\tilde{c}_s), \ldots, \tilde{p}_{L+2}(\tilde{c}_s))'$ be the corresponding vector of basis functions. Next 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 $\hat{\sigma}_{ts}(\tilde{c}_s) = \sqrt{\tilde{p}(\tilde{c}_s)' \hat{\Sigma}_{ts} \tilde{p}(\tilde{c}_s)}$, which is the estimated standard error of $\sum_{k=1}^{L+2} \hat{\beta}_{tsk} \tilde{p}_k(\tilde{c}_s)$. Just as in the linear model, $\hat{\sigma}_{ts}(\tilde{c})$ depends on which other characteristics are included in the model. For example, if two characteristics are highly correlated, the standard deviations of the estimated functions are typically high.

The uniform confidence band for $\tilde{m}_{ts}(\tilde{c}_s) - \int \tilde{m}_{ts}(\tilde{c}_s) d\tilde{c}_s$ is of the form

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

where d_{ts} is a constant. Thus, the width of the confidence band is proportional to the standard deviation of the estimated function. To choose the constant, let $Z \sim N(0, \hat{\Sigma}_{ts})$ and let $\hat{d}_{ts}$ be such that

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

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

$$P\left(\tilde{m}_{ts}(\tilde{c}_s) - \int \tilde{m}_{ts}(\tilde{c}_s) d\tilde{c}_s \in \left[\sum_{k=1}^{L+2} \hat{\beta}_{tsk} \tilde{p}_k(\tilde{c}_s) - \hat{d}_{ts} \hat{\sigma}_{ts}(\tilde{c}_s) \, , \sum_{k=1}^{L+2} \hat{\beta}_{tsk} \tilde{p}_k(\tilde{c}_s) + \hat{d}_{ts} \hat{\sigma}_{ts}(\tilde{c}_s) \right] \; \forall \tilde{c}_s \in [0,1] \right)$$

converges to $1 - \alpha$ as the sample size increases.

To better understand why these bands are useful, suppose that no linear function fits in the confidence band. Then, we can reject the null hypothesis that $\tilde{m}_{ts}(\tilde{c}_s) - \int \tilde{m}_{ts}(\tilde{c}_s) d\tilde{c}_s$

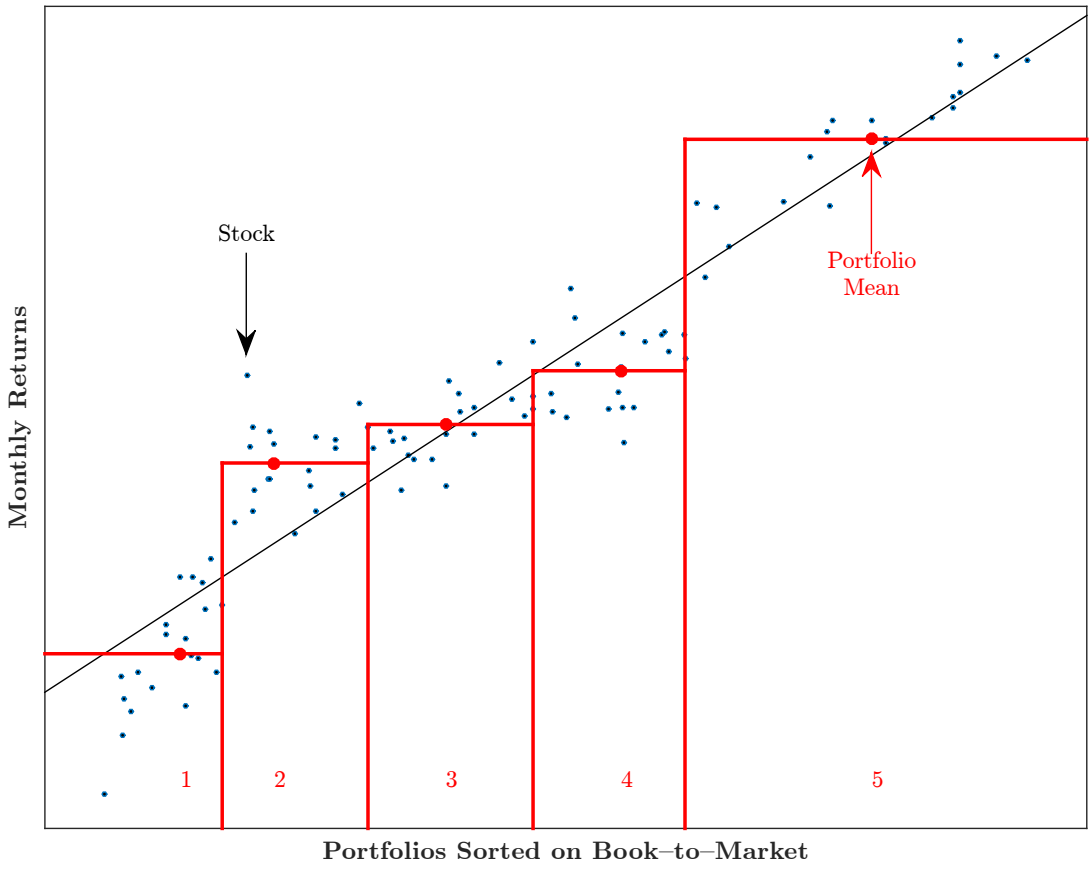
is linear at a significance level of $1-\alpha$. But since $\tilde{m}_{ts}(\tilde{c}_s) - \int \tilde{m}_{ts}(\tilde{c}_s)d\tilde{c}_s$ is linear if and only if $\tilde{m}_{ts}(\tilde{c}_s)$ is linear, we can then also reject the null hypothesis that $\tilde{m}_{ts}(\tilde{c}_s)$ is linear. Similar, by inspecting the band we can test if $\tilde{m}_{ts}(\tilde{c}_s)$ is constant.

