NBER WORKING PAPER SERIES

# DISSECTING CHARACTERISTICS NONPARAMETRICALLY

Joachim Freyberger
Andreas Neuhierl
Michael Weber

Working Paper 23227
http://www.nber.org/papers/w23227

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2017

We thank Jonathan Berk, Philip Bond, Oleg Bondarenko, John Campbell, Jason Chen, Josh Coval, Gene Fama, Ken French, Erwin Hansen, Lars Hansen, Bryan Kelly, Leonid Kogan, Shimon Kogan, Jon Lewellen, Bill McDonald, Stefan Nagel, Stavros Panageas, Lubos Pastor, Seth Pruitt, Alberto Rossi, George Skoulakis, Raman Uppal, Adrien Verdelhan, Amir Yaron and conference and seminar participants at Dartmouth College, FRA Conference 2016, HEC Montreal, McGill, 2017 Revelstoke Finance Conference, Santiago Finance Workshop, Stockholm School of Economics, TAU Finance Conference 2016, Tsinghua University PBCSF, Tsinghua University SEM, the University of Chicago, the University of Illinois at Chicago, the University of Notre Dame, and the University of Washington for valuable comments. Weber gratefully acknowledges financial support from the University of Chicago and the Fama-Miller Center. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Dissecting Characteristics Nonparametrically
Joachim Freyberger, Andreas Neuhierl, and Michael Weber
NBER Working Paper No. 23227
March 2017
JEL No. C14,C52,C58,G12

**ABSTRACT**

We propose a nonparametric method to test which characteristics provide independent 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 is insensitive to outliers. Many of the previously identified return predictors do not provide incremental information for expected returns, and nonlinearities are important. Our proposed method has higher out-of-sample explanatory power compared to linear panel regressions, and increases Sharpe ratios by 50%.

Joachim Freyberger
Department of Economics
University of Wisconsin-Madison
1180 Observatory Drive
Madison, WI 53706
jfreyberger@ssc.wisc.edu

Andreas Neuhierl
University of Notre Dame
College of Business
221 Mendoza
Notre Dame, IN 46556
aneuhier@nd.edu

Michael Weber
Booth School of Business
University of Chicago
5807 South Woodlawn Avenue
Chicago, IL 60637
and NBER
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 more than 300 published factors that have predictive power for the cross section of expected returns.[1] Many economic models, such as the consumption CAPM of Lucas (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.[2] Cochrane (2011) speculates, "To address these questions in the zoo of new variables, I suspect we will have to use different methods."

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

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

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

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

[3] In our empirical application, we estimate quadratic splines.

characteristics provide independent information for the cross section of expected returns. We estimate our model on 36 characteristics including size, book-to-market, beta, and other prominent variables and anomalies on a sample period from July 1963 to June 2015. Only 15 variables, including size, idiosyncratic volatility, and past return-based predictors, have independent explanatory power for expected returns for the full sample period and all stocks. An equally-weighted hedge portfolio going long the stocks with the 10% highest expected returns and shorting the 10% of stocks with the lowest predicted returns has an in-sample Sharpe ratio of close to 3. Only eight characteristics have predictive power for returns in the first half of our sample. In the second half, instead, we find 17 characteristics are significantly associated with cross-sectional return premia. For stocks whose market capitalization is above the 20% NYSE size percentile, only seven characteristics, including size, past returns, and standardized unexplained volume, remain significant return predictors. The in-sample Sharpe ratio is still 1.81 for large stocks.

Second, we compare the out-of-sample performance of the nonparametric model with a linear model. We estimate both models over a period until 1990 and select significant return predictors. We then use 10 years of data to estimate the model 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. 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.[4] We then move the estimation and prediction period forward by one month. The nonparametric model generates an average Sharpe ratio for an equally-weighted hedge portfolio of 3.42 compared to 2.26 for the linear model.[5] The linear model selects 21 characteristics in sample compared to only eight for the nonparametric model, but performs worse out of sample. Nonlinearities are important. We find an increase in

[4] 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.

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

out-of-sample Sharpe ratios relative to the Sharpe ratio of the linear model when we employ the nonparametric model for prediction on the 21 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 eight characteristics we select with the nonparametric model as we do with the 21 characteristics the linear model selects.

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 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 (2015).

## *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. Fama and MacBeth (1973) provide empirical support for the CAPM. Subsequently, researchers identified many variables, such as size (Banz (1981)), the book-to-market ratio (Rosenberg et al. (1985)), leverage (Bhandari (1988)), earnings-to-price ratios (Basu (1983)), or past returns (Jegadeesh and Titman (1993)) that contain additional independent information for expected returns. Sorting stocks into portfolios based on these characteristics often led to rejection of the CAPM because the spread in CAPM betas could not explain the spread in returns. Fama and French (1992) synthesize these findings, and Fama and French (1993) show that a three-factor model with the market return, a size, and a value factor can explain cross sections of stocks sorted on characteristics that appeared anomalous relative to the CAPM. In this sense, Fama and French (1993) and Fama and French (1996) achieve a significant dimension reduction: researchers who want to explain the cross section of stock returns only have to explain the size and value factors.

