The Annals of Applied Statistics
2020, Vol. 14, No. 4, 1622–1650
https://doi.org/10.1214/20-AOAS1351

MONOTONIC EFFECTS OF CHARACTERISTICS ON RETURNS

BY JARED D. FISHER[1], DAVID W. PUELZ[2] AND CARLOS M. CARVALHO[3]

[1]*Department of Statistics, University of California, Berkeley, jared.fisher@berkeley.edu*
[2]*Booth School of Business, University of Chicago, david.puelz@chicagobooth.edu*
[3]*McCombs School of Business, University of Texas at Austin, carlos.carvalho@mccombs.utexas.edu*

This paper considers the problem of modeling a firm's expected return as a nonlinear function of its observable characteristics. We investigate whether theoretically-motivated monotonicity constraints on characteristics and non-stationarity of the conditional expectation function provide statistical and economic benefit. We present an interpretable model that has similar out-of-sample performance to black-box machine learning methods. With this model, the data provide support for monotonicity and time variability of the conditional expectation function. Additionally, we develop an approach for characteristic selection using loss functions to summarize the posterior distribution. Standard unexplained volume, short-term reversal, size, and variants of momentum are found to be significant characteristics, and there is evidence this set changes over time.

1. Introduction. This paper considers the problem of predicting a firm's stock return with observable firm characteristics. These characteristics may be accounting measures, such as market capitalization and book value as well as other observables such as a firm's past performance. In this paper, returns refer to excess returns—the return on an asset that exceeds the "risk-free" rate of return on short-term Treasury bonds. Let r_{it} be the excess return of firm i at time t, and let characteristics be incorporated into the vector $\boldsymbol{x}_{i,t-1}$. The conditional expectation function

$$\mathbb{E}(r_{it} \mid \boldsymbol{x}_{i,t-1}) = f(\boldsymbol{x}_{i,t-1}) \tag{1.1}$$

is the object of interest. This paper accomplishes two goals. First, we develop a flexible Bayesian model for f. We carefully examine the statistical benefits of theoretically-motivated monotonicity constraints and time variation for our case study. These model features are previously unexplored in the finance literature, and we adapt methods from Shively, Sager and Walker (2009) and McCarthy and Jensen (2016) to accomplish this goal. Second, we present a decision-theoretic framework for identifying the most predictive characteristics within the vector $\boldsymbol{x}_{i,t-1}$, extending recent work in posterior summarization (Hahn and Carvalho (2015), Puelz (2018), Puelz, Carvalho and Hahn (2015), Puelz, Hahn and Carvalho (2017), Puelz, Hahn and Carvalho (2020)) to nonlinear models.

Discovery of monotonic relationships in finance began decades ago. Fama and French (1993) found that, on average, smaller firms have higher returns than larger firms. Jegadeesh and Titman (1993) and Jegadeesh and Titman (2001) documented that, on average, previously well-performing firms (past winners) continue to do well in the near future, and past losers have low returns in the future. Patton and Timmermann (2010) develop statistical tests for monotonicity in assets' returns. However, work still remains to understand the modeling impact of these revelations. In statistics, incorporating monotonicity constraints into models is a known but underutilized tool; see Shively, Sager and Walker (2009) and Chipman et al.

Received June 2019; revised March 2020.
Key words and phrases. Cross-section of returns, posterior summarization, Bayesian modeling.

(2019) for recent developments. This paper presents a case study that combines decades-old empirical beliefs of monotonicity with this exciting and new statistical modeling work.

The list of potentially predictive characteristics is long and continues to grow, and numerous studies in finance have shown that these characteristics are independently useful for modeling returns. Harvey, Liu and Zhu (2016) catalog over 300 such characteristics and factors. Recently termed the "Factor Zoo" by Cochrane (2011) due to the sheer number of proposed characteristics and factors, the presence of hundreds is misleading; however, as many characteristics are likely drawing information from the same latent attributes of these firms and the economy. Understanding f as well as its characteristic inputs is a venerable and urgent case study in finance and asset pricing. Hence, this paper will address the following questions: *Which characteristics are important?* And furthermore: *When are these characteristics important?* Of course, the size of $\boldsymbol{x}_{i,t-1}$ and the stationarity of the relationships between characteristics and returns depend on the choice of f, which brings us to a third question: *What do these relationships look like?*

1.1. *Literature and contributions*. A traditional approach for understanding f is modeling the cross-section of firm returns as linearly related to a set of firm characteristics. Finance data, especially company return data, is a low-signal, high-noise environment, and structure is helpful to deal with this tremendous noise. Linear regression represents one extreme of model structure and simplicity, and most papers have at least some regression analysis—most famously the methods presented in Fama and MacBeth (1973). These methods are popular not only because of their structure but also because of their interpretability. Linear regression is widely known, easily estimated and returns a single number representing the relationship between X and Y (the slope). Yet, as Freyberger, Neuhierl and Weber (2020) state: "no a priori reason exists why the conditional mean function should be linear." The core assumption of this standard approach may not hold. Therefore, recent literature considers nonlinear models for the return-characteristic relationship (Freyberger, Neuhierl and Weber (2020), Gu, Kelly and Xiu (2020)). Papers such as Gu, Kelly and Xiu (2020) utilize machine learning (ML) methods to infer nonlinear and joint relationships among characteristics and lie at the other modeling extreme: highly flexible but minimally interpretable. In this paper we show that these minimally-interpretable ML methods (trees/forests specifically) provide surprisingly little advantage in predictive ability within the application area of finance and especially asset pricing.

An alternative, nonparametric, nonlinear approach for modeling f that maintains interpretability is *portfolio sorting*. This is done by cross-sectionally ranking firms based on an explanatory variable and computing the average firm return within each decile (or other quantile). Cattaneo et al. (2020) and Freyberger, Neuhierl and Weber (2020) show that this approach is fitting a step function to return-characteristic relationship, as opposed to a linear fit typically from regression. However, a step function is a simplistic functional form of the the return-characteristic relationship, as it must be assumed constant within deciles and no information is shared across deciles. Fama and French (2008) summarize these issues in saying "sorts are clumsy for examining the functional form of the relation between average returns and an anomaly variable." Additionally, one quickly encounters dimensionality issues. Fitting a mean to each sorted decile of p variables requires 10^p datapoints which very quickly is not plausible. Furthermore, Cattaneo et al. (2020) show that using 10 portfolios (deciles) is not enough and that it is optimal to use more.

The methodology presented in this paper is most similar to Freyberger, Neuhierl and Weber (2020). We model f using additive quadratic splines, and this provides interpretability and flexibility. Our paper differs from Freyberger, Neuhierl and Weber (2020) in four significant ways: We (*i*) characterize uncertainty through a fully Bayesian framework, (*ii*) examine the

theoretical and statistical benefits of monotonicity constraints incorporated through priors, (*iii*) account for time variation through a first-principled, power-weighting density approach and (*iv*) utilize statistical uncertainty to select the meaningful characteristics at each point in time. Standard unexplained volume, short-term reversal, market capitalization (size) and variants of momentum are found to be significant characteristics, and there is evidence this set changes in time. The data also provide support for monotonicity and time variability of the conditional expectation function.

The rest of the paper proceeds as follows. Section 2 details the modeling methodology which relates to contributions (*i*)–(*iii*) above. Section 3 presents a simulation study to describe the merit of using monotonicity for structure and using power-weighting densities for nonstationarity. Section 4 details the posterior summarization approach that is used to select the meaningful characteristics which corresponds to contribution (*iv*) above. Section 5 reports the results from both the modeling and selection processes. Section 6 concludes.

2. Modeling methodology. As discussed in Section 1, our model is comprised of the following components motivated by the case study: interpretability through additivity, flexibility through nonlinearity, minimal/specific structure through monotonicity, uncertainty through Bayesian priors and nonstationarity through weighted densities. We outline each component in detail below.

Interpretability via an additive model. We address the first modeling objective by using an additive model such that each characteristic's effect is separable from the others. Let

$$\mathbb{E}(r_{it}|\boldsymbol{x}_{i,t-1}) = \alpha_t + \sum_{k=1}^{K} f_{kt}(x_{i,k,t-1}), \tag{2.1}$$

where r_{it} is the time t return for firm i, α_t is the intercept term for time t and $\boldsymbol{x}_{i,t-1}$ is a K length vector of firm i's characteristics at time $t-1$, where each characteristic is individually ranked across all n_{t-1} firms at time $t-1$

$$x_{i,k,t-1} = \frac{\text{rank}_{k,t-1}(\text{characteristic}_{i,k,t-1})}{n_t + 1}. \tag{2.2}$$

Thus, $x_{i,k,t-1} \in (0, 1)$ is the empirical quantile of characteristic k for firm i at time $t-1$. This rank transformation is done to eliminate two issues with the predictors variables: (*i*) outliers and (*ii*) changes in the range of characteristics over time. For example, the market capitalization (size) of firms in general has increased, and a one-billion-dollar firm today might be in the 10th percentile of size while 30 years ago it was in the 90th percentile. Using the "empirical percentiles" from the rank transformation eliminates these issues as we only look at a firm's relative place in the distribution of a given characteristic; Freyberger, Neuhierl and Weber (2020) scale characteristics in the same way.

However, we propose a novel adjustment. The intercept in equation (2.1) is the expected return when all x's are zero and, under the rank transformation, $x_{i,k,t-1} = 0$, $\forall k$ means the smallest possible value for x across all variables. The intercept α_t in equation (2.1) would be interpreted as the average return for a "perfectly minimum" firm, that is, a firm with the lowest value of each characteristic across all firms. This firm does not reasonably exist. As such, we shift the x-space by setting

$$x_{i,k,t-1} = \frac{\text{rank}_{k,t-1}(\text{characteristic}_{i,k,t-1})}{n_t + 1} - 0.5 \tag{2.3}$$

such that $x_{i,k,t-1} \in (-0.5, 0.5)$. Now, the intercept α_t represents the average return for a "perfectly median" firm, that is, a firm that has the median value across all characteristics.

Nonlinearity through quadratic splines. We address the second modeling objective through the use of quadratic splines. Typically, this would mean

$$f(x) = \beta_1 x + \beta_2 (x)^2 + \beta_3 (x - \acute{x}_1)_+^2 + \cdots + \beta_{\acute{m}+2} (x - \acute{x}_m)_+^2 \tag{2.4}$$

for m knots, $0 < \acute{x}_1 < \cdots < \acute{x}_m < 1$, where $(y)_+ = \max(0, y)$.

However, our intercept adjustment requires an adjustment to the standard notation. Let f_{kt} be the quadratic spline for characteristic k at time t. For now, we'll drop the ikt subscripts for simplicity. For a given series of $\grave{m} + 1$ nonpositive knots ($\grave{x}_{\grave{m}} < \cdots < \grave{x}_1 < \grave{x}_0 = 0$) and $\acute{m} + 1$ nonnegative knots ($0 = \acute{x}_0 < \acute{x}_1 < \cdots < \acute{x}_{\acute{m}}$), we set

$$f(x) = \beta_1 x + \beta_2 (x)_-^2 + \beta_3 (x - \grave{x}_1)_-^2 + \cdots + \beta_{\grave{m}+2} (x - \grave{x}_{\grave{m}})_-^2, \tag{2.5}$$

$$+ \beta_{\grave{m}+3} (x)_+^2 + \beta_{\grave{m}+4} (x - \acute{x}_1)_+^2 + \cdots + \beta_{\grave{m}+\acute{m}+3} (x - \acute{x}_{\acute{m}})_+^2, \tag{2.6}$$

where the $(y)_+ = \max(0, y)$ and $(y)_- = \min(0, y)$. This can be abbreviated as $f(x) = \boldsymbol{x}^{*\prime}\boldsymbol{\beta}$ where $\boldsymbol{x}^*$ is the carefully constructed quadratic spline basis.

Structure imposed through monotonicity. Theoretical or a priori information can be used to add structure to these splines. We implement this through monotonicity constraints. Without loss of generality we create these splines to be nondecreasing (can be nonincreasing) using the ideas of Shively, Sager and Walker (2009), Section 3, adapted to have both positive and negative knots.