We want to stress that the selection of characteristics in the LASSO does not rely on these confidence bands and we report the confidence bands only to provide intuition and to summarize sampling uncertainty.

A Knot Selection

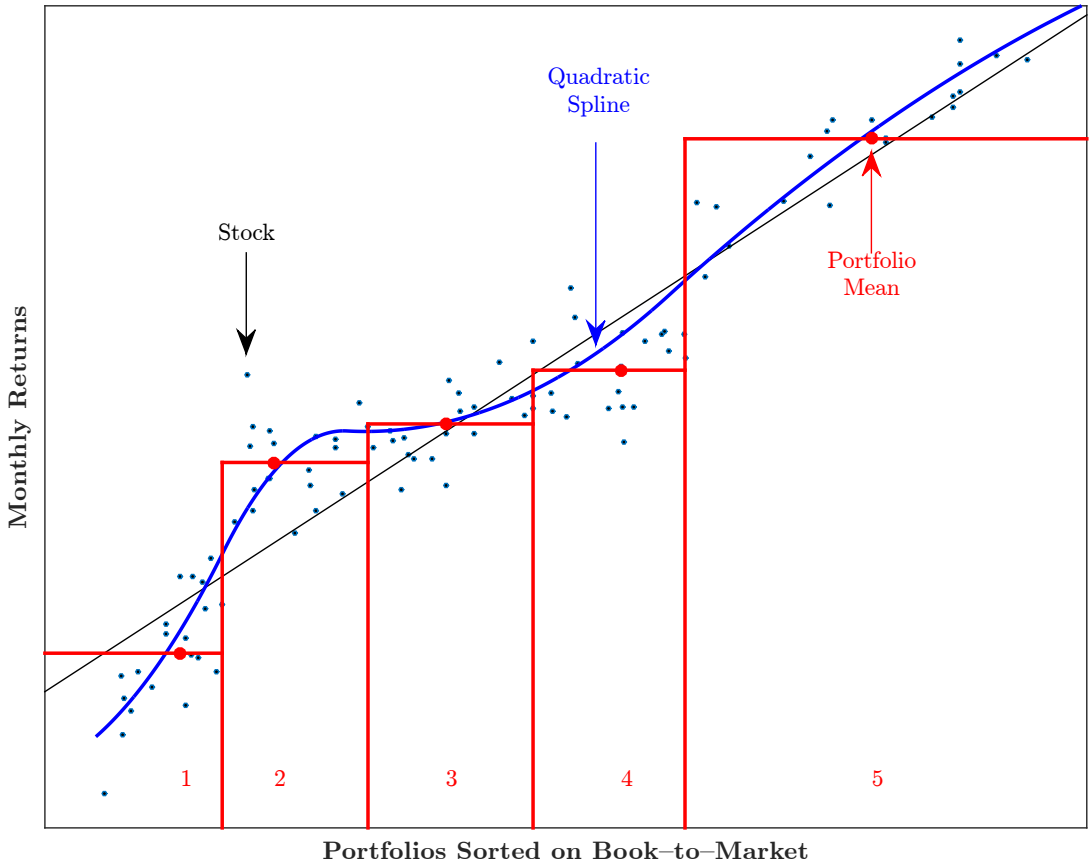
Theory tells us the number of interpolation points should grow as the sample size grows. Empirically, this statement is not too helpful in guiding our choices. We therefore document that the number and identify of characteristics is stable for reasonable variations in the number of knots (see Figure 5 which we discuss below).