Daniel and Titman (1997), on the contrary, argue that characteristics have higher explanatory power for the cross section of expected returns than loadings on pervasive risk factors. Chordia, Goyal, and Shanken (2015) develop a method to estimate bias-corrected return premia from cross-sectional data for individual stocks. They find firm characteristics explain more of the cross-sectional variation in expected returns compared to factor loadings. Kozak, Nagel, and Santosh (2015) show comovements of stocks and associations of returns with characteristics orthogonal to factor exposures do not allow researchers to disentangle rational from behavioral explanations for return spreads. We study which characteristics provide incremental information for expected returns but do not aim to investigate whether rational models or behavioral explanations drive our findings.

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. Figure 3 shows the suggested adjustment over time. However, even employing the higher threshold for the $t$-statistic still leaves approximately 150 characteristics as useful predictors for the cross section of expected returns. Fama and French (2015) take a different route and augment the three-factor model of Fama and French (1993) with an investment and profitability factor (Haugen and Baker (1996) and Novy-Marx (2013)). Fama and French (2016) test the five-factor model on a small set of anomalies and find substantial improvements relative to a three-factor model, but also substantial unexplained return variation across portfolios. Hou et al. (2015) test a q-factor model consisting of four factors on 35 anomalies that are univariately associated with cross-sectional return premia, and find their model can reduce monthly alphas to an average of 0.20%. Barillas and Shanken (2016) develop a new method to directly compare competing factor models.

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 a few are significant predictors for the cross section of expected returns. Green, Hand, and Zhang (2016) extend Lewellen (2015) to many more characteristics for a shorter sample starting in 1980, but confirm his basic conclusion. Although Fama-MacBeth regressions carry a lot of intuition, they do not offer a formal method to select significant return predictors. We build on Lewellen (2015) and provide a framework that allows for nonlinear association between characteristics and returns, provide a formal framework to disentangle significant from insignificant return predictors, and study many more characteristics.

Light, Maslov, and Rytchkov (2016) use partial least squares (PLS) to summarize the predictive power of firm characteristics for expected returns. PLS summarizes the predictive power of all characteristics and therefore does not directly disentangle important from unimportant characteristics and does not reduce the number of characteristics for return prediction. Brandt, Santa-Clara, and Valkanov (2009) parameterize portfolio weights as a function of stock characteristics to sidestep the task to model the joint distribution of expected returns and characteristics. DeMiguel, Martin-Utrera, Nogales, and Uppal (2016) extend the parametric portfolio approach of Brandt et al. (2009) to study which characteristics provide incremental information for the cross section of returns. Specifically, DeMiguel et al. (2016) add long-short characteristic-sorted portfolios to benchmark portfolios, such as the value-weighted market portfolio, and ask which portfolios increase investor utility.

We also contribute to a small literature estimating non-linear asset-pricing models using semi- and nonparametric methods. Bansal and Viswanathan (1993) extend the arbitrage pricing theory (APT) of Ross (1976) and estimate the stochastic discount factor semiparametrically using neural nets. They allow for payoffs to be nonlinear in risk factors and find their APT is better able to explain the returns of size-sorted portfolios. Chapman (1997) explains the size effect with a consumption-based model in which he approximates the stochastic discount factor with orthonormal polynomials in a low number of state variables. Connor, Hagmann, and Linton (2012) propose a nonparametric regression method relating firm characteristics to factor loadings in a nonlinear way. They find momentum and stock volatility have explanatory power similar to size and value for the cross section of expected returns as size and value.

We 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 (2015) assume the irrepresentable condition of Meinshausen and Bühlmann (2006) to achieve model-selection consistency in a single-step LASSO. They use a linear model for high-frequency return predictability using past returns of related stocks, and find their method increases predictability relative to OLS.

Bryzgalova (2016) highlights that weak identification in linear factor models could result in an overstatement of significant cross-sectional risk factors.[6] She proposes a shrinkage-based estimator to detect possible rank deficiency in the design matrix and to identify strong asset-pricing factors. 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. We, instead, are mainly concerned with formal model selection, that is, which characteristics provide incremental information in the presence of other characteristics.

[6] See also Jagannathan and Wang (1998), Kan and Zhang (1999), Kleibergen (2009), Gospodinov, Kan, and Robotti (2014), Kleibergen and Zhan (2015), and Burnside (2016).

