Online Appendix: Dissecting Characteristics Nonparametrically

Not for Publication

Joachim Freyberger — University of Wisconsin-Madison
Andreas Neuhierl — University of Notre Dame
Michael Weber — University of Chicago

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 (2018).

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.2.1 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."

A.2.2 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. 1.3 below.

A.3.1 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

[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.

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)).

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 1.3.

A.3.2 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 1.3 (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.

A.3.3 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, that is,

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

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. Becuase of 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 2, we estimate our model over subsamples and also estimate rolling specifications to investigate the variation in the conditional mean function over time.

A.3.4 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 2.

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 1.3.

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 1.3 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 $\tilde{m}_{ts}(\tilde{c}_s) - \int \tilde{m}_{ts}(\tilde{c}_s) d\tilde{c}_s$

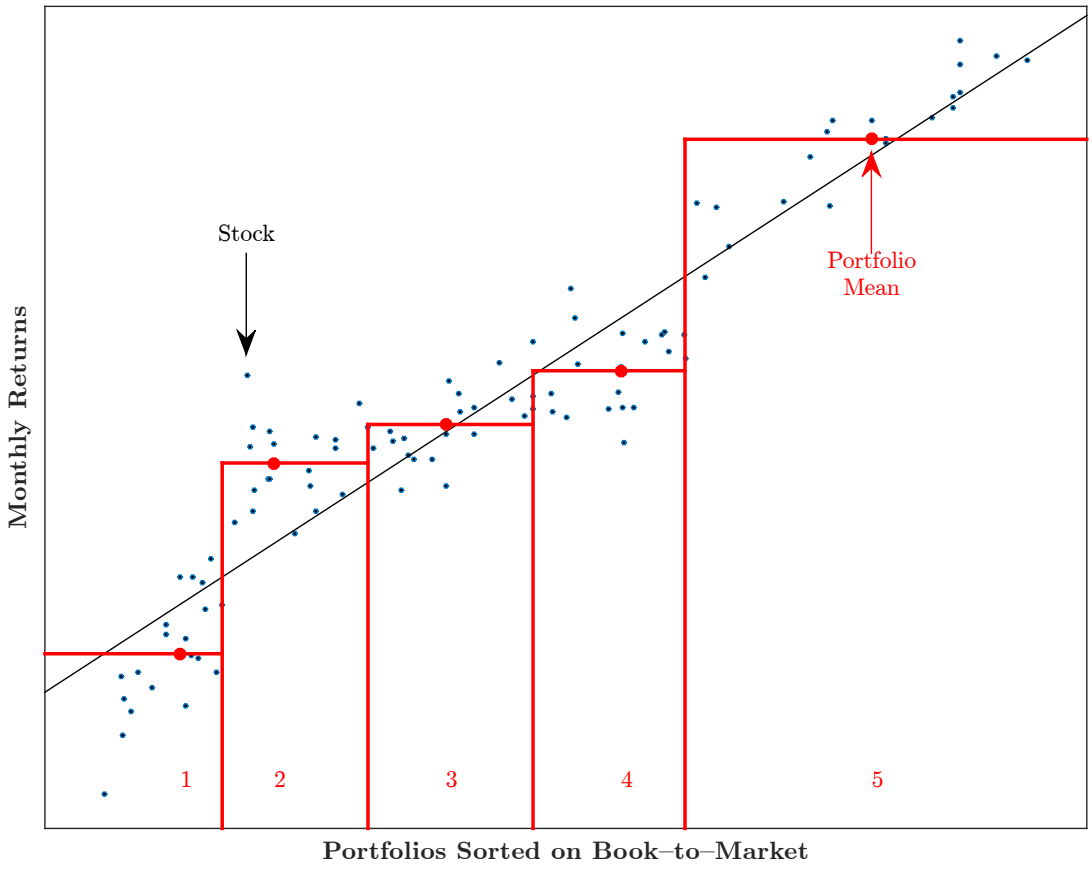
is linear at a significance level of $1-\alpha$. But since $\tilde{m}_{ts}(\tilde{c}_s) - \int \tilde{m}_{ts}(\tilde{c}_s)d\tilde{c}_s$ is linear if and only if $\tilde{m}_{ts}(\tilde{c}_s)$ is linear, we can then also reject the null hypothesis that $\tilde{m}_{ts}(\tilde{c}_s)$ is linear. Similar, 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.6.1 Knot Selection