Figure A.1: **5 Portfolios Sorted on Book-to-Market**



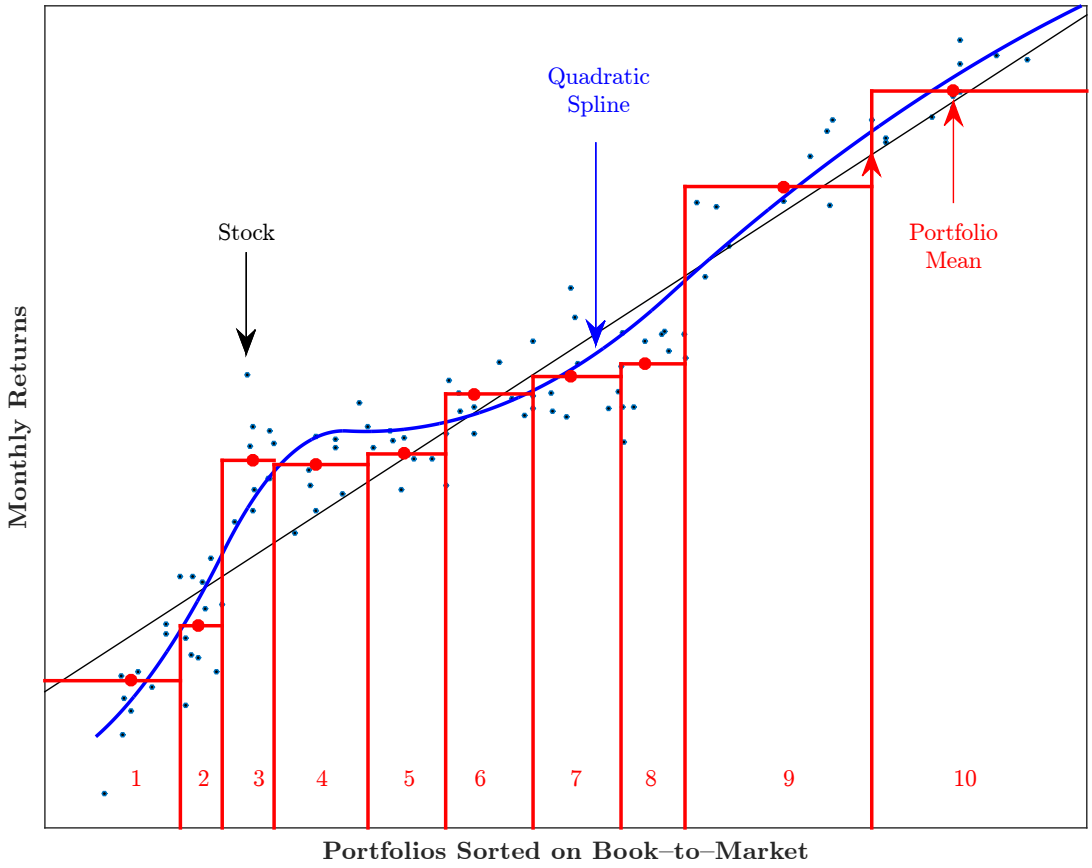
This figure plots returns on the y-axis against the book-to-market ratio on the x-axis as well as portfolio mean returns for simulated data.