By definition, the spline is monotonic nondecreasing if the first derivative is nonnegative for all x: $f'(x) \geq 0$. While specifications are in Appendix B, we suffice it here to say that the above restriction yields $\grave{m} + \acute{m} + 3$ linear constraints to satisfy, which can be summarized in a lower triangular matrix. We label this matrix $\boldsymbol{L}$ such that $\boldsymbol{0} \leq \boldsymbol{L\beta} = \boldsymbol{\gamma}$, and we see that $\boldsymbol{L}$ acts as a projection matrix, projecting our more complicated constraints on $\boldsymbol{\beta}$ to the simple nonnegative constraints on $\boldsymbol{\gamma}$. Hence,

$$f(x) = \boldsymbol{x}^{*\prime}\boldsymbol{\beta} = \boldsymbol{x}^{*\prime} L^{-1} L\boldsymbol{\beta} = \boldsymbol{w}'\boldsymbol{\gamma}, \tag{2.7}$$

where $\boldsymbol{w}' = \boldsymbol{x}^{*\prime} L^{-1}$ is now our modified spline basis. Returning the use of subscripts ikt, equation (2.1) is now

$$\mathbb{E}(r_{it} | \boldsymbol{x}_{i,t-1}) = \alpha_t + \sum_{k=1}^{K} \boldsymbol{w}'_{ikt} \boldsymbol{\gamma}_{kt}. \tag{2.8}$$

We allow our splines to be monotonic if there is prior information about the direction of a relationship between a firm characteristic and its stock return. For example, if we believe that a smaller firm will, on average, have higher returns than a larger firm, regardless of their absolute size (Fama and French (1993)), then we believe size should have a monotonic relationship with expected returns. Monotonicity is one of the less intrusive structures we can assume to reign in the flexibility, and potential overfit, of splines. We demonstrate that enforcing monotonicity has statistical benefits as well as a useful interpretation. When the data is especially noisy, monotonicity is helpful in decreasing the variability of the inferred relationship between stock returns and characteristics.

Bayesian model specification. With equation (2.8) introduced, we can describe the statistical model on our uncertainty. Let

$$r_{it} = \alpha_t + \sum_{k=1}^{K} \boldsymbol{w}'_{ikt} \boldsymbol{\gamma}_{kt} + \epsilon_{it} \tag{2.9}$$

with $\epsilon_{it} \sim N(0, \sigma^2)$.

We now set a prior on the coefficients γ. To protect against overspecifying the number of knots, we include shrinkage as an important part of this prior. Let $I_{jkt} = 1$ indicate that $\gamma_{jkt} > 0$ and $I_{jkt} = 0$ indicate that $\gamma_{jkt} = 0$, where j indexes the $\grave{m} + \acute{m} + 3$ coefficients. Thus, I_{jkt} is a Bernoulli random variable with prior probability $P(I_{jkt} = 1) = p_{jk}$. This leads us to the conditional prior on γ_{jkt},

$$(\gamma_{kj}|I_{kj} = 1, \cdot) \sim N_+(0, c_k\sigma^2), \tag{2.10}$$

where N_+ indicates a truncated Normal distribution with support on positive numbers (to change this entire setup to monotonic decreasing splines, we would simply change the support to negative numbers and appropriately adjust the definition of I_{jkt} above).

This setup allows us to let the data select the knots for the splines. By overspecifying the number of potential knots, the data will inform the model as to which knots should be included ($I_{jkt} = 1$) and which should not ($I_{jkt} = 0$). This way, we include more knots than needed, and the shrinkage instigated by our priors will remove knots that are not supported by the data.

Following Shively, Sager and Walker (2009), we place uninformative priors on $\sigma^2 \sim U(0, 10^3)$ and $\alpha \sim N(0, 10^{10})$, as well as set $p_{jk} = 0.2, \forall j, k$. c_k is chosen, $\forall k$, to be 2253.689, the average number of firms in a quarter across all quarters.

Nonstationarity incorporated through power-weighted densities. While using all historic data (i.e., using all data up to and including time $t - 1$ to forecast time t events) is an option, this does not allow the parameters to adjust to different trends over time (nonstationarity). Hence, we look at two approaches. First, we look at the traditional rolling-window method, where a model uses the most recent M time periods only, dropping all time periods older than the cutoffs. In this paper, akin to much of the empirical finance literature, we use $M = 120$ months as well as $M = 60$ and 36 months to show the effect of different window lengths. There are, however, methods that allow the window length to be different for different assets (Ang and Kristensen (2012)) though here we simply apply the same nonstationarity approach for all firms over all points in time.

Second, we use the power-weighted likelihood approach of McCarthy and Jensen (2016). For $\omega_t \in [0, 1]$, such that $\omega_1 \leq \omega_2 \leq \cdots \leq \omega_\tau$, the likelihood at time $\tau \in \{1, \ldots, T\}$ discounts the impact of past data: $p(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_\tau|\Theta_\tau) = \prod_{t=1}^{\tau} p(\boldsymbol{r}_t|\Theta_\tau)^{\omega_t}$, to allow more recent data to receive more weight than older data, we choose $\omega_t = \delta^{\tau - t}$, for $\delta \in (0, 1]$. Hence, for $\delta = 0.99$, yesterday's ω is 99% of today's. Thus, these likelihoods have an asymptotic effective sample size of $\frac{1}{1-\delta}$, for example, $\frac{1}{1-0.99} = 100$.

McCarthy and Jensen (2016) point out that this is a simpler alternative to specifying a model for the evolution process itself. They also point out that the rolling-window method is a special case of these power-weights, such that $\omega_1 = \cdots = \omega_{\tau-120} = 0$ and $\omega_{\tau-119} = \cdots = \omega_\tau = 1$.

3. Simulation. ***Why monotonicity?*** When modeling functional phenomena, if the underlying generative function is in fact monotonic, then assuming monotonicity will improve the model. Specifically, the uncertainty about the fitted curve will be smaller, or, in other words, the posterior will be more precise. In Figure 1, we present a monotonic increasing mean function. The gray data points are randomly generated with heteroskedastic noise. Here, we model the data using varying monotonicity constraints. Posterior curve draws are shown in pink, and the posterior mean curve is in red. We see that, while the unconstrained quadratic spline fits the underlying function reasonably well, the monotonic constrained spline fits better. Lastly, enforcing inappropriate constraints, namely a nonincreasing constraint in this case, disables the model. Hence, adding wise constraints help models ignore more of the noise and better detect signal.

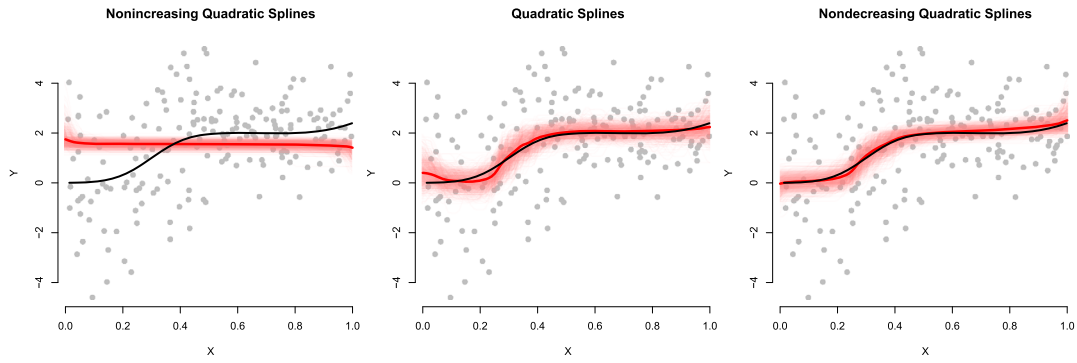

FIG. 1. *Data generated from a monotonically increasing mean function. Shown are three spline fits to the simulated data*: (*left*) *nonincreasing quadratic splines*, (*middle*) *quadratic splines without constraints and* (*right*) *nondecreasing quadratic splines*.

We believe there is signal in the firm characteristics data we're analyzing, but there is a lot of noise, so this property of the model is desirable. When there is a weak signal (a barely-nonzero generative function) but low noise, models perform about equally with and without monotonicity. Again, we show in the top row of Figure 2 where the generative curve is in black, data generated with homoskedastic noise in gray, the posterior draws in pink and the posterior mean curve in red.

However, as noise increases, the unconstrained spline tends to overfit to the data as in the bottom row of Figure 2, where the noise of the generative model is twice that of the top row. Note that the posterior uncertainty around the nondecreasing curve is visibly smaller than

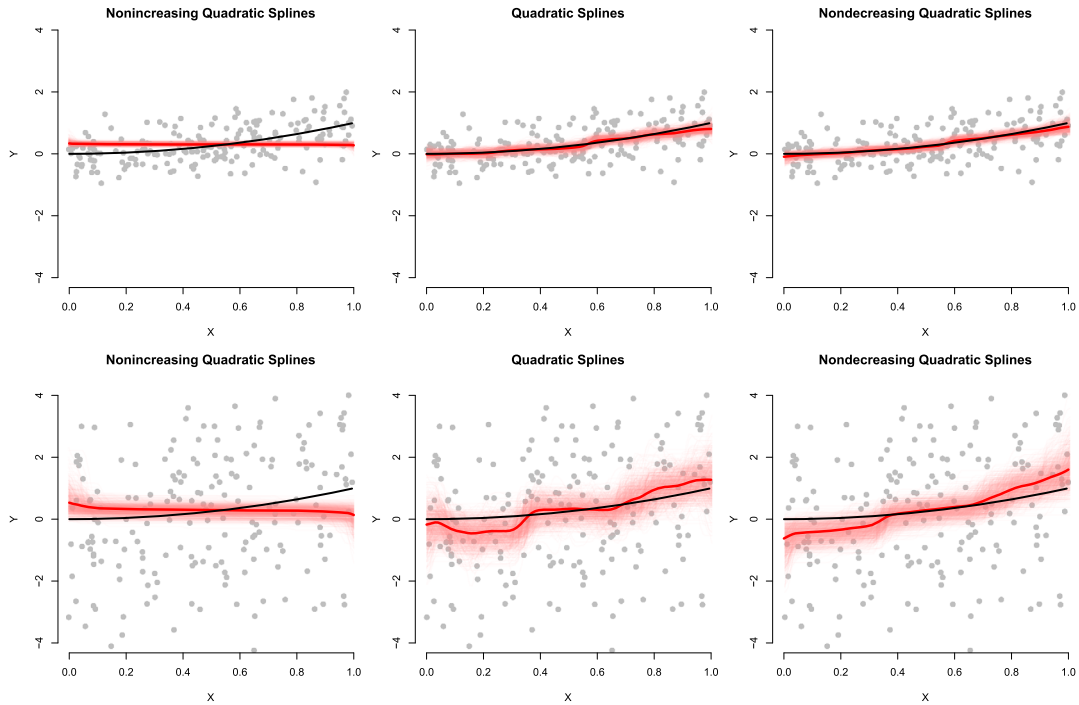

FIG. 2. *Data generated from a monotonically increasing mean function. Shown are three spline fits to the simulated data*: (*left*) *nonincreasing quadratic splines*, (*middle*) *quadratic splines without constraints and* (*right*) *nondecreasing quadratic splines. The top row is a low-noise environment, and the bottom row is a high-noise environment*.

the unconstrained spline. The unconstrained model can fit to the noise of the data instead of the underlying true function. In the bottom row the mean of the nondecreasing spline almost match the spirit of true function. Finally, we of course see that inappropriate constraints (nonincreasing) force the resulting model to fail entirely to model the underlying phenomena.

Why discount past information? Often, forecasts of future returns use all historical data, equally weighted. However, if the function of interest changes over time, then the more time between a past observation and the future time of interest, the less relevant that observation is.

As an example, consider the function $f(x, a) = ax^2$ as $a \to 0$. In Figure 3, we plot this parabola for $a \in \{10, \ldots, 1, 0\}$, such that at time 1 the function is $10x^2$, and at time 11 the function is 0. This is a parabola flattening over time, as illustrated by the random data points and their mean generative curve are fading from black to white. The pink curves are the draws from a power-weighted-discounted model with $\delta = 0.8$, where at time τ, $\omega_t = 0.8^{\tau - t}$, which implies an [asymptotic] effective sample size of $\frac{1}{1-\delta} = 5$, and the red curve is the posterior mean curve. The light blue lines are the MCMC sample curves from a historic-window model (all past time periods are equal weighted, so sample size is 11 time points), and the blue curve is their posterior mean. As displayed in Figure 3, allowing for time variation permits the model to better track the current state of a relationship that changes over time, as the red curve is closer to current function (flat) than the blue curve is.