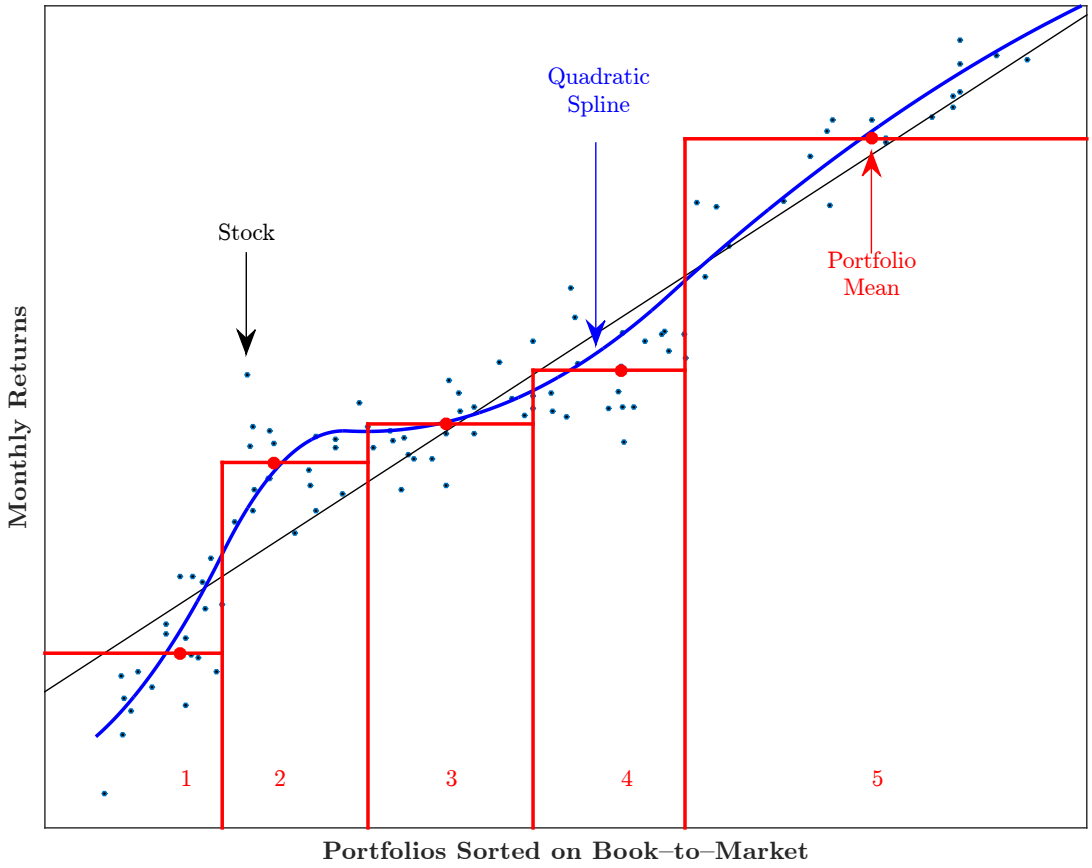
Theory tells us the number of interpolation points should grow as the sample size grows. Empirically, this statement is not too helpful in guiding our choices. We therefore document that the number and identify of characteristics is stable for reasonable variations in the number of knots (see Figure 3 which we discuss below).