Figure A.2: **5 Portfolios Sorted on Book-to-Market and Nonparametric Estimator**



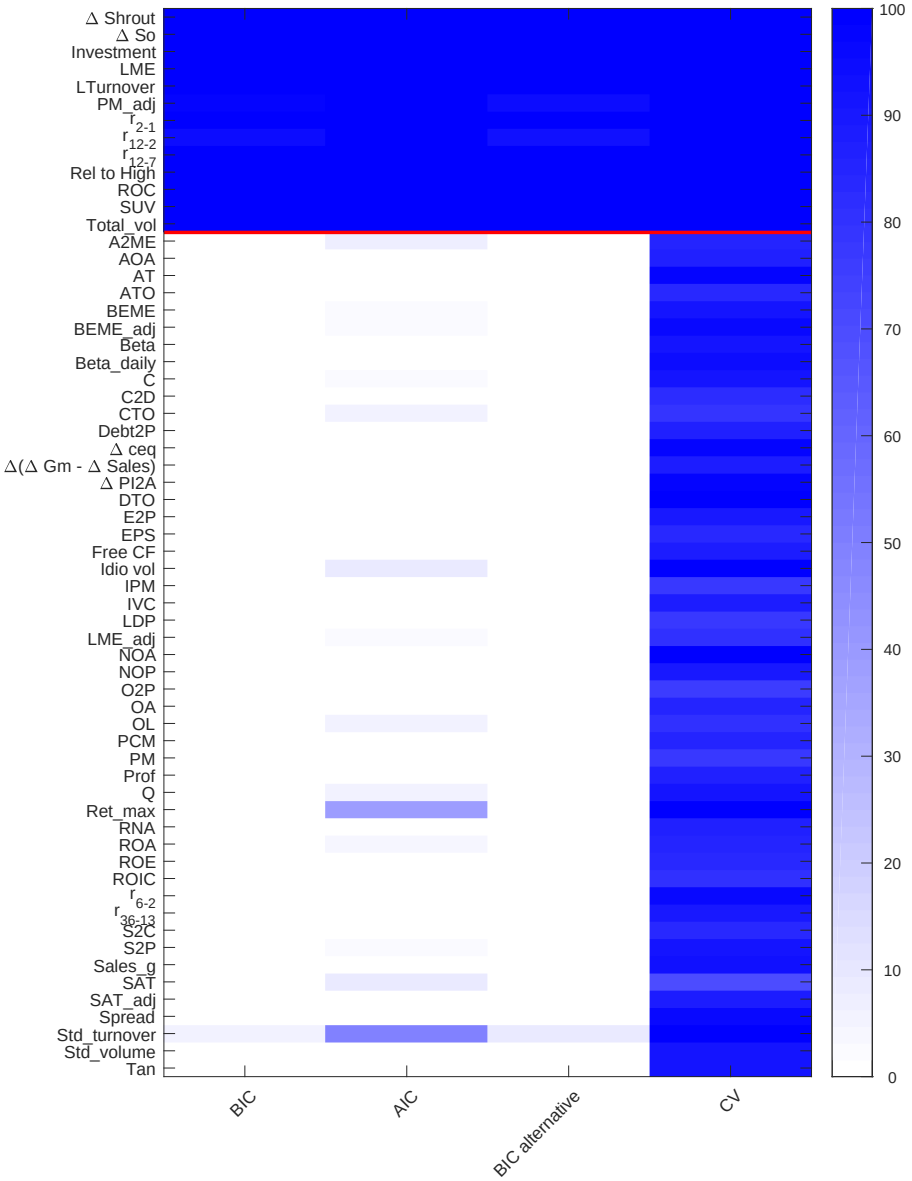
This figure plots returns on the y-axis against the book-to-market ratio on the x-axis as well as portfolio mean returns and a nonparametric conditional mean function for simulated data.