It is important to highlight the differences between a rolling-window model and our proposed alternative. First, the rolling-window method is a special case of this model (see McCarthy and Jensen (2016)). Second, completely forgetting past data is not a desirable property. While older data is clearly not as valuable or pertinent as recent data, its value is not zero. Furthermore, if a 10-year rolling window is used, then data from 120 months ago is valued the same as today's, while data from 121 months ago is thrown away, as shown in the figure below. The arbitrary cutoff between 120 and 121 months does not reflect the true value of information on either side of that threshold. We propose that, in the case of time-varying phenomena, the importance of data decays as the data ages, akin to our power-weighting specified above. The exception to this are structural shocks that may occur, but even a 120-month rolling window will take 120 months to fully adapt. If adapting to shocks is the desired property, a structural break model should be used (Bekaert, Harvey and Lumsdaine

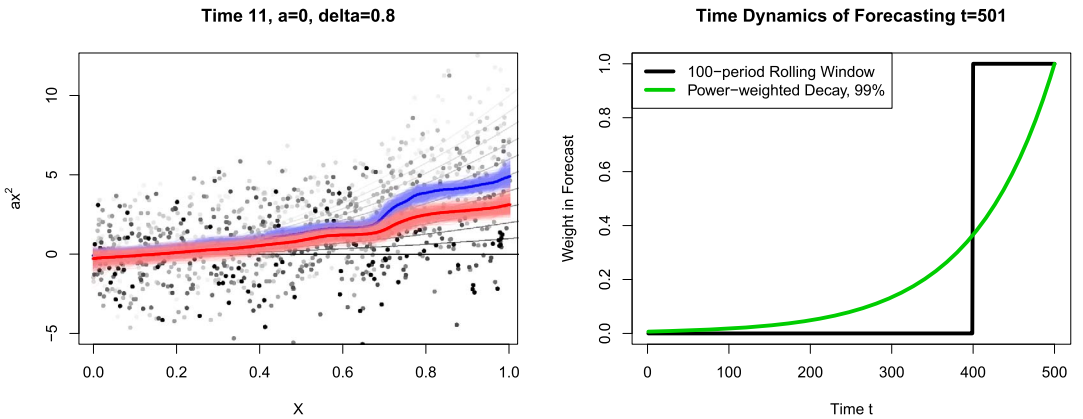

FIG. 3. (*left*) *Generated data from a parabolic function which collapses to a constant function over* 11 *time points. White to black points display the simulated data, with darker colors corresponding to points generated more recently. Also displayed are the function fits when treating each time point equitably* (*blue*) *vs. having decaying weights for increasingly further away time points* (*red*). (*right*) *Example weighting scheme for rolling-window weighting* (*black*) *versus time-decay weighting* (*green*).

(2002), Lettau and Nieuwerburgh (2008), Pastor and Stambaugh (2001)). However, incorporating the uncertainty around potential future structural breaks in one's model is a challenge (Pettenuzzo and Timmermann (2011), Smith and Timmermann (2018)) but will help better reflect model/parameter uncertainty if incorporated correctly. Smith, Timmermann and Zhu (2019) actually perform variable selection across structural breaks in a Bayesian model, though we pursue a different path to variable selection as described below.

4. Selection methodology. This section develops an approach for selecting meaningful firm characteristics. The aim is to identify characteristics of a firm that are predictive of its return and how this set varies in time. This approach builds upon the decision-theoretic selection procedure first proposed in Hahn and Carvalho (2015) and developed for econometric applications in Puelz, Hahn and Carvalho (2017), Puelz, Hahn and Carvalho (2020).

Rewriting the model as a predictive regression. As a first step, we rewrite our model as a predictive regression. Focusing on time t in the cross-section, the fully-specified model for the vector of n_t firm returns R_t is

$$
\begin{aligned}
&\mathrm{R}_t \sim \mathrm{N}\big(\alpha_t \mathbf{1}_{n_t} + \mathbf{X}_{t-1}\boldsymbol{\beta}_t, \sigma_t^2 \mathbb{I}_{n_t}\big) \quad \text{with } \mathbf{X}_{t-1}\boldsymbol{\beta}_t = \mathbf{W}_{t-1}\boldsymbol{\gamma}_t, \\
&\qquad \alpha_t \sim N\big(0, 10^{10}\big), \qquad \sigma_t^2 \sim U\big(0, 10^3\big), \\
&\qquad (\gamma_{jkt} | I_{jkt} = 1) \sim N_+\big(0, c_k \sigma_t^2\big), \qquad (\gamma_{jkt} | I_{jkt} = 0) = 0, \\
&\qquad I_{jkt} \sim \mathrm{Bn}(p_{jkt} = 0.2),
\end{aligned}
\tag{4.1}
$$

where $\mathbf{X}_{t-1}\boldsymbol{\beta}_t = \mathbf{X}_{t-1}\mathrm{diag}_K(\mathbf{L})^{-1}\mathrm{diag}_K(\mathbf{L})\boldsymbol{\beta}_t = \mathbf{W}_{t-1}\boldsymbol{\gamma}_t$. Note that $\mathrm{diag}_K(\mathbf{L})$ is a block diagonal matrix of size $K(\grave{m} + \acute{m} + 3) \times K(\grave{m} + \acute{m} + 3)$ where each lower triangular block is the projection matrix $\mathbf{L}$. Also, $\mathbf{X}_{t-1}$ is matrix of size $n_t \times K(\grave{m} + \acute{m} + 3)$, and $\boldsymbol{\beta}_t$ is vector of size $K(\grave{m} + \acute{m} + 3)$. Therefore, each firm is given a row in $\mathbf{X}_{t-1}$, and each $\grave{m} + \acute{m} + 3$ block of $\boldsymbol{\beta}_t$ corresponds to the coefficients on the spline basis for a particular characteristic, k. Incorporating the intercept directly into the characteristic matrix, we can write the generating model compactly as

$$
\mathrm{R}_t \sim \mathrm{N}\big(\mathbb{X}_{t-1}\mathbf{B}_t, \sigma_t^2 \mathbb{I}_{n_t}\big), \tag{4.2}
$$

where $\mathbb{X}_{t-1} = \big[\mathbf{1}_{n_t} \quad \mathbf{X}_{t-1}\big]$ and $\mathbf{B}_t = \big[\alpha_t \quad \boldsymbol{\beta}_t\big]$.

After rewriting our model more compactly, we delve into the second main contribution of this paper—firm characteristic selection in light of uncertainty. As described in the Introduction, there are many firm characteristics available for predicting returns. This leads to a natural question, which small subset of characteristics is *most relevant* for predicting the cross-section of firm returns? Further, does this subset vary over time?

Two components: Predictive uncertainty and loss. Suppose we have fit Model (4.1) using standard Monte Carlo methods. We now have access to the posterior distribution over all parameters: $p(\Theta_t \mid \text{past data} = \mathbf{R}_t)$. Also, conditional upon these posterior draws, we can simulate from the predictive distribution, providing draws from the joint distribution of future firm returns $\tilde{\mathrm{R}}_t$ and model parameters Θ_t, written as: $p(\tilde{\mathrm{R}}_t, \Theta_t \mid \text{past data} = \mathbf{R}_t)$. Uncertainty from the predictive is the first input for the selection procedure.

The second component is a rule for comparing models to one another; we call this our loss function. With both predictive uncertainty and a loss function in hand, we can ask and answer the pivotal question: *In light of uncertainty*, how do simpler models with fewer characteristics compare to the model including all characteristics? The decision-theoretic blend of these two components, a Bayesian model and a loss function, will allow us to discern which characteristics are important while taking uncertainty of all forms into account.

Optimizing expected loss and model selection. We formalize this methodology by first deriving the expected loss function. A natural measure for our characteristic selection goal is the log density of Regression (4.2). Note that (4.2) is not being used in a statistical capacity for model estimation but rather as a measure of how well a sparse representation of the linear predictor represents future data. The log density may be written as

$$\mathcal{L}(\tilde{\mathrm{R}}_t, \mathbf{A}_t) = \frac{1}{2}(\tilde{\mathrm{R}}_t - \mathbb{X}_{t-1}\mathbf{A}_t)^T(\tilde{\mathrm{R}}_t - \mathbb{X}_{t-1}\mathbf{A}_t), \tag{4.3}$$

where $\tilde{\mathrm{R}}_t$ is future return data at time t and $\mathbf{A}_t$ is the "action" to be taken by the data analyst. This action is intended to represent a sparse summary of the regression vector $\mathbf{B}_t$. In order to encourage sparsity in $\mathbf{A}_t$, we include an additional penalty function Φ with parameter λ_t,

$$\mathcal{L}_{\lambda_t}(\tilde{\mathrm{R}}_t, \mathbf{A}_t) = \frac{1}{2}(\tilde{\mathrm{R}}_t - \mathbb{X}_{t-1}\mathbf{A}_t)^T(\tilde{\mathrm{R}}_t - \mathbb{X}_{t-1}\mathbf{A}_t) + \Phi(\lambda_t, \mathbf{A}_t). \tag{4.4}$$

We now integrate the loss function over all uncertainty given by the predictive distribution of asset returns conditioned on observed data: $p(\tilde{\mathrm{R}}_t \mid \mathbf{R}_t) = \int p(\tilde{\mathrm{R}}_t \mid \Theta_t, \mathbf{R}_t)p(\Theta_t \mid \mathbf{R}_t)\,d\Theta_t$. We do this integration in two steps, first over $\tilde{\mathrm{R}}_t \mid \Theta_t$ and second over Θ_t,

$$\begin{aligned}\mathcal{L}_{\lambda_t}(\mathbf{A}_t) &= \mathbb{E}_{\Theta_t}\mathbb{E}_{\tilde{\mathrm{R}}_t|\Theta_t}\left[\frac{1}{2}(\tilde{\mathrm{R}}_t - \mathbb{X}_{t-1}\mathbf{A}_t)^T(\tilde{\mathrm{R}}_t - \mathbb{X}_{t-1}\mathbf{A}_t) + \Phi(\lambda_t, \mathbf{A}_t)\right] \\ &\propto 2\overline{\mathbf{B}}_t^T\mathbb{X}_{t-1}^T\mathbb{X}_{t-1}\mathbf{A}_t + \mathbf{A}_t^T\mathbb{X}_{t-1}^T\mathbb{X}_{t-1}\mathbf{A}_t + \Phi(\lambda_t, \mathbf{A}_t) + \text{constants}.\end{aligned} \tag{4.5}$$

After integration we notice that the posterior mean of the coefficients, $\overline{\mathbf{B}}_t$, appears in the first term, while the expectations pass over the second and third terms.

We complete the square and drop constants to obtain the final form of the integrated loss function,

$$\mathcal{L}_{\lambda_t}(\mathbf{A}_t) = \|\mathbb{X}_{t-1}\mathbf{A}_t - \mathbb{X}_{t-1}\overline{\mathbf{B}}_t\|_2^2 + \Phi(\lambda_t, \mathbf{A}_t). \tag{4.6}$$

For a fixed time t, Loss (4.6) has the same form as the one derived for linear regression models in Hahn and Carvalho (2015). The third and final step is to choose a penalty function Φ and optimize the loss function for a range of λ_t for each time t.

For this paper we choose $\Phi(\lambda_t, \mathbf{A}_t) = \lambda_t \sum_{k=1}^{K}\|\mathbf{A}_t^k\|_2$, where $\mathbf{A}_t^k$ is the kth $\grave{m} + \acute{m} + 3$ block of the vector $\mathbf{A}_t$ after neglecting the intercept. The group lasso algorithm of Yuan and Lin (2006) is then used to minimize the integrated loss. This provides a way to jointly penalize groups of covariates. In the context of our financial application, this "group penalization" permits the selection of firm characteristics by grouping the coefficients of a single quadratic spline together in the penalty.

In order to see this, recall the structure of the sparse action $\mathbf{A}_t$. It is a $K(\grave{m} + \acute{m} + 3) + 1$ length vector where the kth $\grave{m} + \acute{m} + 3$ block (excluding the intercept) corresponds to the spline basis for firm characteristic k. By using the approach outlined in Yuan and Lin (2006), we group together the spline bases for each characteristic. Then, Loss (4.6) is minimized for varying penalty parameter choices, such that we can look at a range of quadratic spline models built from one characteristic up to the 36 characteristics available.

Posterior summary plots. These sparse models are optimal under our choice of loss and fixed level of regularization given by the penalty parameter, and we can compare them in light of the statistical uncertainty from the Bayesian model. Denoting the collection of sparse optimal models $\{\mathbf{A}_{\lambda_t}^*\}$, we study the distribution of the *difference in loss* of a reduced model and the full model,

$$\Delta_{\lambda_t} = \mathcal{L}(\tilde{\mathrm{R}}_t, \mathbf{A}_{\lambda_t}^*) - \mathcal{L}_0(\tilde{\mathrm{R}}_t, \mathbf{A}_0^*), \tag{4.7}$$

where $\mathcal{L}$ is as defined in equation (4.3). Note that, as $\mathcal{L}$ is a random variable, so is Δ. Crucially, this metric incorporates statistical uncertainty through the predictive and optimality through consideration of the set $\{\mathbf{A}^*_{\lambda_l}\}$.