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



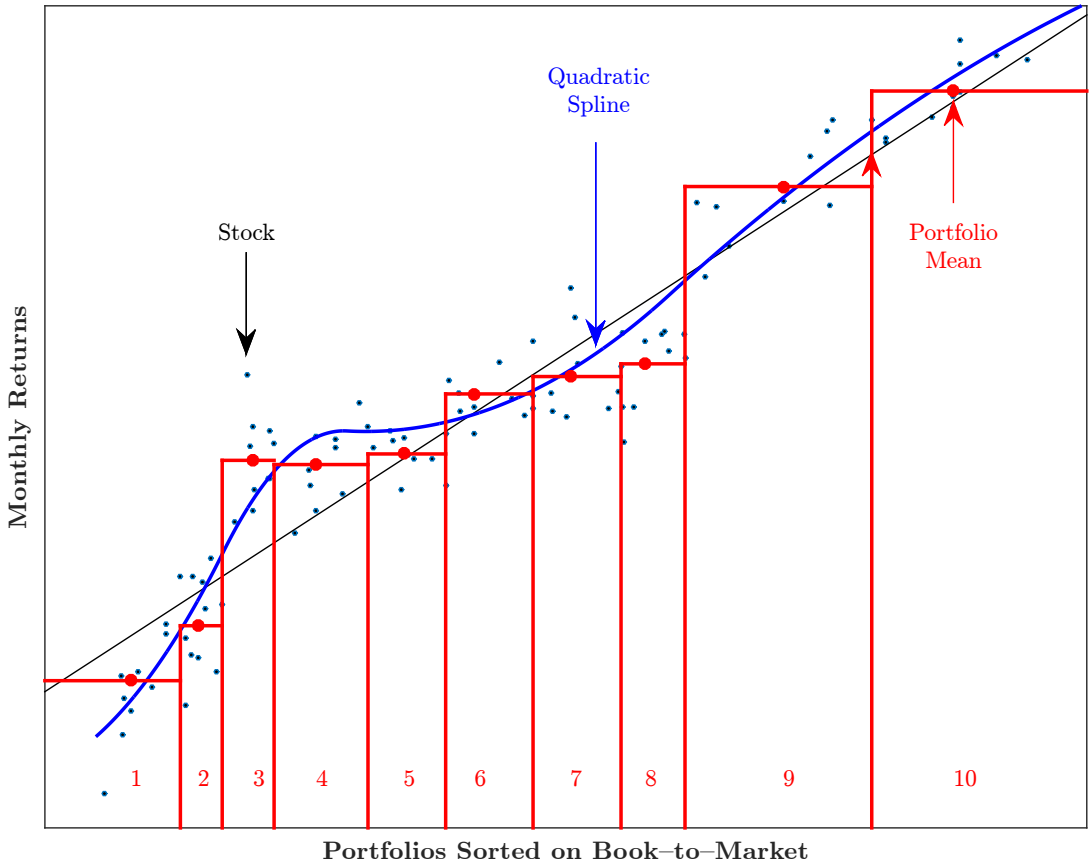
This figure plots returns on the y-axis against the book-to-market ratio on the x-axis as well as portfolio mean returns for simulated data.

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



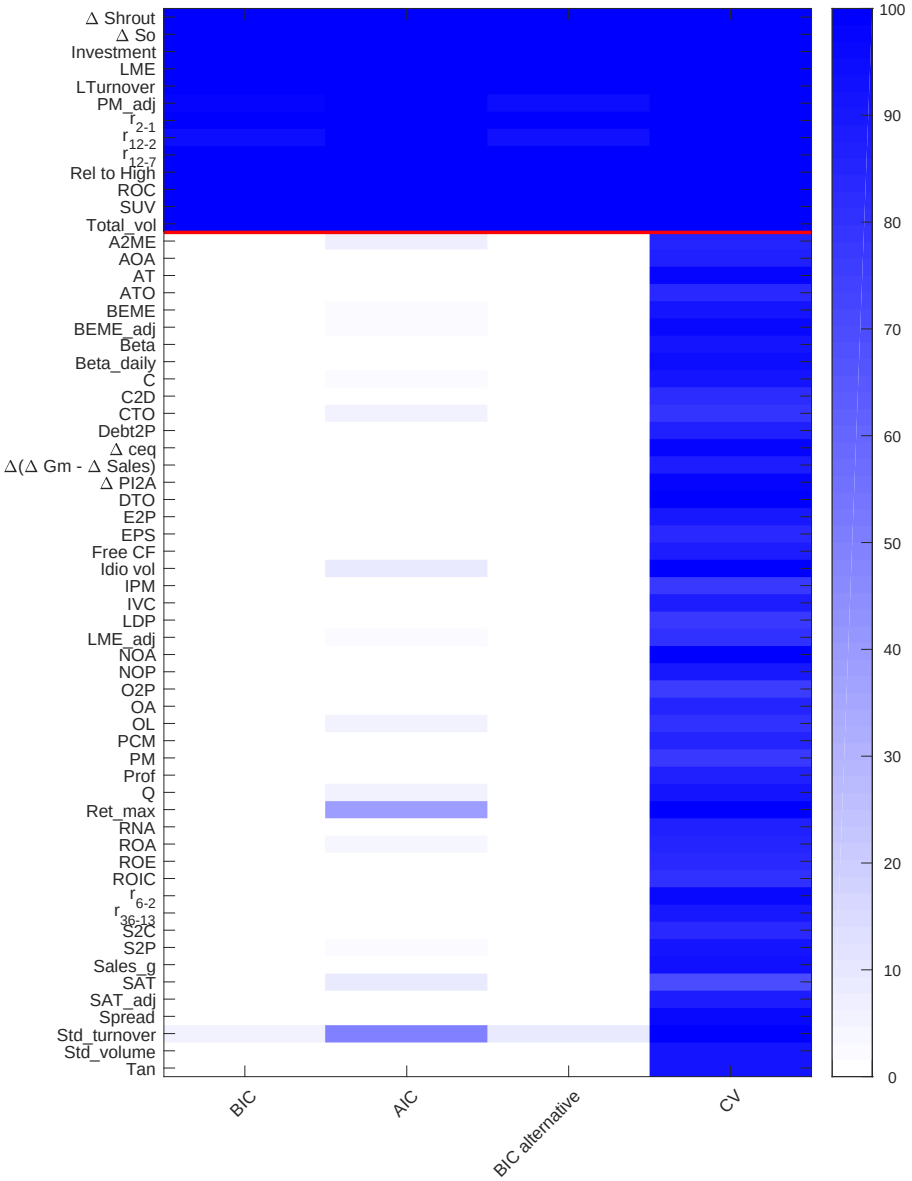
This figure plots returns on the y-axis against the book-to-market ratio on the x-axis as well as portfolio mean returns and a nonparametric conditional mean function for simulated data.