# II Current Methodology

## A Expected Returns and the Curse of Dimensionality

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

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

We often use portfolio sorts to approximate equation (1). We typically sort stocks into 10 portfolios and compare mean returns across portfolios. Portfolio sorts are simple, straightforward, and intuitive, but they also suffer from several shortcomings. First, we can only use portfolio sorts to analyze a small set of characteristics. Imagine sorting stocks jointly into five portfolios based on CAPM beta, size, book-to-market, profitability, and investment. We would end up with $5^5 = 3125$ portfolios, which is larger than the number of stocks at the beginning of our sample.[7] 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."[8] Nonetheless, portfolio sorts are by far the most commonly used technique to analyze which characteristics have predictive power for expected returns.

Instead of (conditional) double sorts, we could sort stocks into portfolios and perform

[7] 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.

[8] 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.

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,

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

Linear regressions allow us to study the predictive power for expected returns of many characteristics jointly, but they also have potential pitfalls. First, no a priori reason explains why the conditional mean function should be linear.[9] Fama and French (2008) estimate linear regressions as in equation (2) to dissect anomalies, but raise concerns over potential nonlinearities. They make ad hoc adjustments and use, for example, the log book-to-market ratio as a predictive variable. Second, linear regressions are sensitive to outliers. Third, small, illiquid stocks might have a large influence on point estimates because they represent the majority of stocks. Researchers often use ad hoc techniques to mitigate concerns related to microcaps and outliers, such as winsorizing observations and estimating linear regressions separately for small and large stocks (see Lewellen (2015) for a recent example).

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

[9] 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.

## 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.[10]

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, $C_{it-1}$. 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$^{\text{th}}$ percentile of the firm size distribution. We would sort that stock in the 5$^{\text{th}}$ 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

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

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

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

which is the intercept in a (time-series) regression of the difference in portfolio returns, $P_{tl} - P_{tl'}$, on a constant.[11]

Alternatively, we can run a pooled time-series cross-sectional regression of excess

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

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

returns on dummy variables, which equal 1 if firm $i$ is in portfolio $l$ in period $t$. We denote the dummy variables by $\mathbf{1}(C_{it-1} \in I_{tl})$ and write,

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

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

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

It then follows that

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

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

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

and

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

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

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

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

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

Second, we could use the weighted least squares estimator,

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

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

# III Nonparametric Estimation

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

Suppose we knew the conditional mean function $m_t(c) \equiv E[R_{it} \mid C_{it-1} = c]$.[12] Then,

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

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

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

functional form more generally.

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

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

In nonparametric estimation, we call indicator functions of the form $\mathbf{1}(C_{it-1} \in I_{tl})$ 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.[13]

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

We see in Figure 4 that most of the dispersion in book-to-market ratios and returns is in the extreme portfolios. Little variation 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 5 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 6 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

[13] We formally define our estimator in Section III. D below.

of portfolios.

## A Multiple Regression & Additive Conditional Mean Function

Both portfolio sorts and regressions theoretically allow us to look at several characteristics simultaneously. Consider small ($S$) and big ($B$) firms and value ($V$) and growth ($G$) firms. We could now study four portfolios: $(SV), (SG), (BV)$, 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 $m_t(c) \equiv E[R_{it} \mid C_{it} = c]$ fully nonparametrically with many regressors results in a slow rate of convergence and imprecise estimates in practice.[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 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

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

[15] 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 $s \neq s'$; therefore, the additive model does not allow for interactions between characteristics; for example, the predictive power of the book-to-market ratio for expected returns does not vary with firm 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.

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

## *B* Comparison of Linear & Nonparametric Models

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

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

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

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

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

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

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

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

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

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

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

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

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

Assume the following return-generating process:

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

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

We now compare the performance of the linear and nonparametric model for the "10-1" hedge portfolio. The table below shows monthly returns, standard deviations, and

Sharpe ratios from a simulation for $2,000$ stocks and 240 periods for both models.[16]