An important feature of this approach is the ability to identify important return predictors and how this set may vary *over time*. The time variation and connection across time periods is driven by the power-weighted density approach and embedded in the posterior (recall that the rolling-window model is a special case of the power-weighted density approach). Therefore, although the minimization of the integrated loss is performed myopically at each point in time, the variation of optimal sparse models across time may be studied.

5. Case study. Our case study focuses on a rich data set from an earlier version of Freyberger, Neuhierl and Weber (2020). It is a monthly panel where we observe a cross-section of firms, their excess return as well as 36 lagged characteristics of each. The full dataset spans 623 months of returns, July 1962 through May 2014, and includes 1,404,048 observations. We train our models on the first 12 years and then test and update the models on the remaining data. Thus, the results shown cover 1974–2014. The posterior distributions and training models are updated annually.

The characteristics are listed in Table 5 as well as the direction of monotonicity we impose for each. We will examine three different sets of splines: "Splines-0" with no constraints (non-monotonic), "Splines-6" with some constraints and "Splines-24" with many constraints. The "Splines-6" model applies the rather established evidence from the financial literature (Fama and French (2016)) and constrains size, book-to-market, profitability, investment, momentum (Jegadeesh and Titman (1993), Jegadeesh and Titman (2001)) and intermediate momentum to be monotonic. For the "Splines-24" model we impose monotonicity constraints on every variable whose constraint had reasonable support in the literature; thus, 24 of the 36 variables are constrained. The supporting papers are also listed in Table 5.

We examine three different benchmarks: ordinary least squares (OLS), random forests (Breiman (2001)) and Bayesian additive regression trees or "BART" (Chipman, George and McCulloch (2010)). The latter two are nonparametric ensemble learning models with competitive predictive ability across many applications. Each of these models are given all 36 characteristics, and fit with different training window sizes: 36 months, 60 months, 120 months and the historic window. Tree-based models are interesting in this case, as they are effectively portfolio sorts (i.e., step functions) with breaks between portfolios (i.e., partitions) chosen by the data.

Our random forest model uses 300 trees, six layers deep, and tries five characteristics at each split node. This specification allows for deep, complex and diverse trees. These tuning parameter values were chosen similar to the models in Gu, Kelly and Xiu (2020).[1] Like Gu, Kelly and Xiu (2020), we use 300 trees per forest. Each forest then has two tuning parameters. First, each tree in the forest has a certain depth (anywhere from one to six layers) and, the deeper the tree, the more complex the model in terms of large differences in leaf values, nonlinearities and interactions. Second, each forest can have varying numbers of variables explored at each split in the tree (either three, five, 10, 20 or 30 characteristics per split), and this controls how diverse the set of covariates is in each tree. This provides 30 different combinations of tuning parameters, and we fit all 30 with a 120-month rolling window. The tuning parameter values that yielded the highest ex post out-of-sample R^2 were chosen. The `ranger` package (Wright and Ziegler (2017)) was used to fit the models in R.

[1] Gu, Kelly and Xiu (2020) reselect their tuning parameters in a rolling manner, whereas here we make a single choice ex post.

Our BART model was fit using the newly developed `XBART` package in R (He, Yalov and Hahn (2019)) with the tuning parameters recommended in their paper. Unlike random forests, the default number of trees is a function of the data. Also, BART specifies a prior on tree depth which implies the trees in an ensemble can be of varying depth depending on what the data dictate.[2]

Additionally, we fit nine other specifications of our additive quadratic splines model across the three different sets of monotonic constraints mentioned above and across three different specifications of discounting. Using the McCarthy and Jensen (2016) approach, we implement power-weighted discounting with $\delta = 0.980$ which corresponds to an effective sample size of 50 months, $\delta = 0.990$ for roughly 100 months and $\delta = 0.998$ for about 500 months, thus providing only slight discounting.

In the analysis to follow, we present two sets of results. First, we detail the modeling results, including out-of-sample performance and partial effect function plots. Second, we present characteristic selection results (conditional upon the models).

5.1. *Modeling results.* ***The impact of model specification.*** In this section we display the forecasting ability of the model specifications considered. Since our selection process requires a posterior distribution as an input, this analysis helps in identifying the model (equivalently, the monotonicity and time-variation specifications) among our 21 different spline specifications that possesses the best predictive ability.

We first look at the out-of-sample (OOS) R^2, given in a percentage scale,

$$R^2_{\text{OOS}} = 100\left(1 - \frac{\sum_{i,t}(r_{it} - \hat{r}_{it})^2}{\sum_{i,t} r_{it}^2}\right), \tag{5.1}$$

where $\hat{r}_{it}$ are the model forecasts.[3] The results for the entire testing period (1974–2014) are given in Panel A of Table 1 and provides two main takeaways. First, longer window sizes are best for forecast accuracy, with the historic window model performing best. Second, our spline models perform somewhere in between that of random forests and BART. However, additive splines do not include interactions between characteristics and, therefore, are "interpretable," meaning we can easily display and understand the partial effect of each characteristic. Additionally, having a small number of monotonic constraints (Splines-6) performs uniformly better than no constraints (Splines-0), though this effect size may be small. A slight improvement from no to some constraints is to be expected, as observed in Figures 1 and 2. In summary, the important discoveries are that, (*i*) spline models can have similar forecast performance to tree/forest models without being a black-box and (*ii*) monotonic constraints can yield some improvement.

Table 2 shows the OOS results of Table 1 with the test period divided into thirds.[4] In terms of R^2_{OOS}, the longer window length models perform better out-of-sample. Also, BART performs well regardless of the time period. We do see that random forests surpass the performance of the splines in the latest subperiod. Again, if the sole purpose of modeling were to improve forecasting, then a black-box machine learning method is best. However, if one is interested in an interpretable model (which we define as being able to visualize the partial

[2]XBART does not use the standard BART posterior sampling techniques but rather a "grow-from-root" backfitting procedure that is efficient and faster than a standard MCMC sampler. The suggested number of trees for the XBART sampler is $0.25 * \log(n)^{\log(\log(n))}$, where n is the number of observations the model uses for training.

[3]As in Gu, Kelly and Xiu (2020), the total sum of squares in the denominator assumes returns are mean zero.

[4]Table 6 in the Appendix breaks the testing period into only two subperiods and shows, essentially, the same patterns changing over time with less granularity.

TABLE 1

Out-of-sample model performance when forecasting monthly returns over the 1974–2014 *period. Out-of-sample* R^2 *is calculated as done in Gu, Kelly and Xiu* (2020). *Sharpe ratios use the returns from equal-weighted decile portfolios, buying stock in the firms whose returns are forecasted to be in the top decile and shorting the bottom decile. Bolded values indicate the best two models for a given window length*

	Panel A: OOS R^2				Panel B: Sharpe ratio			
	Window size (months)				Window size (months)			
	All	120	60	36	All	120	60	36
OLS	0.57	0.43	−0.07	−0.57	2.45	2.26	1.71	1.48
Random Forest	0.74	0.62	0.14	−0.43	3.11	2.61	1.89	1.45
BART	**1.22**	**0.98**	0.23	−0.58	**3.44**	**3.29**	**2.83**	2.35
Splines-0	0.87	0.68	0.24	−0.16	3.02	2.98	2.60	2.30
Splines-6	**0.87**	**0.68**	**0.25**	**−0.13**	3.05	2.99	2.64	**2.35**
Splines-24	0.81	0.67	**0.27**	**−0.09**	**3.14**	**3.22**	**2.71**	**2.41**

effect of individual characteristics), then splines are a reasonable approach with performance in the same ballpark as black-box methods.

Tables 1 and 2 also show the economic impact of the models in terms of Sharpe ratio. The challenge with looking too closely at point forecasts (via R^2) is that slight differences may not matter much: the signal to noise ratio is small. Therefore, an alternative model comparison approach is to compute portfolio metrics. Like Freyberger, Neuhierl and Weber (2020), these Sharpe ratios are calculated using the returns of an "equal-weighted hedge portfolio going long the 10% of stocks with highest predicted returns and shorting the 10% of stocks with lowest predicted returns." These returns are forecasted monthly, and the portfolio is rebalanced monthly. We present the annualized Sharpe ratio from these monthly returns in the right panel of Table 1. We again see that BART's forecasting accuracy includes good estimation of which firms will be in the tails (which leads to large Sharpe ratios). However, we also see that including more monotonic constraints in Splines-24 leads to large Sharpe ratios, on par with and sometimes exceeding BART's performance. In Table 2 we also see that the Sharpe ratios decline over all specifications in the last third of the data, and this is likely a result of the discovery of many characteristics/factors in recent decades.[5]

Why would the random forest, a more flexible model, not outperform splines? The bias-variance tradeoff and low signal-to-noise environment in finance data are key concepts affecting these results. The highly flexible nonlinear model given by the random forest is overwhelmed with noise, and its resulting performance is poorer than it might have been otherwise. This underscores a need for structured models in these applications beyond their ease of interpretability. BART has bias toward smaller trees which protects against overfitting. Similarly, bias induced by our structured monotonic spline models leads to the splines typically performing similar to, or outperforming, random forests out-of-sample, though both structure and interpretability are minimal in random forests. In the following sections we explore the return-characteristic relationships that our fitted models provide.

[5]The Sharpe ratios we report are rather high as we examine characteristics that were ex post discovered to be useful predictors of returns. Thus, an investor in the past could not have done this. This is illustrated empirically as we do see a drop in Sharpe ratios between the 1988–2000 period and the 2001–2014 period. For our purposes, we are interested in the relative magnitude of one feature set's model's Sharpe ratio compared to another. Hence, these numbers should only be interpreted for this out-of-sample comparison (or perhaps contemporary out-of-sample exercises with past data).

TABLE 2

Out-of-sample model performance when forecasting monthly returns over over the January 1974 *to December* 1987, *January* 1988 *to December* 2000 *and January* 2001 *to May* 2014 *periods. Out-of-sample* R^2 *is calculated as done in Gu, Kelly and Xiu* (2020) *and given in percentage points. Sharpe ratios use the returns from equal-weighted decile portfolios, long on the firms with forecasted in the top decile and shorting the bottom decile. Bolded values indicate the best two models for a given window length*

	OOS R^2				Sharpe ratio			
	Window size (months)				Window size (months)			
	All	120	60	36	All	120	60	36
				1974–1987				
OLS	0.70	0.45	0.34	0.01	3.17	3.01	2.79	2.64
Random Forest	0.64	0.40	0.23	−0.12	**3.91**	3.74	3.03	2.68
BART	**1.32**	**1.01**	**0.74**	**0.47**	**3.92**	**3.83**	**3.74**	**3.53**
Splines-0	0.89	0.63	0.53	0.33	3.84	3.74	3.57	**3.45**
Splines-6	**0.90**	**0.64**	**0.54**	**0.35**	3.77	3.79	**3.64**	3.44
Splines-24	0.84	0.61	0.53	0.35	3.84	**3.79**	3.59	3.38
				1988–2000				
OLS	0.51	0.37	0.21	−0.18	2.87	2.47	2.22	1.82
Random Forest	0.69	0.59	0.40	0.06	3.97	3.08	2.48	1.74
BART	**1.34**	**1.18**	**0.89**	**0.54**	**4.45**	**4.25**	**3.99**	**3.00**
Splines-0	0.92	0.78	0.67	0.32	3.82	3.80	3.61	2.95
Splines-6	**0.93**	**0.79**	**0.68**	**0.34**	3.90	3.78	3.55	2.97
Splines-24	0.86	0.73	0.64	0.33	**4.10**	**4.28**	**3.82**	**3.05**
				2001–2014				
OLS	0.62	0.54	−0.59	−1.29	1.66	1.61	0.78	0.61
Random Forest	**0.88**	**0.79**	**−0.20**	−1.18	**2.06**	1.67	0.95	0.56
BART	**1.05**	**0.74**	−0.80	−2.43	**2.37**	**2.20**	**1.50**	1.14
Splines-0	0.82	0.60	−0.39	−0.96	1.98	1.97	1.41	1.22
Splines-6	0.82	0.60	−0.39	**−0.91**	2.02	1.99	1.46	**1.32**
Splines-24	0.76	0.65	**−0.29**	**−0.77**	2.02	**2.14**	**1.49**	**1.32**

Table 3 shows the performance of spline models under different time-variation methods. Recall that the effective sample sizes of $\delta = 0.998, 0.990, 0.980$ are 500, 100, 50, respectively. Hence, $\delta = 0.990$ is most directly comparable to the 120-month rolling window, and $\delta = 0.980$ is in between the effective sample sizes of the 60- and 36-month rolling windows. Using $\delta = 0.998$ provides a slightly discounted version of the historic window, and we see that these two approaches perform similarly over every period and over both metrics, with the historic window having slightly better R^2_{OOS} values. For medium effective sample sizes, power-weighting with $\delta = 0.990$ outperforms the 120-month rolling window on both metrics in the early period but that gap closes in the more recent periods. In fact, during the 1988–2000 and 2001–2014 subperiods, Splines-24 fit with the 120-month rolling window produces rather high Sharpe ratios. With the smaller effective samples sizes, we see that power-weighting with $\delta = 0.980$ has better R^2_{OOS} than the 60-month rolling window on the first and last subperiod but comparable in the second subperiod. In terms of Sharpe ratio, power-weighting is slightly better in first subperiod, rolling window is better in the second subperiod and power-weighting is better in the final subperiod. By comparison, 36-month rolling window performs poorly everywhere. As a method for dealing with nonstationarity, it does not use enough data and is thus too flexible.