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



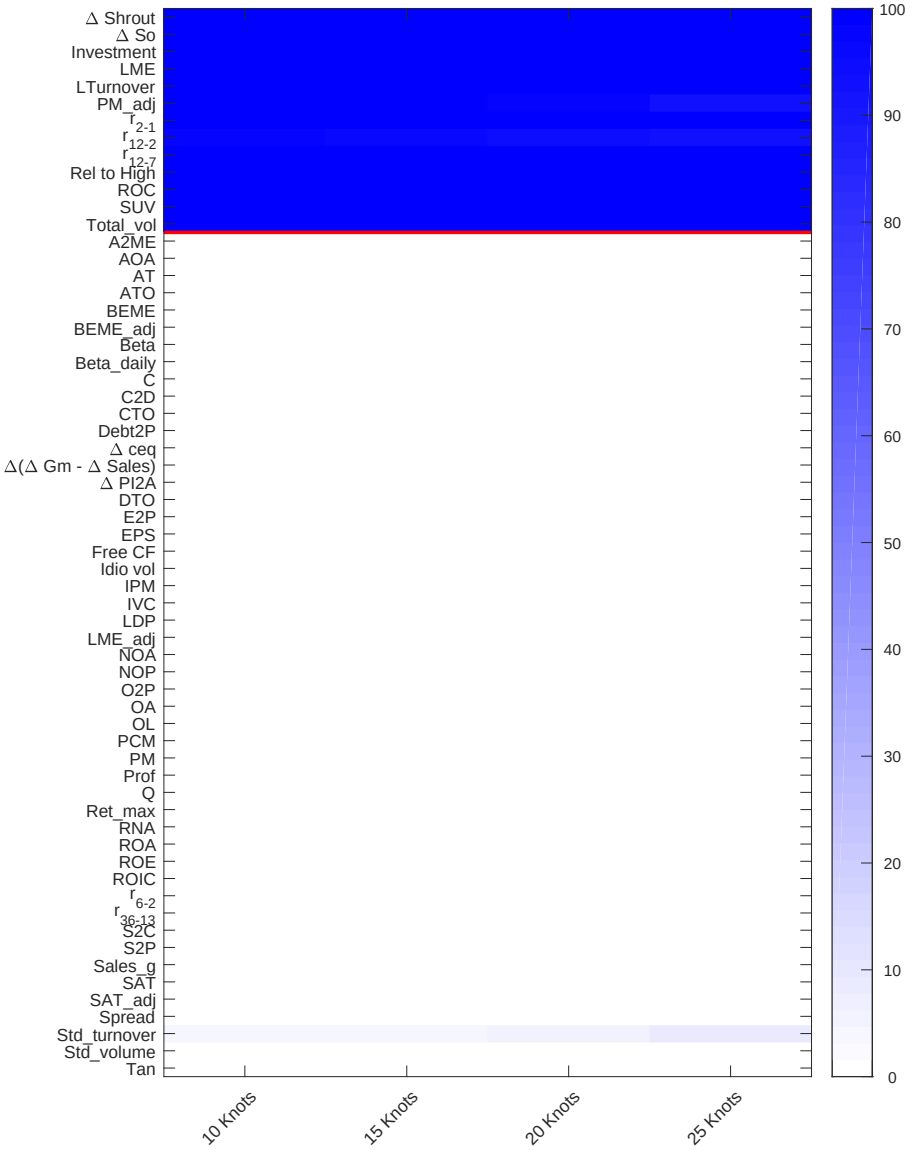
This figure plots returns on the y-axis against the book-to-market ratio on the x-axis as well as portfolio mean returns and a nonparametric conditional mean function for simulated data.