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

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

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

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

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

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

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

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

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

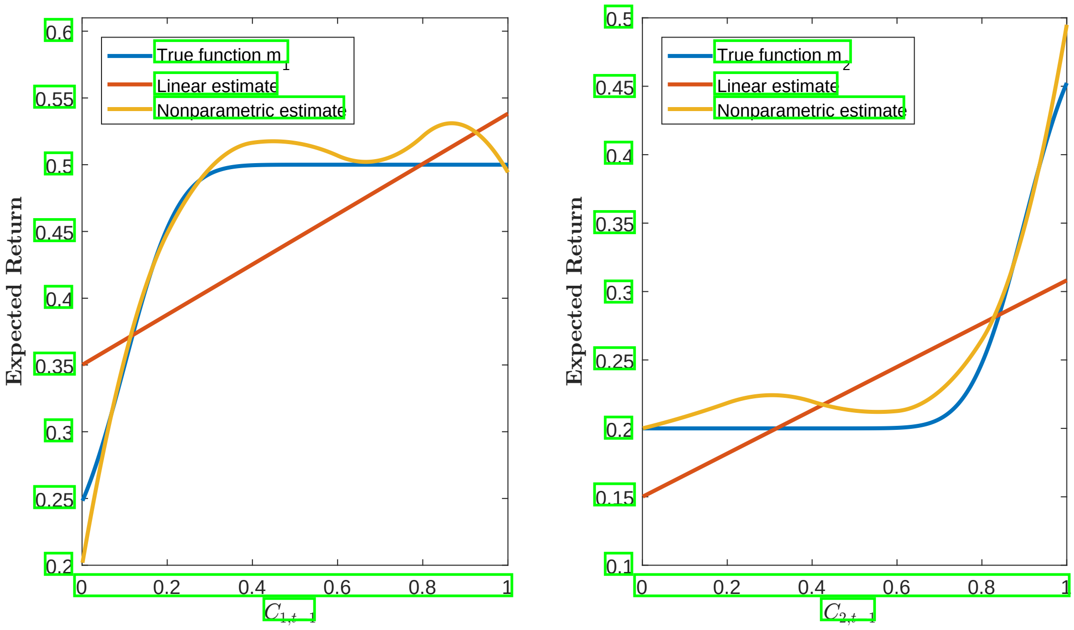


Figure 1: Regression functions and estimates

of the true data-generating process even though returns increase monotonically in both characteristics.[17]

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

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

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

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

[17] We thank Seth Pruitt for raising this point.

will play an important role.

## C Normalization of Characteristics

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

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

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, $\tilde{C}_{s,it-1} \in [0, 1]$. 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 $\alpha$ quantile of $\tilde{C}_{s,it-1}$ is $\alpha$. We use this particular transformation because portfolio sorting maps into our estimator as a special case.[18]

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

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

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

In summary, the transformation does not impose any additional assumptions, directly relates to portfolio sorting, and works well in finite samples because it appears more robust to outliers.[19]

## D Adaptive Group LASSO

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

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

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,

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

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

while setting functions for a given characteristic to 0 if the characteristic does not help predict expected returns. Therefore, the procedure achieves model selection; that is, it discriminates between the functions $\tilde{m}_{ts}$, which are constant, and the functions that are not constant.[20]

In portfolio sorts, we approximate $m_{ts}$ 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 $t_1, \ldots, t_{L-1}$ as knots and choose $t_l = l/L$ for all $l = 1, \ldots, L$ 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 $m_t$, we use *quadratic* splines; that is, we approximate $m_t$ 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 these properties, i.e.,

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

where $p_k(c)$ are known basis functions.[21]

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

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

[20] 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).

[21] 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.

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 $b_{sk}$ estimates and $\lambda_1$ is a penalty parameter.

The first part of equation (5) is just the sum of the squared residuals as in ordinary least squares regressions; the second part is the LASSO group penalty function. Rather than penalizing individual coefficients, $b_{sk}$, the LASSO penalizes all coefficients associated with a given characteristic. Thus, we can set the point estimates of an entire expansion of $\tilde{m}_t$ to 0 when a given characteristic does not provide independent information for expected returns. Due to the penalty, the LASSO is applicable even when the number of characteristics is larger than the sample size. In the application, we choose $\lambda_1$ in a data-dependent way to minimize a Bayes Information Criterion (BIC) proposed by Yuan and Lin (2006).

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

We first define the following weights:

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

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

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

We again choose $\lambda_2$ to minimize a BIC.

Huang et al. (2010) show $\breve{\boldsymbol{\beta}}_t$ is model-selection consistent; that is, it correctly selects the non-constant functions with probability approaching 1 as the sample size grows large. However, the estimators have unfavorable statistical properties because they are not oracle efficient. We re-estimate the model for the selected characteristics with OLS to address this problem.

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

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

If the cross section is sufficiently large, model selection and estimation can be performed period by period. Hence, the method allows for the importance of characteristics and the shape of the conditional mean function to vary over time. For example, some characteristics might lose their predictive power for expected returns. 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 V, we estimate our model over subsamples and also estimate rolling specifications to investigate the variation in the conditional mean function over time.

## *E* Confidence Bands

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

Let $p(\tilde{c}) = (p_1(\tilde{c}), \ldots, p_{L+2}(\tilde{c}))'$ be the vector of spline functions and let $\Sigma_{ts}$ 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 $\Sigma_{ts}$ and define $\hat{\sigma}_{ts}(c) = \sqrt{p(c)'\hat{\Sigma}_{ts} p(c)}$.