Figure A.3: **10 Portfolios sorted on Book-to-Market and Nonparametric Estimator**



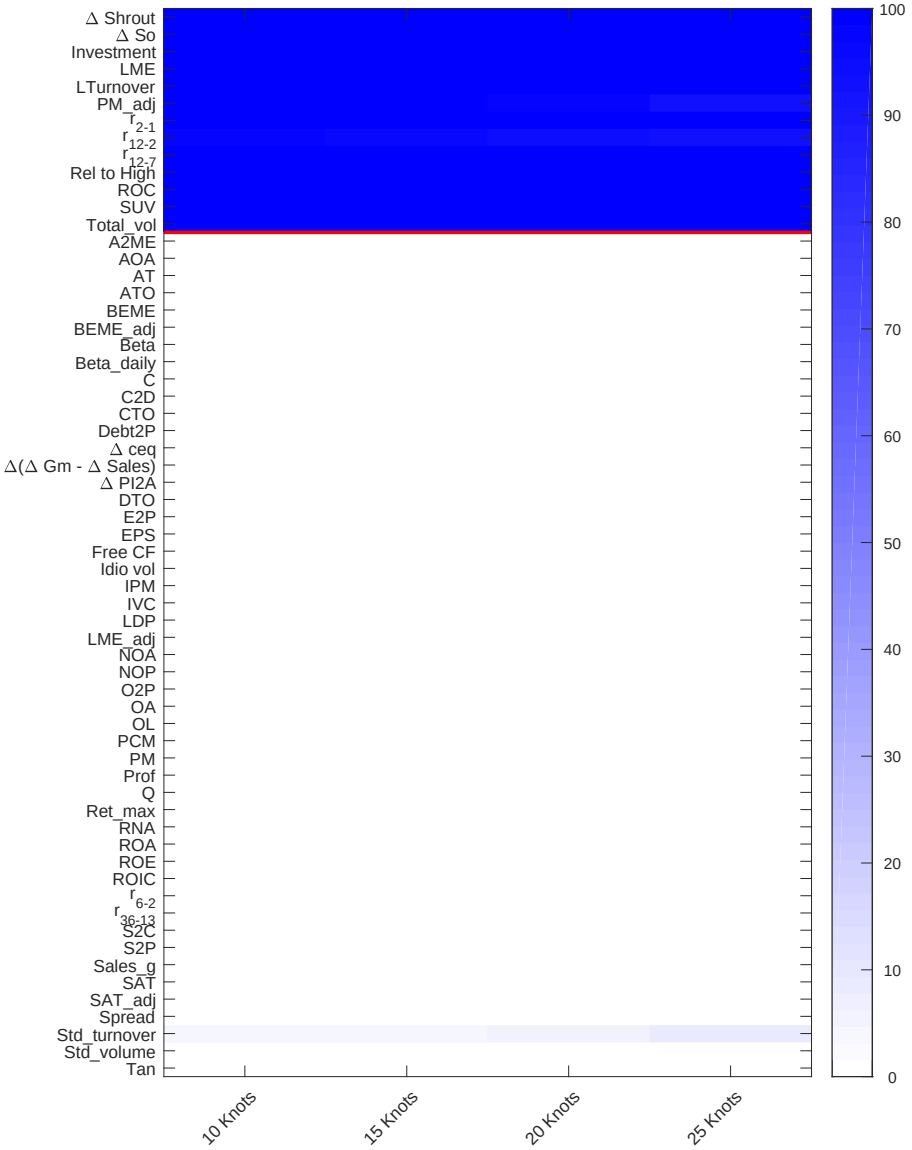
This figure plots returns on the y-axis against the book-to-market ratio on the x-axis as well as portfolio mean returns and a nonparametric conditional mean function for simulated data.