In summary, there are two conclusions for time variation in this dataset. First, larger (effective) sample sizes are preferred in terms of out-of-sample performance. The main exception is that 120-month rolling windows can produce better Sharpe ratios than the historic window

TABLE 3

Out-of-sample model performance when forecasting monthly returns over over the January 1974 *to December* 1987, *January* 1988 *to December* 2000 *and January* 2001 *to May* 2014 *periods. Out-of-sample* R^2 *is calculated as done in Gu, Kelly and Xiu* (2020) *and given in percentage points. Sharpe ratios use the returns from equal-weighted decile portfolios, long on the firms with forecasted in the top decile and shorting the bottom decile*

	OOS R^2							Sharpe ratio						
		Window size (months)			Discounting weight δ				Window size (months)			Discounting weight δ		
	Historic	120	60	36	0.998	0.990	0.980	Historic	120	60	36	0.998	0.990	0.980
							1974–1987							
Splines-0	0.89	0.63	0.53	0.33	0.87	0.80	0.69	3.84	3.74	3.57	3.45	3.73	3.83	3.75
Splines-6	0.90	0.64	0.54	0.35	0.87	0.79	0.68	3.77	3.79	3.64	3.44	4.04	3.96	3.66
Splines-24	0.84	0.61	0.53	0.35	0.83	0.76	0.65	3.84	3.79	3.59	3.38	3.91	3.87	3.68
							1988–2000							
Splines-0	0.92	0.78	0.67	0.32	0.91	0.82	0.67	3.82	3.80	3.61	2.95	3.89	3.70	3.38
Splines-6	0.93	0.79	0.68	0.34	0.92	0.83	0.67	3.90	3.78	3.55	2.97	3.93	3.67	3.44
Splines-24	0.86	0.73	0.64	0.33	0.83	0.75	0.62	4.10	4.28	3.82	3.05	4.07	3.75	3.59
							2001–2014							
Splines-0	0.82	0.60	−0.39	−0.96	0.78	0.62	0.31	1.98	1.97	1.41	1.22	1.97	1.98	1.72
Splines-6	0.82	0.60	−0.39	−0.91	0.78	0.61	0.31	2.02	1.99	1.46	1.32	2.02	1.98	1.79
Splines-24	0.76	0.65	−0.29	−0.77	0.72	0.54	0.28	2.02	2.14	1.49	1.32	2.00	1.86	1.92

for the Splines-24 model. Second, we see that power-weighting methods perform similarly to traditional window methods in most cases, particularly in recent years. However, if a high degree of time-varying flexibility is desired (i.e., a small effective sample size), then power-weighting is likely the preferred approach.

The return-characteristic relationship. Figure 4 shows the partial effects of 16 characteristics from the full posterior of the historic window model. Specifically, we use Splines-6 with the subset of monotonic constraints, as it has the best fit in terms of forecast error; see Table 1. Each individual pane shows the partial effect of a characteristic, assuming the other 35 characteristics are held at their medians. The first thing to note are the strong effects of size, momentum, short-term reversal, standard unexplained volume and price to 52-week-high. We also see that some effects are clearly not monotonic: turnover, idiosyncratic volatility and price to 52-week-high. This provides important insight into why excessive monotonic constraints can hurt the model. Additionally, there are nearly flat partial effects in the full posterior, such as book-to-market which is a staple in empirical finance work. We look into this phenomenon in Section 5.2.

The return-characteristic relationship, over time. We next look at partial effects given at different points in time for the different models in Figure 5. This figure shows the partial effect of firm size on returns when holding all other variables at the median, estimated by multiple linear regression (OLS), a random forest, BART and our different spline models. The effects and their uncertainty are given for January of 1974, 1994 and 2014, each using a 120-month window of training data (rolling-window models). Generally, we see the size effect growing stronger over time: the smallest firms see much larger average returns than all other firms. This effect is blurred by standard regression's assumption of a strictly linear relationship. Random forests pick up this small-firm phenomenon, but the resulting curve is extreme and does not change much over time compared to BART and the spline models. The spline and BART partial effects look similar. The splines, with shrinkage on each knots' coefficient, smoothes over the noise picked up in BART's estimates. Without this degree of shrinkage, the spline estimates of characteristics' partial effects can contain undesired waves.

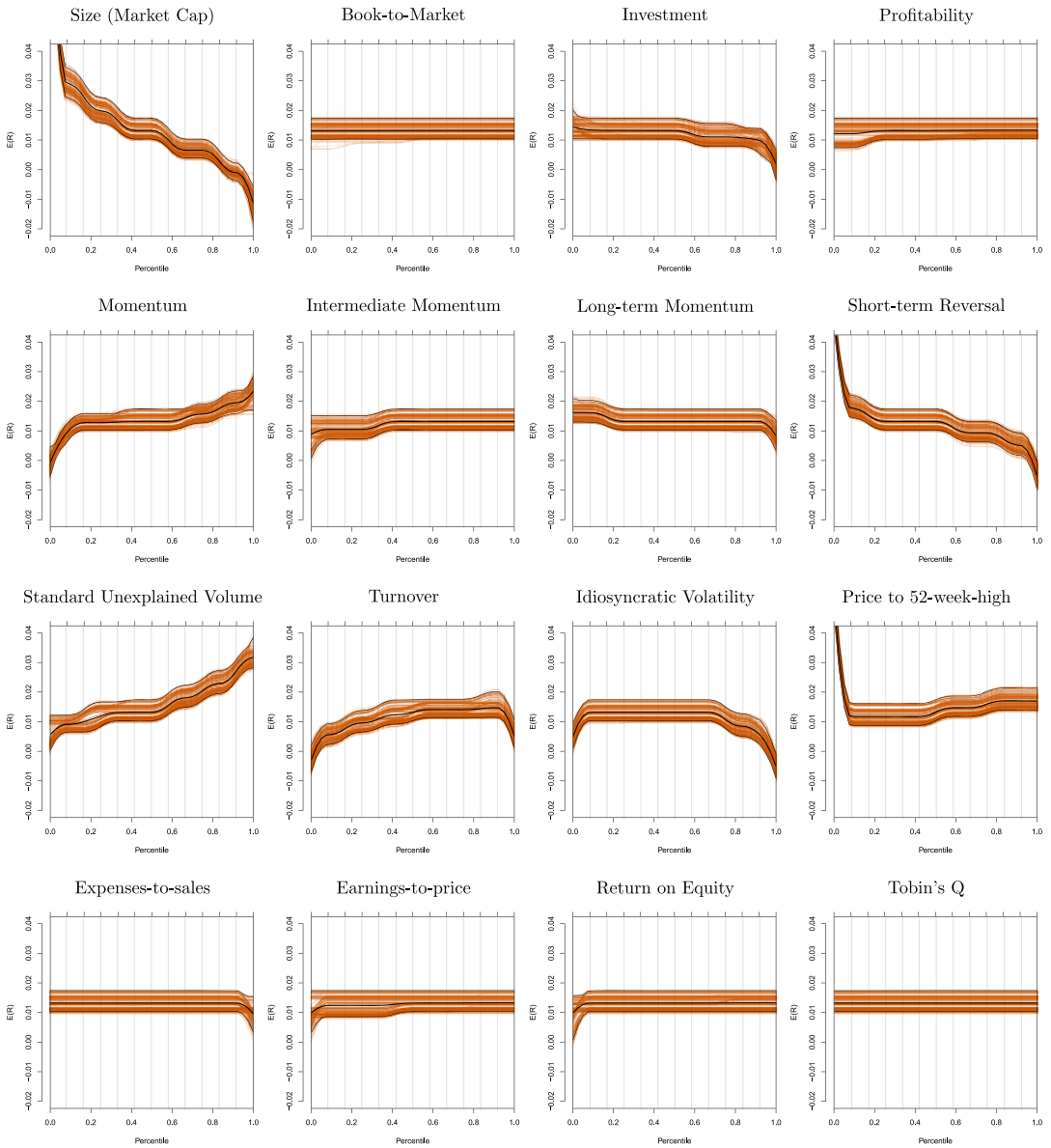

FIG. 4. *Effects of characteristics on returns over the historic window of* 1974–2014 *period* (*each observation equally-weighted over time*) *using the Splines-*6 *model. Here, only six variables are constrained to be monotonic*: *size*, *book-to-market*, *investment and profitability on the first row* (*Fama and French* (2016)) *as well as momentum and intermediate momentum on the second row. The remainder of the second row is composed of other functions of past returns. The third is composed of various pronounced effects, while the fourth row contains characteristics with much smaller or no effects in the full posterior. The three black curves are the posterior mean and the* 95% *credible bands. The transparent orange curves are each of the MCMC draws, such that darker orange areas reflect greater posterior density. The vertical gray lines show where the knots are placed. The horizontal axes are the percentiles of the characteristic. The vertical axes are the expected returns.*

For example, the partial effect estimates in Freyberger, Neuhierl and Weber (2020) would be smoother (less wavy within noisy data) if additional shrinkage were imposed. Smoothing over noise is also aided by imposing monotonic constraints where appropriate. Furthermore, this figure visually demonstrates two sources of uncertainty and variance reduction, more data and more structure (the bias part of the bias/variance tradeoff). There are more firms in 2014 than in 1974, and thus the 95% credible or confidence bands decrease as we move from

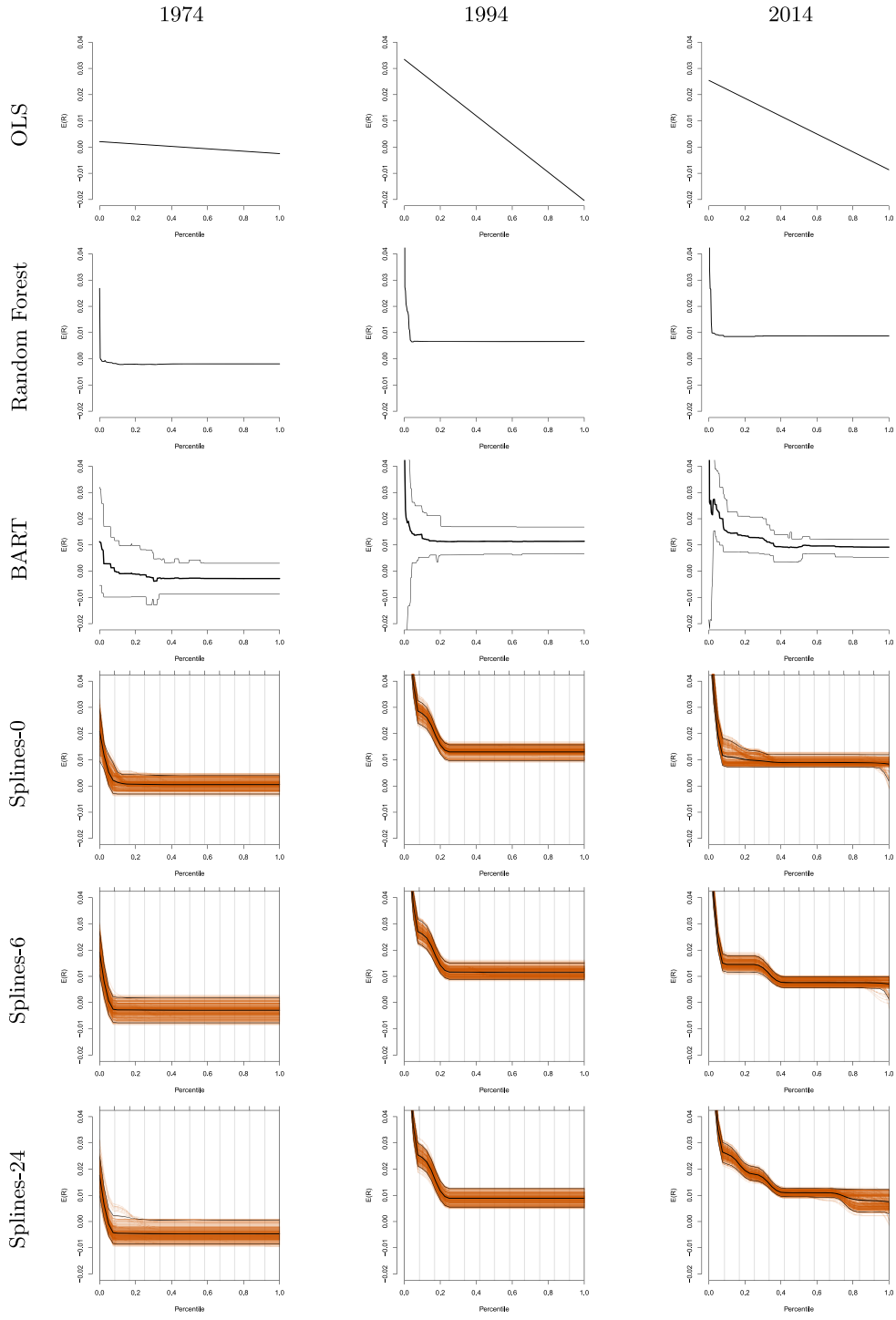

FIG. 5. *Comparing functions of size (market cap) over time. Assumes all other variables are held at their medians* (0.5). *Each uses a* 120*-month rolling window. The horizontal axes are the percentiles of size. The vertical axes are the expected returns. When three black curves are present, they are the posterior mean and the* 95% *credible bands. The transparent orange curves are each of the MCMC draws, such that darker orange areas reflect greater posterior density. The vertical gray lines show where the knots are placed.*