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

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

where $d_{ts}$ is a constant.

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

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

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

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

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

as the sample size increases.

## F Interpretation of the Conditional Mean Function

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

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

Then,

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

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

Let $c_s$ be a fixed value of a given transformed characteristic $s$, such as the mean or the median. Then,

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

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

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

# IV Data

Stock return data come from the Center for Research in Security Prices (CRSP) monthly stock file. We follow standard conventions and restrict the analysis to common stocks of firms incorporated in the United States trading on NYSE, Amex, or Nasdaq with market prices above USD 5.

Market equity (**ME**) is the total market capitalization at the firm level. **LME** is the total market capitalization at the end of the previous calendar month. **LTurnover** is the ratio of total monthly trading volume over total market capitalization at the end of the previous month. The bid-ask spread (**spread**) is the average daily bid-ask spread during the previous month. We also construct lagged returns over the previous month ($\mathbf{r_{2-1}}$), the previous 12 months leaving out the last month ($\mathbf{r_{12-2}}$), intermediate momentum ($\mathbf{r_{12-7}}$), and long-run returns from three years ago until last year ($\mathbf{r_{36-13}}$). We follow Frazzini and Pedersen (2014) in the definition of Beta (**Beta**), and idiosyncratic volatility (**Idio vol**) is the residual from a regression of daily returns on the three Fama and French factors in the previous month as in Ang, Hodrick, Xing, and Zhang (2006).

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

**AT** are total assets, **ATO** are sales scaled by net operating assets, and cash (**C**) is cash and short-term investments over total assets. **CTO** is capital turnover, **D2A** is depreciation and amortization over total assets, and **DPI2A** is the change in property, plant, and equipment. **E2P** is the earnings-to-price ratio. We define expenses to sales (**FC2Y**) as the sum of advertising expenses; research and development expenses; and selling, general, and administrative expenses over sales, and **Free CF** is net income and depreciation and amortization less the change in working capital and capex. **Investment** is the growth rate in total assets, **Lev** is the ratio of total debt to total debt and shareholders' equity, and **NOA** are net operating assets to lagged total assets. We define operating accruals (**OA**) as in Sloan (1996), and operating leverage (**OL**) is the ratio of cost of goods sold and selling, general, and administrative expenses over total assets. We define the price-to-cost margin (**PCM**) as sales minus cost of goods sold over sales, the profit margin (**PM**) as operating income after depreciation to net sales, gross profitability

(**Prof**) as gross profits over book value of equity, and **Q** is Tobin's Q. **Rel to high** is the closeness to the 52-week high price and **RNA** is the return on net operating assets. The return-on-assets (**ROA**) is income before extraordinary items to total assets and the return-on-equity (**ROE**) is the ratio of income before extraordinary items to lagged book value of equity. **S2P** is the ratio of sales to market capitalization, **SGA2S** is the ratio of selling, general, and administrative expenses to net sales, **spread** is the monthly average bid-ask spread, and **SUV** is standardized unexplained volume.

To alleviate a potential survivorship bias due to backfilling, we require that a firm has at least two years of Compustat data. Our sample period is July 1963 until June 2015. Table 1 reports summary statistics for various firm characteristics and return predictors. We calculate all statistics annually and then average over time. On average we have 2.5 million observations in our analysis.

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

# V Results

We now study which of the 36 characteristics we describe in Section IV provide independent information for expected returns, using the adaptive group LASSO for selection and estimation.

## A Selected Characteristics and Their Influence

Table 2 reports average annualized returns with standard errors in parentheses of 10 equally-weighted portfolios sorted on the characteristics we study. Most of the 36 characteristics have individually 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 3). Twenty-one sorts have annualized hedge returns of more than 5%, and 12 characteristics are even associated with excess returns of more than 10%. Twenty-one 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 independent information for expected returns.

To tackle the multi-dimensionality challenge, we now estimate the adaptive group LASSO with four, nine, 14, and 19 knots. The number of knots corresponds to the smoothing parameter we discuss in Section III. Nine knots corresponds to 10 portfolios in sorts.

Figure 7 and Figure 8 show the mean function, $\tilde{m}(\tilde{C}_{it-1})$, for Tobin's Q, return-on-assets, profitability, and investment. The left panels report the unconditional mean functions, whereas the right panels plot the association between the characteristic and expected returns conditional on all selected characteristics.