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



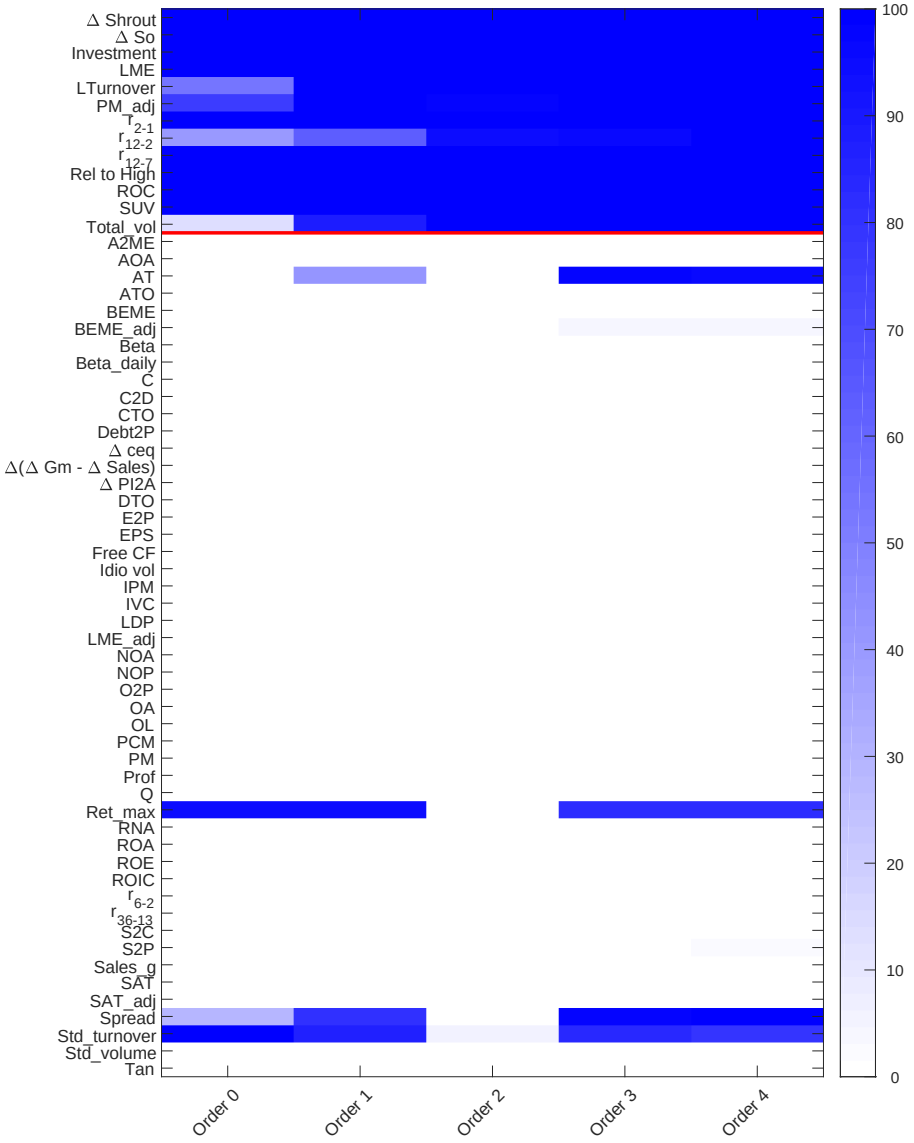
The figure graphically shows for the nonlinear adaptive group LASSO for different information criteria the frequency with which characteristics from the universe of 62 firm characteristics we discuss in Section A.1 of the online appendix are selected by each information criteria. The darker the color, 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)**