the left panes to the right panes. The further reduction in variance we see as we progressively move from the middle right panes to the bottom right panes comes from adding monotonic constraints. This is especially seen in the tighter probability bands in the 2014 panes: as more monotonicity constraints are added to the model in general, the estimate of the size effect becomes more certain, even though the splines for size never appear nonmonotonic.

Figure 6 shows the annual progression of the splines. There are two different color schemes to denote two different sets of axes. The first panel illustrates a common pattern—little change over time. The effect of standard unexplained volume is fairly consistent over time with slight fluctuations. Next, while frequently included in our models and most other papers' asset pricing models, book-to-market does not have a very strong effect, though this could change with different control variables. We see some value premium (high expected returns of high book-to-market firms) in the late 1970s and low returns of growth firms (low percentiles of book-to-market) in the 2000s, which, as this is a 120-month rolling window, likely reflects the burst in the dot-com bubble. The positive returns seen by the smallest firms (size) increase halfway through the period, while there is little effect on the large firms until the Great Recession, as seen in the red dip near the end of last decade. While the effect of short-term reversal (firms performing well last month tend to underperform this month and vice versa) on the low percentile/worst firms' positive returns is fairly stable over time, the negative effect on returns of last month's winners depletes over time.

5.2. *Selection results.* ***Which characteristics are important?*** We first look at which characteristics are important over the whole time period, 1974–2014. To do this, we use the historic window model where each observation over the 41-year period receives equal weight. Also, Splines-6 (the small set of monotonic constraints) is used, as it has the best fit in terms of forecast error; see Table 1. Then, posterior summarization is performed as detailed in Section 4, except that we look at the whole period as a single time step; thus, $\mathbb{X}_{t-1}$ contains data from over the whole time period. We do this by looking at the distribution of the difference in loss of a reduced model, and the full model defined in expression (4.7).

A convenient feature of the selection methodology is the ability to undertake a full sample analysis such as this. Beyond the loss function, the remaining components for the method are the predictive distribution calculated at the end of our sample and a subsample of our data to build $\mathbb{X}_{t-1}$.[6] Then, the expected loss is computed, optimized and the optimal sparse models are compared in light of predictive uncertainty. This *separation* of inference from characteristic selection is a helpful tool for exploratory analysis within our case study.

In Figure 7 we show a difference in loss metric Δ_{λ_t} in the left panel and the probability that a sparse model has less loss than the full model, $P(\Delta_{\lambda_t} < 0)$, in the right panel for a sequence of models with varying numbers of included characteristics. One can think of λ_t as indexing models of varying sizes. The models in this sequence are minimum loss models for each number of included characteristics.[7] Figure 7 shows that using 27 or more characteristics has a very high probability of having the same or smaller loss than the full model. The left panel of Figure 7 also shows that the timely inclusion of expenses-to-sales, investment, return on equity and Tobin Q's all lead to significant movements (toward zero) in the distribution of Δ_{λ_t}.

The red line in the right panel shows our threshold of 0.05, meaning models above the threshold have at least a 5% chance of having equal or better loss than the full/dense model.

[6]We summarize the posterior with respect on the a random month from each of the 41 years in the test set, as using all firm-year observations in $\mathbb{X}_{t-1}$ from equation (4.2) to summarize the posterior currently does not work on a 16GB RAM machine.

[7]Equation (4.6) is optimized for hundreds of values of λ_t. Of all the models with p covariates selected into the model, the chosen model has the minimum loss among all models with p covariates.

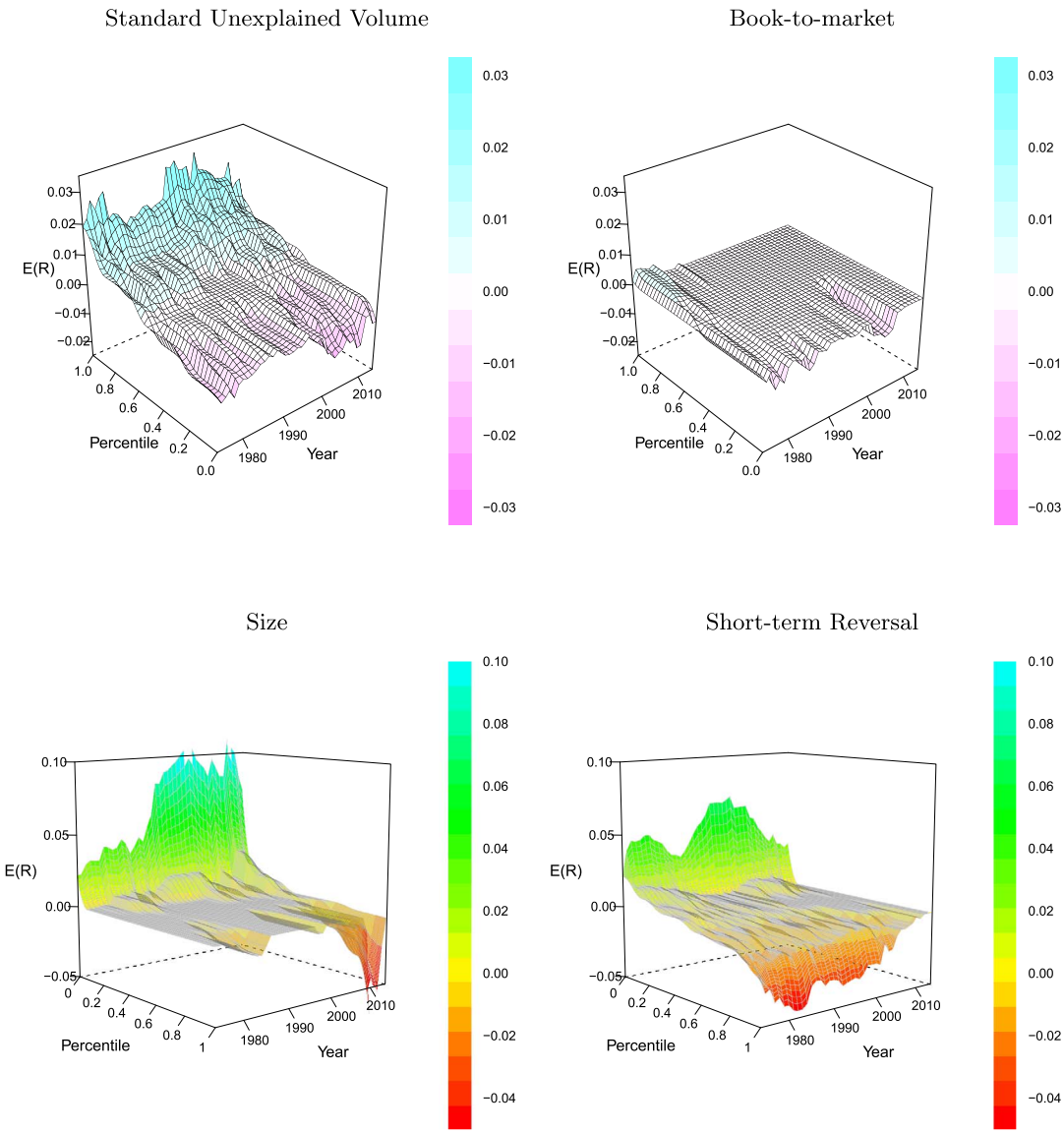

FIG. 6. *Splines of most included effects, throughout the test period*, 1974–2014. *For the given characteristic, the splines for January in each year are placed in order, yielding a response surface of the rank-transformed characteristic and time vs. monthly expected returns. The blue/pink color scheme has an increasing percentile axis. Colors are assigned to buckets of expected returns* 50 *basis points wide, such that regions of expected returns between* −25 *and* +25 *basis points are white. These basis point changes are with respect to a firm with the median value of the characteristic in a given year. Hence, the white areas of the plot reflect percentiles of firms that do not vary significantly from the median firm. The red/green/cyan color scheme flips the percentile axis, so the curve is viewable and zooms out along the E(R) axis essentially doubling the limits and halving the granularity of the color spectrum. The spline estimates come from the Splines-*24 *model fit on* 120*-month rolling windows, such that book-to-market is constrained to be nondecreasing and size is nonincreasing.*

We select the sparsest model over the threshold which has 25 characteristics. These are given in Table 4 in order of inclusion: Table 4 shows us that, while the variables from Fama and French (2016) are present (i.e., investment, book-to-market, size and profitability), they do not come first; standard unexplained volume and short-term reversal are the first characteristics to enter the sparsest models. Both of these have large effects over the sample, as we saw in Figure 4.

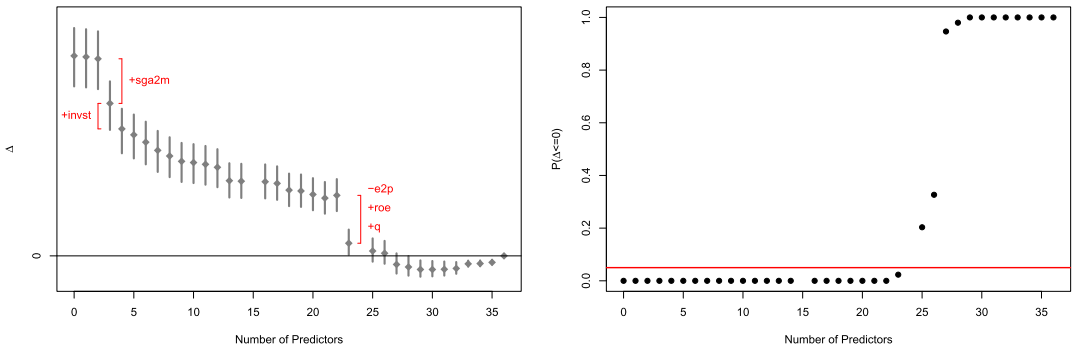

FIG. 7. *Posterior summary plots over the full test period. The left panel shows the distribution of the difference in loss for models of differing numbers of characteristics relative to the full model. In red are shown the variables that, when added* (+) *or removed* (−) *from the model, cause significant changes in the loss distribution. The right panel shows the probabilities of having the same or better loss than the fully-dense model of all* 36 *predictors. The red threshold in the right panel is* 0.05*, and the model immediately above the threshold is selected.*

Freyberger, Neuhierl and Weber (2020) find that, out of 62 characteristics, around 13–16 characteristics are selected into their main model over the full period from 1965-2014. Like us, their selected set includes: investment, size, turnover, short-term reversal, momentum, intermediate momentum, standard unexplained volume and price to 52-week high. Additionally, their top 13 characteristics include some not in our dataset: percent change in shares outstanding, log change in the split adjusted shares outstanding, ROC,[8] total volatility and adjusted profit margin.[9] The next three that are sometimes included are book-to-market, net

TABLE 4

Selected variables variable for a threshold of 0.05 *on* $P(\Delta_{\lambda_t} < 0)$*, ordered by order of inclusion in the model. These are variables included from the model that does no worse than the full model with* 5% *probability. (tie) denotes variables that come in to the model at the same time. Note that in Figure* 7 *there is no optimal model with* 15 *or* 24 *predictors, thus why we see ties at* 15*th and* 24*th above. The other ties are instances of two variables coming into the model as another leaves. However, variables that leave the model and do not come back in for the ideal set of* 25 *are not included in the overall ranking*

1. Standard unexplained volume	14. Free cashflow
2. Short-term reversal	15. Cash-to-assets (tie)
3. Expenses-to-sales	15. Price-to-cost margin (tie)
4. Investment	17. Price to 52-week-high
5. Book-to-market	18. Return on assets
6. Momentum	19. Idiosyncratic volatility
7. Intermediate momentum	20. Profit margin (tie)
8. Size (Market cap)	20. Sales-to-price (tie)
9. Depreciation-to-assets	22. Tobin's Q (tie)
10. Long-term momentum	22. Return on equity (tie)
11. Net-operating assets	24. Change in PP&E and inventory (tie)
12. Turnover	24. Profitability (tie)
13. Leverage	

[8]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).