Stocks with low Q, low return on assets, investment, but high profitability have higher expected returns than stocks with high Q, return on assets, investment, or low profitability unconditionally. These results are consistent with our findings for portfolio sorts in Table 2. Portfolio sorts result in average annualized hedge portfolio returns of around 14%, 5%, 13%, and 6% for sorts on Q, return on assets, investment, and profitability, respectively. Profitability, Q, and investment also have t-stats relative to the Fama-French three-factor model substantially larger than the threshold Harvey et al. (2016) suggest (see Table 3).

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 all other firm characteristics we study. 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 with the characteristics. The constant conditional mean functions imply Q, return on assets, investment, and profitability have no marginal predictive power for expected returns once we condition on other firm characteristics.

The examples of Tobin's Q, return-on-asset, profitability, and investment show the importance of conditioning on other characteristics to infer on the predictive power of characteristics for expected returns. We now study this question for 36 firm characteristics using the adaptive group LASSO.

Table 4 reports the selected characteristics of the nonparametric model for different numbers of knots, sets of firms, and sample periods. We see in column (1) that the baseline estimation for all stocks over the full sample period using 14 knots selects 15 out of the universe of 36 firm characteristics. The assets-to-market cap, total assets, beta, capital intensity, earnings to price, fixed costs to sales, idiosyncratic volatility, lagged market cap, lagged turnover, the closeness to the 52-week high, momentum, short-term reversal, long-term reversal, SG&A -to-market cap, and standardized unexplained volume all provide incremental information conditional on all other selected firm characteristics.

When we allow for a wider grid in column (2) with only nine knots, we also select cash to total assets, free cash flow, intermediate momentum, and the average bid-ask spread. We instead select the same characteristics when we impose a finer grid and estimate the group LASSO with 19 interpolation points (see column (3)).

We estimate the nonparametric model only on large stocks above the 10%-, 20%-, and 50%-size quintile of NYSE stocks in columns (3) to (5), reducing the sample size from more than 1 million observations to around 300,000. Assets-to-market cap, total assets, beta, fixed-costs to sales, and idiosyncratic volatility lose their predictive power for returns for a sample of firms above the 10%-size threshold compared to all stocks in column (1), whereas operating accruals becomes a significant return predictor. For firms above the 20%-size threshold of NYSE firms, we also see capital intensity, earnings to price, lagged turnover, and momentum lose the predictive power, but intermediate momentum becomes a significant return predictor. For the largest firms, only seven characteristics have significant incremental predictive power for expected returns, including the book-to-market ratio, the closeness to the high price, past return-based predictors, SG&A to market cap, and standardized unexplained volume.[22]

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 eight characteristics have

[22]The number of knots increases with the sample size. The penalty function instead increases in the number of knots. In the nonparametric model with nine knots, the penalty is proportional to 12 times the number of selected characteristics, which is why we select fewer characteristics with more knots.

predictive power for expected returns in the sample until 1990. All the characteristics we select in the first half of our sample still provide incremental information for expected returns in the second half starting in 1991, but now nine additional characteristics gain predictive power for expected returns.

Size, the closeness to the previous 52-week high, short-term reversal, and standardized unexplained volume are the most consistent return predictors across different sample periods, number of interpolation points, and sets of firms. Table A.1 in the online appendix reports selected characteristics for additional specifications.

Figure 9 and Figure 10 plot the conditional and unconditional mean functions for short-term reversal, momentum, size, and the assets-to-market-cap ratio. We see in Figure 9 both for reversal and momentum a more monotonic association between the characteristic distribution and expected returns once we condition on other characteristics in the right panel relative to the unconditional association in the left panels. Size matters for returns for all firms in the right panel of Figure 9 and the conditional association is more pronounced than the relationship in the left panel. This finding is reminiscent of Asness, Frazzini, Israel, Moskowitz, and Pedersen (2015), who argue "size matters, if you control your junk". We see in the lower panels, the assets-to-market-cap ratio is unconditionally positively associated with expected returns. Once we condition on other characteristics, though, the association flips sign, and stocks with a high ratio have lower expected returns compared to stocks with a lower ratio.

Figure 11 plots the conditional and unconditional mean functions for the book-to-market ratio and idiosyncratic volatility. Both book-to-market and idiosyncratic volatility have a monotonic, positive unconditional association with expected returns across the whole characteristic distribution. Conditional on other characteristics, book-to-market loses the predictive power for returns and only stocks with the highest idiosyncratic volatility are significantly associated with returns.

This section shows that many of the univariate significant return predictors do not provide incremental predictive power for expected returns once we condition on other stock characteristics.

## B 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 12 to Figure 15 show the conditional mean function for our baseline nonparametric model for all stocks and nine 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 normalize the conditional mean function to equal 0 when the normalized characteristic equals 0.5 (the median characteristic in a given month).