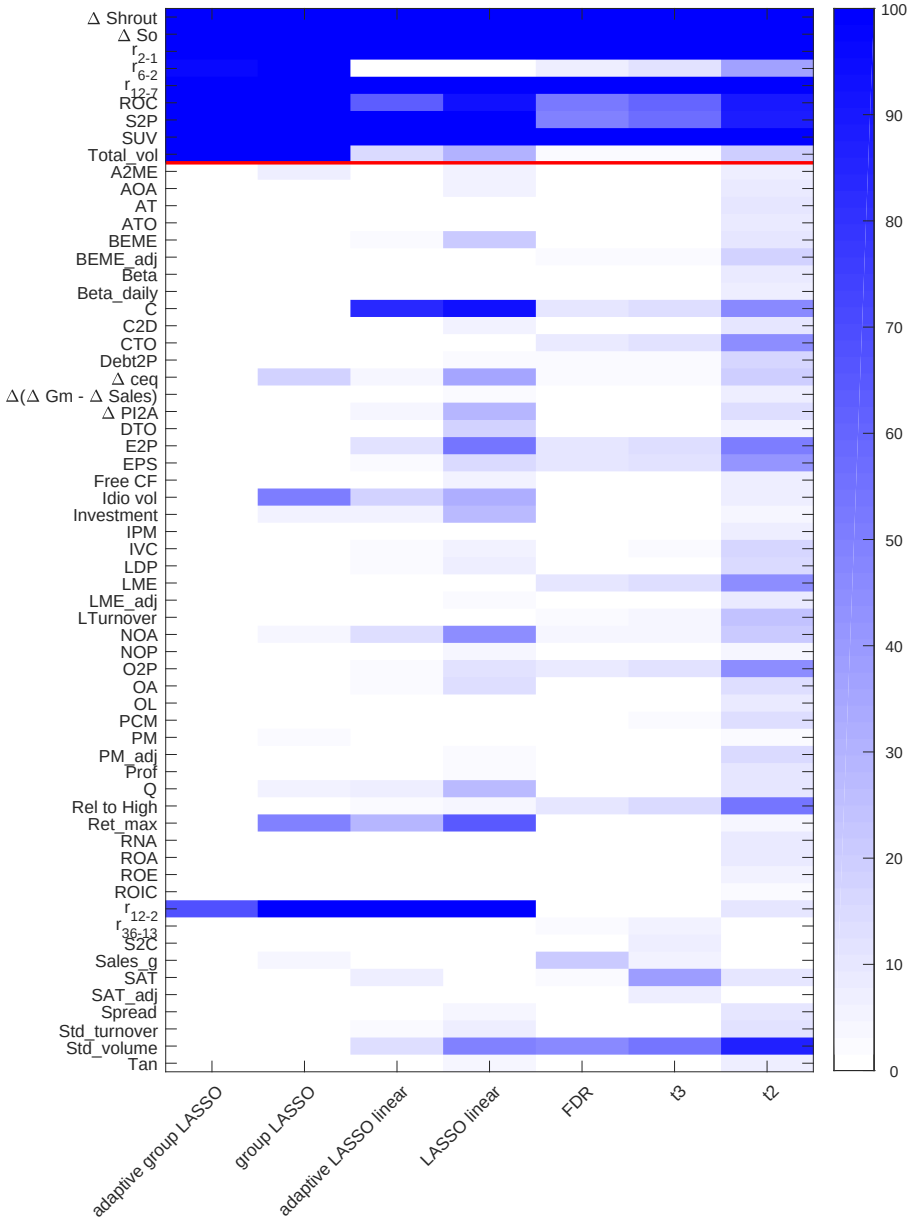
The figure graphically shows for the nonlinear adaptive group LASSO for different number of knots the frequency with which characteristics from the universe of 62 firm characteristics we discuss in Section A.1 of the online appendix are selected by each information criteria. The darker the color, the 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%