Figure A.4: **Selected Characteristics in Simulations: Empirical Data-Generating Process (different Information Criteria)**



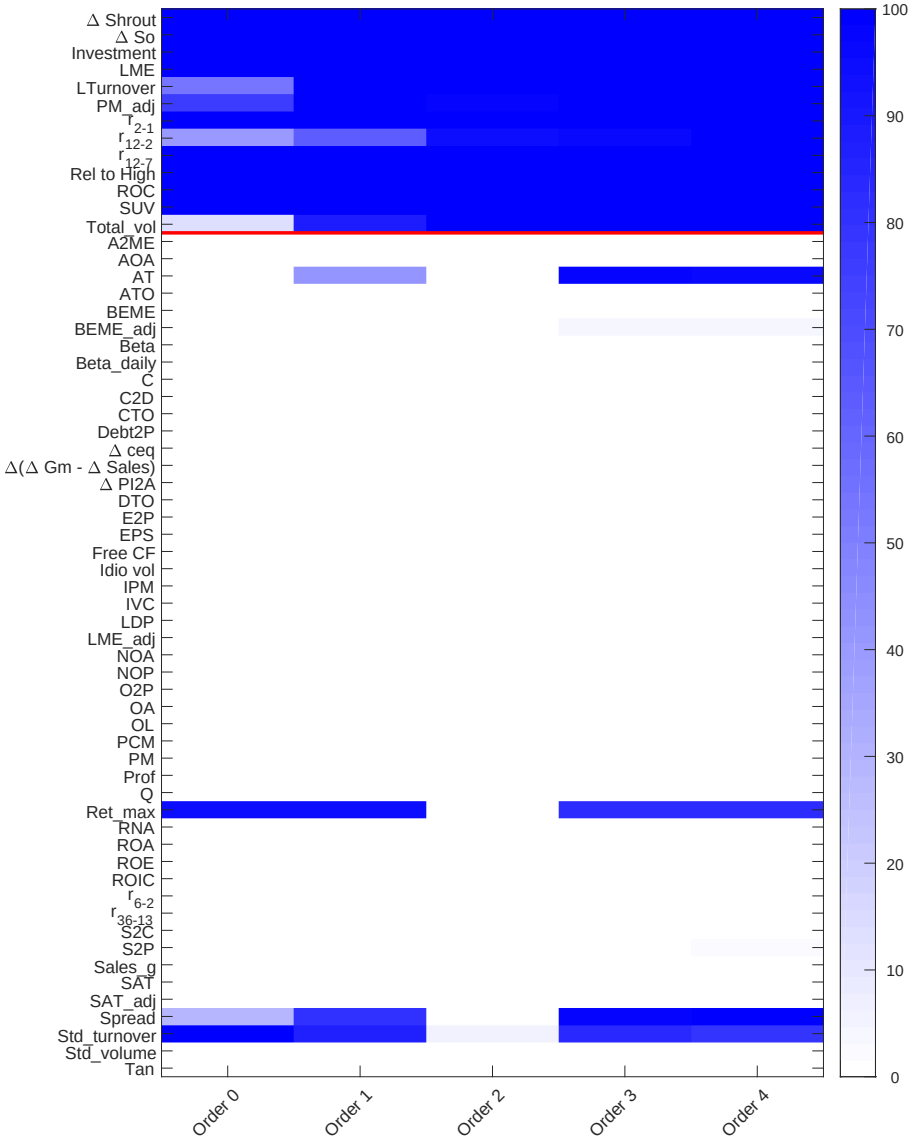
The figure graphically shows for the nonlinear adaptive group LASSO for different information criteria the frequency with which characteristics from the universe of 62 firm characteristics we discuss in Section A.1 of the online appendix are selected by each information criteria. The darker the color, the more frequently a given selection method selects a given characteristic. The true model is nonlinear and consists of the 13 characteristics above the red vertical line. The average number of selected characteristics for the different information criteria across 500 simulations are: BIC: 12.99; AIC: 14.59; BIC alternative: 12.97; CV (cross validation): 56.34. The sample period is January 1965 to June 2014.