We see in Figure 12 that the conditional mean function is non-constant throughout the sample period for lagged market cap. Small firms have higher expected returns compared to large firms, conditional on all other significant return predictors. Interestingly, the size effect seems largest during the end of our sample period, contrary to conventional wisdom (see Asness et al. (2015) for a related finding). The bottom panel shows that firms with higher total assets have higher expected returns conditional on other firm characteristics, contrary to the unconditional association (see Table 2).

We see in the top panel of Figure 13 that intermediate momentum has a significant conditional association with expected returns throughout the sample period. Interestingly, past intermediate losers tend to outperform once we condition on other characteristics. In the bottom panel, we see momentum 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 14 shows the effect of long-term reversal on expected returns has been strongest in the modern sample period because past losers tend to appreciate more than they did historically. The bottom panel shows the association of idiosyncratic volatility and returns has been flat until the early 1990s and only afterwards did stocks with the highest level of idiosyncratic volatility earn substantially higher returns than all other stocks.

## C Out-of-Sample Performance and Model Comparison

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

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 1963 to 1990 and carry out model selection with the adaptive group LASSO with nine knots, but also use the adaptive LASSO for model selection in the linear model over the same sample period and with the same number of knots. We then use 10 years of data to estimate the model on the selected characteristics. In the next months, we take the selected characteristics and predict one-month-ahead returns and construct a hedge portfolio going long stocks with the 10% highest predicted returns and shorting stocks with the 10% lowest predicted returns. We 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 December 1980 until November 1990 to get estimators $\hat{\boldsymbol{\beta}}$.[23] We then take the estimated coefficients and characteristics data from January 1981 until December 1990 to predict returns for January 1991 and form two portfolios for each method. We buy the stocks with the 10% highest expected returns and sell the stocks with the 10% lowest predicted returns. We then move our estimation sample forward by one month from February 1981 until January 1991, get new estimators $\hat{\boldsymbol{\beta}}$, and predict returns for February 1991.

Table 5 reports the out-of-sample Sharpe ratios for both the nonparametric and linear models for different sample periods, number of knots, and equally- and value-weighted portfolios. For a sample from 1991 to 2014 and nine knots, the nonparametric model generates an out-of-sample Sharpe ratio for an equally-weighted hedge portfolio of 3.42

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

compared to 2.26 for the linear model (compare columns (1) and (3)).[24] The linear model selects 21 characteristics in sample compared to only eight for the nonparametric model, but performs worse out of sample.[25]

We see a substantial drop in out-of-sample Sharpe ratios both for the nonparametric and linear model when we study value-weighted portfolios (see columns (2) and (4)). The difference in Sharpe ratios between value- and equally-weighted portfolios is similar to Lewellen (2015), who finds Sharpe ratios of value-weighted portfolios 50% smaller than the equally-weighted counterpart. Most studies in empirical finance winsorize the data, including size. When we winsorize size at the 1% and 5% levels, we find out-of-sample value-weighted Sharpe ratios of 1.38 and 1.73 for the nonparametric model (results not reported).

Column (5) studies a linear model that also employs the rank transformation we discuss in Section III. The linear model now selects even 27 of the 36 characteristics, but the out-of-sample Sharpe ratio is similar to the linear model for non-transformed characteristics.

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 on the 21 characteristics the linear model selects (see column (6)). The linear model appears to overfit the data in sample. The Sharpe ratio for the linear model when we use the eight characteristics we select with the nonparametric model is identical to the one we find when we use the 21 characteristics the linear model selects (see column (7)).

We see in columns (8) and (9) that Sharpe ratios drop by more than 50% for both models when we exclude firms below the $10^{\text{th}}$ percentile of NYSE stocks (around 200,000 firm-months observations out of 650,000). Lewellen (2015) also finds Sharpe ratios for an equally-weighted hedge portfolio that are lower by 50% when he excludes "all but tiny stocks." Columns (10) to (12) show Sharpe ratios larger than 3 for an out-of-sample

[24] The linear model we estimate and the results are similar to Lewellen (2015). Specifically, he finds for a linear model out-of-sample Sharpe ratios of 1.65 and 0.85 for equally- and value-weighted hedge portfolios and an out-of-sample period of 1974 to 2013.

[25] The linear model might be misspecified and therefore select more variables (see discussion and simulation results in Section III).

period starting in 1975 using the first 10 years for model selection and estimation and for our baseline sample starting in 1991, or using 4 or 14 knots rather than 9 as in column (1).

We also studied whether the return forecasts of the nonparametric model actually picks up cross-sectional variation in ex-post realized returns at the firm level. A return estimate that provides an unbiased forecast of returns should predict ex-post returns with a slope of 1. For all stocks and a sample from 1991 until 2014, we find the return forecasts from the nonparametric model have a slope estimate of 0.78 and explain 3.11% of the ex-post variation in returns at the firm level. These estimates compare favorably to Lewellen (2015), who typically finds smaller point estimates and $R^2$s of below 1%.