[9]Our data only contains profit margin, while Freyberger, Neuhierl and Weber (2020) include both profit margin and "adjusted" profit margin, the latter of which is their chosen variable.

operating assets and long-term momentum. The methods of Gu, Kelly and Xiu (2020) do not explicitly select a subset of characteristics, but they do look at how often each characteristic is used by each of their machine learning methods. They find that, out of 94 characteristics (and 74 industry indicators), the 10 characteristics that are most important in the 1987–2016 out-of-sample period are, in order: short-term reversal, size, 12-month momentum, change in six-month momentum, maximum daily return, industry momentum, return volatility, dollar trading volume, sales-to-price and turnover. Another related paper is Han et al. (2019), and they use a combination lasso approach on 99 characteristics over the time period 1975–2018. They find that 20–30 characteristics are used in their linear model at any point in time. Similarly, their top 10 most often included characteristics are, in order: short-term reversal, tobacco/alcohol/gaming industry indicator, tax income-to-book income ratio, organizational capital, corporate investment, consecutive quarters with an increase in earnings over the same quarter in the prior year, sales-to-receivables ratio, convertible debt indicator, price delay and industry-adjusted cash flow-to-price ratio. Across each of these studies, with similar but different time periods and characteristic sets, we see that short-term reversal, size, investment, momentum and a measure of volatility are important characteristics. Of note is that book-to-market, for all its acclaim in the finance literature, is not often a top-ranking characteristic.

The selection procedure also allows us to answer practical questions about individual characteristics. For example, the book-to-market characteristic has a relatively flat partial effect in the full posterior; see Figure 6. The posterior summary selection procedure can be used here to identify which groups of characteristics are interacting with book-to-market to produce this effect. This is done by optimizing the expected loss in expression (4.6) while constraining the coefficients of the book-to-market basis to be in all reduced models. Optimizing this loss over a range of penalty parameters produces a set of sparse models, all of which include the book-to-market characteristic.

Figure 8 displays the book-to-market partial effect from three of the reduced models along the solution path. "Book-to-market" is the sparsest model; "Book-to-market + some" is the model containing the book-to-market coefficients and 27 other characteristics, including the variants of momentum and volatility, size, profitability, investment and leverage. "Book-to-market + all" includes the remaining characteristics and flattens the partial effect even further. The characteristics entering this stage are: total assets to size, asset turnover, sales to lagged total assets, earnings to price, fixed costs to sales, costs of ratio of goods sold plus SG&A to total assets, return on net operating assets and CAPM beta. These three curves in Figure 8 show us that the first 27 other characteristics are responsible for about half of the dampening of the magnitude of book-to-market's effect, and the remaining eight characteristics are responsible for the other half. Hence, these last eight characteristics work the most to flatten the book-to-market effect in the full posterior. It should also be noted that the magnitude of the vertical axis in Figure 8 is fairly small, suggesting that, while book-to-market clearly has an increasing relationship with returns, the magnitude of that relationship is small.

When are characteristics important? To answer our second question, we implement the same posterior summarization as mentioned previously, using the same 0.05 threshold, but now in five year increments. Employing the posterior from the lightly time-discounted Splines-6 model ($\delta = 0.998$) of the listed January as well as its most current 60 months of data, we plot the selected covariates in Figure 9. Black cells indicate selected characteristics and light grey cells represent excluded characteristics. The most visibly striking thing herein is the lack of predictors selected in 1985. Apparently few characteristics were predictive of returns in the early 1980s. However, this paper does not put us in a position to make a causal statement as to why these changes happen, but we will comment on what happens.

In regards to specific characteristics, we first note that standard unexplained volume (suv), short-term reversal (r_{2-1}), momentum (r_{12-2}) and book-to-market (beme) are the only variables selected every period. Eight more were selected in every period except the early

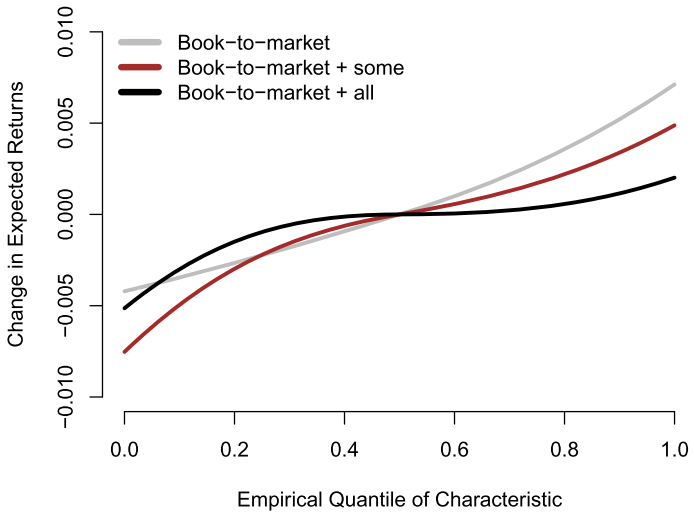

FIG. 8. *The partial effect plot of the book-to-market characteristic for three models along the posterior summary solution path. Each set of betas used to construct the three plots are taken from the expected loss optimization in expression* (4.6) *that is used to summarize the full model posterior. As more characteristics are included* (*gray to brown to black*), *the partial effect of book-to-market flattens*, *suggesting other characteristics* (*notated in the text*) *interact to dampen the book-to-market effect*.

1980s: price to 52-week-high (rel_to_high_price), idiosyncratic volatility (idio_vol), intermediate momentum (r_{12-7}), net operating assets (noa), turnover (lturnover), size (lme), investment and debt-to-assets (d2a). In reference to the literature (Jegadeesh and Titman (1993), Jegadeesh and Titman (2001); Fama and French (2016)) and our smaller set of monotonic constraints in Splines-6, there appears to be support for size, book-to-market, investment, momentum and intermediate momentum, but not for profitability (prof).

Figure 10 displays the time evolution of the partial effect functions and provides context for the time-varying selection results in Figure 9. Note that these dynamic estimates are a novel feature of our methodology and, along with monotonicity, represent two new contributions of the paper. Shown are the function evolutions for three characteristics, size (lme), profitability (prof) and leverage (lev). Note that the shape, location and statistical uncertainty of these estimates change over time. The location is driven primarily from the evolving intercept term in the CEF. The uncertainty in the early periods is large, reflecting a smaller number of cross-sectional observations compared to more recent time periods. Effects that stray away from constant will carry more influence in the Loss function (4.3) used for selection. Figure 10 show that the size characteristic possesses a strong partial effect for most time periods. In contrast, profitability remains flat throughout time and, as a consequence, is rarely selected in Figure 9. The final characteristic is leverage, and, while its effect is initially constant, it moves toward nonconstant in later periods. We see that it is selected in more recent periods from Figure 9 as a result.

6. Conclusion. The intersection of flexible modeling in Bayesian statistics and characteristic selection in finance is the focus area of this paper. We develop a statistical method for modeling returns based on the joint distribution of characteristics as well as provide a way to identify significant ones in light of statistical uncertainty. Our case study concludes that thoughtful model construction is important when dealing with finance data. Our conclusions suggest that model structure (through additivity and monotonicity) provides the dual benefit of interpretability and similar out-of-sample performance to highly-flexible, but minimally-interpretable, machine learning methods.

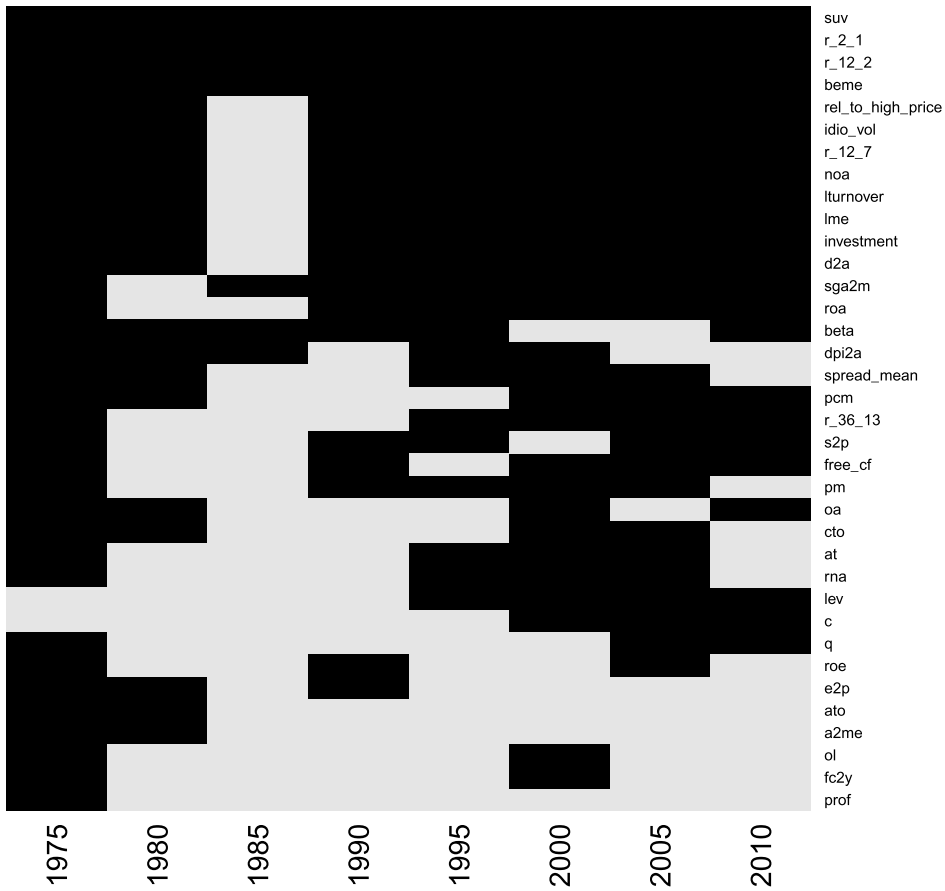

FIG. 9. *Using the same* 0.05 *threshold shown in Figure 7, we find the sparsest model with at least* 0.05 *probability of having no more loss than the fully dense model. This is done in five year increments, ending in January of the years listed on the horizontal axis. Variables on the vertical axis are ordered according to the frequency of their appearance. Black cells indicate selection, while light grey cells indicate exclusion from the sparse model.*

Specifically, there are three important contributions made by our model in this paper. First, our flexible and interpretable model is Bayesian and, thus, accounts for the different sources of uncertainty. Second, the model supplements the flexibility of quadratic splines, as shown in Freyberger, Neuhierl and Weber (2020), with theoretically-supported monotonic constraints being one of the least imposing forms of structure. Our results repeatedly show that the addition of monotonic constraints improves upon a model without such structure. Third, we modify the monotonic splines of Shively, Sager and Walker (2009) to be time dependent in order to model the nonlinear yet possibly-dynamic relationships of returns and characteristics. We carefully investigate time variation in our model using the methods of McCarthy and Jensen (2016) to discount past data. We find evidence for monotonicity even after conditioning on many other available characteristics. This conclusion is supported statistically and economically by an analysis across 21 model specifications.