Figure A.5: **Selected Characteristics in Simulations: Empirical Data-Generating Process (different Knot Numbers)**



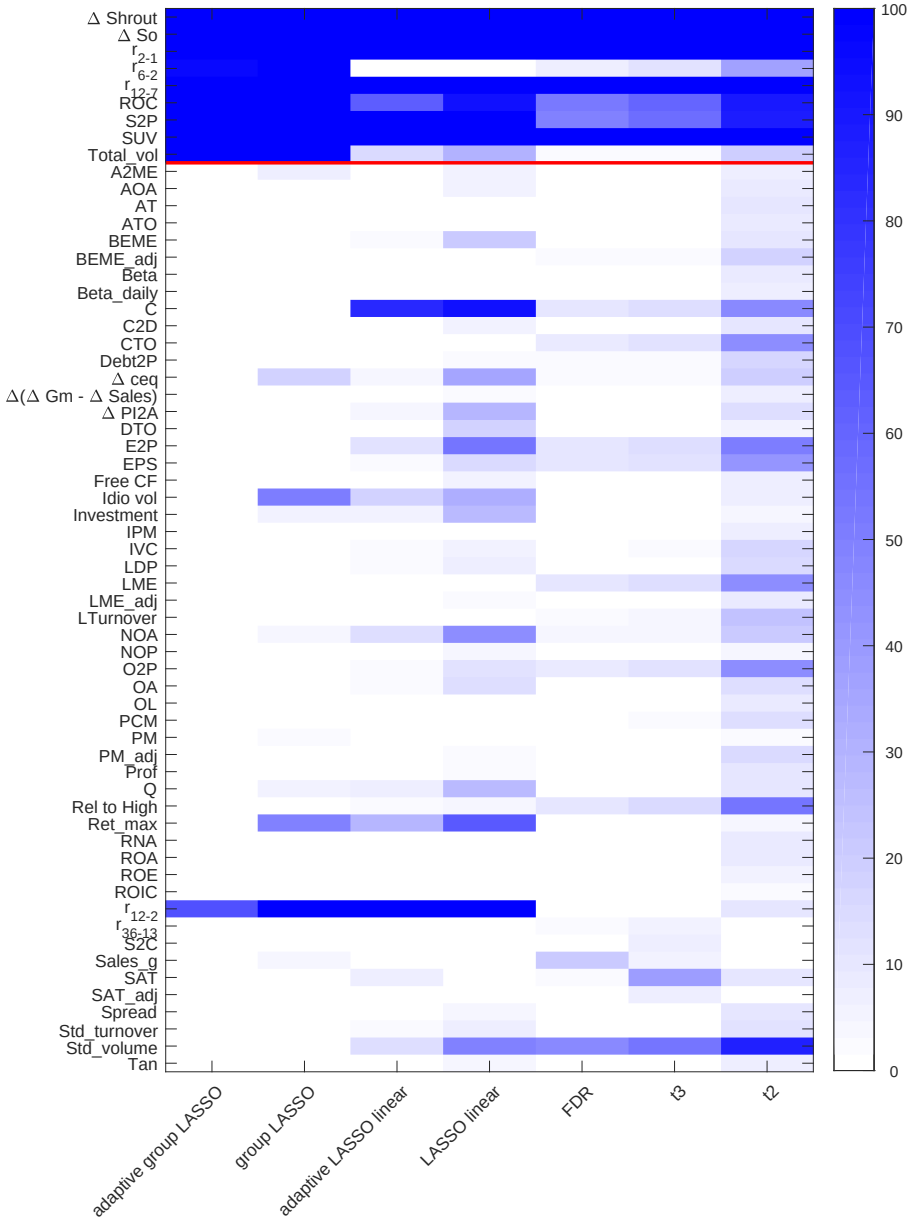
The figure graphically shows for the nonlinear adaptive group LASSO for different number of knots the frequency with which characteristics from the universe of 62 firm characteristics we discuss in Section A.1 of the online appendix are selected by each information criteria. The darker the color, the more frequently a given selection method selects a given characteristic. The true model is nonlinear and consists of the 13 characteristics above the red vertical line. The average number of selected characteristics for the different information criteria across 500 simulations are: 10 knots: 13.03; 15 knots: 13.01; 20 knots: 12.99; 25 knots: 12.94. The sample period is January 1965 to June 2014.