# VI Conclusion

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

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

Our results are as follows: (i) Out of 36 characteristics, only 7 to 15 provide independent information depending on the number of interpolation points (similar to the number of portfolios in portfolio sorts), sample period, and universe of stocks (large versus small stocks). (ii) Substantial time variation is present in the predictive power of characteristics. (iii) The nonparametric model selects fewer characteristics than the linear model in sample and has a 50% higher Sharpe ratio out of sample.

We see our paper as a starting point only and ask the following questions. Are the characteristics we identify related to factor exposures? How many factors are important? Can we achieve a dimension reduction and identify $K$ factors that can summarize the $N$

independent dimensions of expected returns with $K << N$ similar to Fama and French (1993) and Fama and French (1996)?

# References


Ang, A., R. J. Hodrick, Y. Xing, and X. Zhang (2006). The cross-section of volatility and expected returns. *The Journal of Finance 61*(1), 259–299.

Asness, C. S., A. Frazzini, R. Israel, T. J. Moskowitz, and L. H. Pedersen (2015). Size matters, if you control your junk. *Unpublished Manuscript, University of Chicago*.

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

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

Bansal, R. and S. Viswanathan (1993). No arbitrage and arbitrage pricing: A new approach. *The Journal of Finance 48*(4), 1231–1262.

Banz, R. W. (1981). The relationship between return and market value of common stocks. *Journal of Financial Economics 9*(1), 3–18.

Barillas, F. and J. Shanken (2016). Which alpha? *Review of Financial Studies (forthcoming)*, hhw101.

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.

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.

Bryzgalova, S. (2016). Spurious factors in linear asset pricing models. Technical report, Unpublished Manuscript, Stanford University.

Burnside, C. (2016). Identification and inference in linear stochastic discount factor models with excess returns. *Journal of Financial Econometrics 14*(2), 295–330.

Bustamante, M. C. and A. Donangelo (2016). Product market competition and industry returns. *Review of Financial Studies (forthcoming)*.

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

Chapman, D. A. (1997). Approximating the asset pricing kernel. *The Journal of Finance 52*(4), 1383–1410.

Chinco, A., A. D. Clark-Joseph, and M. Ye (2015). Sparse signals in the cross-section of returns. *Available at SSRN 2606396*.

Chordia, T., A. Goyal, and J. A. Shanken (2015). Cross-sectional asset pricing with individual stocks: betas versus characteristics. *Unpublished Manuscript, Emory*

*University*.

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.

Connor, G., M. Hagmann, and O. Linton (2012). Efficient Semiparametric Estimation of the Famaï¿½French Model and Extensions. *Econometrica 80*, 713–754.

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 (2016). Flexible prices and leverage. *Unpublished Manuscript, University of Chicago*.

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

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

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 22*(5), 1915–1953.

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 K. R. French (2016). Dissecting anomalies with a five-factor model. *Review of Financial Studies 29*(1), 69–103.

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.

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.

Gospodinov, N., R. Kan, and C. Robotti (2014). Misspecification-robust inference in linear asset-pricing models with irrelevant risk factors. *Review of Financial Studies 27*(7), 2139–2170.

Green, J., J. R. Hand, and F. Zhang (2016). The characteristics that provide independent information about average us monthly stock returns. *Review of Financial Studies (forthcoming)*.

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. Klemela, 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*.

Jagannathan, R. and Z. Wang (1998). An asymptotic theory for estimating beta-pricing models using cross-sectional regression. *The Journal of Finance 53*(4), 1285–1309.

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.

Kan, R. and C. Zhang (1999). Two-pass tests of asset pricing models with useless factors. *Journal of Finance 54*(1), 203–235.

Kleibergen, F. (2009). Tests of risk premia in linear factor models. *Journal of econometrics 149*(2), 149–173.

Kleibergen, F. and Z. Zhan (2015). Unexplained factors and their effects on second pass r-squared's. *Journal of Econometrics 189*(1), 101–116.

Kozak, S., S. Nagel, and S. Santosh (2015). Interpreting factor models. *Unpublished Manuscript, University of Michigan*.

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

Light, N., D. Maslov, and O. Rytchkov (2016). Aggregation of information about the cross section of stock returns: A latent variable approach. *Review of Financial Studies (forthcoming)*.

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

Lucas, 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.

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

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

Rosenberg, B., K. Reid, and R. Lanstein (1985). Persuasive evidence of market inefficiency. *The Journal of Portfolio Management 11*(3), 9–16.

Ross, S. A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory 13*(3), 341–360.

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.

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.