Figure A.6: **Selected Characteristics in Simulations: Empirical Data-Generating Process (different Order Splines)**



The figure graphically shows for the nonlinear adaptive group LASSO for different order splines the frequency with which characteristics from the universe of 62 firm characteristics we discuss in Section A.1 of the online appendix are selected by each information criteria. The darker the color, the more frequently a given selection method selects a given characteristic. The true model is nonlinear and consists of the 13 characteristics above the red vertical line. The average number of selected characteristics for the different information criteria across 500 simulations are: 0 order: 13.08; 1 order: 15.55; 2 order: 12.99; 3 order: 16.63; 4 order: 16.61. The sample period is January 1965 to June 2014.

Figure A.7: **Selected Characteristics in Simulations: Empirical Data-Generating Process (large Firms)**



The figure graphically shows for different model selection methods the frequency with which characteristics from the universe of 62 firm characteristics we discuss in Section A.1 of the online appendix are selected by each method for firms above the 20^{th} size percentile. The darker the color, the more frequently a given selection method selects a given characteristic. The true model is nonlinear and consists of the 9 characteristics above the red vertical line. The average number of selected characteristics for the different methods across 500 simulations are: adaptive group LASSO: 9.12; group LASSO: 12.50; adaptive LASSO linear model: 9.95; LASSO linear model: 14.60; FDR: 7.60; t3: 8.28; t2: 16.15. The sample period is January 1965 to June 2014.

Table A.1: **Out-of-Sample Predictability in Simulation: Robustness Nonlinear Model**

This table reports results from an out-of-sample prediction exercise for different model selection methods and data generating processes. Column (1) reports first the out-of-sample R^2 of regressing ex-post realized returns on ex-ante predicted returns for the true model and then the out-of-sample R^2 for the different model selection techniques relative to the true out-of-sample R^2. Column (2) reports the root mean squared prediction error (RMSPE) of the true model and the % differences between the RMSPEs of the true model and the different specifications. The sample period is January 1965 to June 2012 for model selection and 2013 to 2014 for out-of-sample prediction. We simulate each model 500 times. Panel A reports results for different information criteria, Panel B for different number of knots, and Panel C for different order splines. We use the nonparametric adaptive group LASSO for model selection with the BIC of Yuan and Lin (2006), 20 knots, and order 2 splines as baseline model.

	Relative R^2 (1)	Relative RMSPE (2)
	Panel A: Different Information Criteria	
BIC	88.61%	0.092%
AIC	87.91%	0.098%
BIC alternative	88.52%	0.092%
CV	60.66%	0.593%
	Panel B: Different Knots Numbers	
10 knots	86.96%	0.102%
15 knots	88.51%	0.090%
20 knots	88.61%	0.092%
25 knots	81.24%	0.166%
	Panel C: Different Order Splines	
Order 0	69.36%	0.241%
Order 1	84.01%	0.127%
Order 2	88.61%	0.092%
Order 3	90.53%	0.078%
Order 4	91.58%	0.070%

Table A.2: **Out-of-Sample Predictability in Simulation: Large Firms**

This table reports results from an out-of-sample prediction exercise for different model selection methods and data generating processes. Column (1) reports first the out-of-sample R^2 of regressing ex-post realized returns on ex-ante predicted returns for the true model and then the out-of-sample R^2 for the different model selection techniques relative to the true out-of-sample R^2. Column (2) reports the root mean squared prediction error (RMSPE) of the true model and the % differences between the RMSPEs of the true model and the different specifications. The sample period is January 1965 to June 2012 for model selection and 2013 to 2014 for out-of-sample prediction. We simulate each model 500 times. Large firms are all firms above the 20^{th} size percentile.

	(Relative) R^2 (1)	(Relative) RMSPE (2)
True parametric model	0.0062	0.0822%
True nonparametric model	90.13%	0.036%
Adaptive group LASSO	89.31%	0.039%
Group LASSO	87.06%	0.049%
Adaptive LASSO linear	79.40%	0.064%
LASSO linear	79.31%	0.064%
FDR	75.57%	0.076%
t3	76.21%	0.074%
t2	78.02%	0.068%