The fourth contribution and the second half of this paper is the development of a posterior-summary selection procedure for our model. Using this new approach, we are able to examine the practical significance of characteristics and how these effects vary in time. We find about two dozen firm characteristics that have been important over the last four decades. However, we note that the timing of the importance of each characteristic varies and, two, that the magnitude of each characteristic's effect ranges from negligible to large, and this too can vary over time. With these methods we find that characteristics with the largest effects on

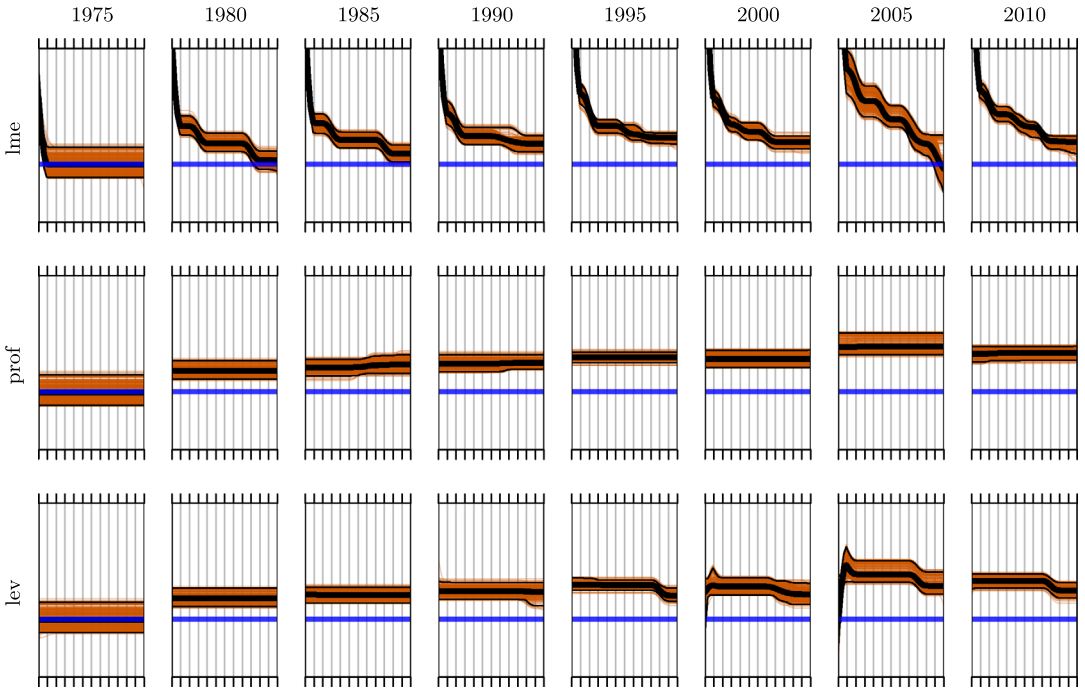

FIG. 10. *Comparing the time evolution of partial effect functions, every five years from* 1975 *to* 2010. *The model used is the same used for the characteristic selection in Figure* 9*—Splines-*6 *with discount factor* $\delta = 0.998$. *The characteristics shown are size* (*lme*), *profitability* (*prof*) *and leverage* (*lev*). *For readability, the axes are suppressed but are on the same scale as Figure* 4, *and the blue line is* $r = 0$.

expected returns are size, short-term reversal and standard unexplained volume. We find that, while the specifics of these effects change over time, their importance does not diminish. We also find that book-to-market, investment and momentum are also important over all time, although their effect sizes in the full posterior are not nearly as large as the former three.

APPENDIX A: DATA DESCRIPTION

TABLE 5

References for direction of relationship. Papers in the literature were referenced for established directions of monotonic relationships. The relationship is classified as either positive (*monotonic increasing*), *negative or unclear. The unclear category contains nonmonotonic variables, variables whose literature is undecided on the direction as well as variables whose relationship with returns is unclear.*

Variable	Description	Papers	Monotonic direction
a2me	Total assets to size	Bhandari (1988)	Unclear
at	Total assets	Gandhi and Lustig (2015)	Unclear
ato	Asset turnover: Sales to lagged net operating assets	Soliman (2008)	Positive
beme	Book to market ratio	Lewellen (2015)	Positive
beta	CAPM Beta	Lewellen (2015)	Negative
c	Cash to total assets	Palazzo (2012)	Positive
cto	Sales to lagged total assets	Haugen and Baker (1996)	Unclear
d2a	Depreciation and amortization (DP) to total assets (AT)	Gorodnichenko and Weber (2016)	Unclear

TABLE 5
(*Continued*)

Variable	Description	Papers	Monotonic direction
dpi2a	Change in PP&E and inventory over lagged assets (AT)	Lyandres, Sun and Zhang (2007)	Negative
e2p	Earnings to price	Basu (1983)	Positive
fc2y	Fixed costs to sales	D'Acunto et al. (2018)	Unclear
free_cf	Free cash flow to book equity	Hou, Karolyi and Kho (2011)	Positive
idio_vol	Idiosyncratic volatility from Fama-French 3 factor model	Ang et al. (2005)	Negative
investment	Percent change in total assets	Cooper, Gulen and Schill (2008)	Negative
lev	Leverage	Bhandari (1988); Fama and French (1992); Lewellen (2015)	Positive
lme	Size: Market equity defined as stock price times shares outstanding	Fama and French (1992); Fama and French (2008); Lewellen (2015)	Negative
lturnover	Volume to shares outstanding (turnover)	Datar, Naik and Radcliffe (1998)	Negative
noa	Net-operating assets over lagged assets (AT)	Hirshleifer et al. (2004)	Negative
oa	Operating accruals	Sloan (1996)	Unclear
ol	Costs of goods sold + SG&A to total assets	Novy-Marx (2010)	Positive
pcm	Price-to-cost margin: Sales minus costs of goods sold to sales	Bustamante and Donangelo (2017); Gorodnichenko and Weber (2016); D'Acunto et al. (2018)	Positive
pm	Profit margin: OI after depreciation over sales	Soliman (2008)	Positive
prof	Profitability: Gross profitability over BE	Ball et al. (2015); Lewellen (2015)	Positive
q	Tobin's Q		Unclear
r_{12-2}	Momentum	Lewellen (2015)	Positive
r_{12-7}	Intermediate momentum	Novy-Marx (2012)	Positive
r_{2-1}	Short-term reversal		Unclear
r_{36-13}	Long-term reversal	De Bondt and Thaler (1985)	Unclear
rel_to_high_price	Price to 52-week-high price	George and Hwang (2004)	Positive
rna	Return on net operating assets: OI after depreciation to lagged net operating assets	Soliman (2008)	Positive
roa	Return on assets: Income before extraordinary items to lagged AT	Balakrishnan, Bartov and Faurel (2010)	Positive
roe	Return on equity: Income before extraordinary items to lagged BE	Haugen and Baker (1996)	Positive
s2p	Sales to price	Fama and French (1992); Lewellen (2015)	Positive
sga2m	Expenses-to-sales: Ratio of expenses (XSGA) to net sales (SALE)		Unclear
spread_mean	Average daily bid-ask spread	Chung and Zhang (2014)	Unclear
suv	Standard unexplained volume	Garfinkel (2009)	Unclear

APPENDIX B: STATISTICAL FORMULATION AND COMPUTATION

Model summary. In summary,

$$\begin{aligned}
\boldsymbol{r}_t|\boldsymbol{\theta}_t &\sim N\left(\alpha_t \mathbf{1}_{n_t} + \sum_{k=1}^{K} f_{kt}(\boldsymbol{x}_{k,t-1}), \sigma_t^2 I_n\right),\\
f_{kt}(\boldsymbol{x}_{k,t-1}) &= X_{k,t-1}\boldsymbol{\beta}_{kt} = X_{k,t-1}L^{-1}L\boldsymbol{\beta}_{kt} = W_{kt}L\boldsymbol{\beta}_{kt} = W_{kt}\boldsymbol{\gamma}_{kt},\\
\alpha_t &\sim N(0, 10^{-2}),\\
\sigma_t^2 &\sim U(0, 10^3),\\
(\gamma_{jkt}|I_{jkt} = 1, \cdot) &\sim N_+(0, c_k\sigma_t^2),\\
(\gamma_{jkt}|I_{jkt} = 0) &= 0,\\
I_{jkt} &\sim \mathrm{Bn}(p_{jk} = 0.2).
\end{aligned}$$

Spline conditions. This yields $\grave{m} + \acute{m} + 3$ conditions to satisfy:

$$\begin{aligned}
0 \le f'(-0.5) &= \beta_1 + 2\beta_2(-0.5) + 2\beta_3(-0.5 - \grave{x}_1) + \cdots\\
&\quad + 2\beta_{\grave{m}+2}(-0.5 - \grave{x}_{\grave{m}}),\\
0 \le f'(\grave{x}_{\grave{m}}) &= \beta_1 + 2\beta_2(\grave{x}_{\grave{m}}) + 2\beta_3(\grave{x}_{\grave{m}} - \grave{x}_1) + \cdots\\
&\quad + 2\beta_{\grave{m}+1}(\grave{x}_{\grave{m}} - \grave{x}_{\grave{m}-1}),\\
&\vdots\\
0 \le f'(\grave{x}_2) &= \beta_1 + 2\beta_2(\grave{x}_2) + 2\beta_3(\grave{x}_2 - \grave{x}_1),\\
0 \le f'(\grave{x}_1) &= \beta_1 + 2\beta_2(\grave{x}_1),\\
0 \le f'(0) &= \beta_1,\\
0 \le f'(\acute{x}_1) &= \beta_1 + 2\beta_{\grave{m}+3}(\acute{x}_1),\\
0 \le f'(\acute{x}_2) &= \beta_1 + 2\beta_{\grave{m}+3}(\acute{x}_2) + 2\beta_{\grave{m}+4}(\acute{x}_2 - \acute{x}_1),\\
&\vdots\\
0 \le f'(\acute{x}_{\acute{m}}) &= \beta_1 + 2\beta_{\grave{m}+3}(\acute{x}_{\acute{m}}) + 2\beta_{\grave{m}+4}(\acute{x}_{\acute{m}} - \acute{x}_1) + \cdots\\
&\quad + 2\beta_{\grave{m}+\acute{m}+2}(\acute{x}_{\acute{m}} - \acute{x}_{\acute{m}-1}),\\
0 \le f'(0.5) &= \beta_1 + 2\beta_{\grave{m}+3}(0.5) + 2\beta_{\grave{m}+4}(0.5 - \acute{x}_1) + \cdots\\
&\quad + 2\beta_{\grave{m}+\acute{m}+3}(0.5 - \acute{x}_{\acute{m}}),
\end{aligned}$$

which can be vectorized as a system of $\grave{m} + \acute{m} + 3$ linear inequalities, and these inequalities serve as our monotonicity conditions.

The MCMC sampler. To sample all parameters at time $\tau \in \{1, \ldots, T\}$, iterate through the following, conditional upon the most recent draws of other parameters:

1. Draw $\alpha_\tau \sim N(m_\alpha, v_\alpha)$:
   - $m_\alpha = \frac{v_\alpha}{\sigma^2}\sum_{t=1}^{\tau} \omega_t \mathbf{1}_{n_t}'(\boldsymbol{r}_t - \sum_{k=1}^{K} W_{kt}\boldsymbol{\gamma}_{k\tau})$ and

- $v_\alpha = (\frac{1}{\sigma^2}\sum_{t=1}^{\tau}\omega_t n_t + \frac{1}{10^{-2}})^{-1}$.

2. Draw $\sigma_\tau^2 \sim IG(a_\sigma, b_\sigma)$, where:
   - $a_\sigma = \frac{1}{2}(\sum_{t=1}^{\tau} n_t\omega_t + \sum_{j=1}^{m}\sum_{k=1}^{K} I_{jk\tau}) - 1$ and
   - $b_\sigma = \frac{1}{2}(\sum_{t=1}^{\tau}\omega_t \boldsymbol{e}_t'\boldsymbol{e}_t + \sum_{j=1}^{m}\sum_{k=1}^{K}\frac{\gamma_{jk\tau}^2}{c_k})$ for the residual $\boldsymbol{e}_t = \boldsymbol{r}_t - \alpha_\tau \mathbf{1}_{n_t} - \sum_{k=1}^{K} W_{kt} \times \boldsymbol{\gamma}_{k\tau}$.

3. For coefficients $j = 1, \ldots, m+2$ and characteristics $k = 1, \ldots, K$:
   (a) Draw $I_{jk\tau} \sim \text{Bernoulli}(p^*_{jk\tau})$ where:
   - $p^*_{jk\tau} = \frac{\hat{p}_{jk\tau}}{\hat{p}_{jk\tau}+(1-p_{jk})}$,
   - $\hat{p}_{jk\tau} = 2p_{jk}c_k^{-\frac{1}{2}} v_{\gamma_{jk\tau}}^{\frac{1}{2}} \exp\{\frac{1}{2\sigma^2 v_{\gamma_{jk\tau}}} m^2_{\gamma_{jk\tau}}\}[1 - \Phi(0|m_{\gamma_{jk\tau}}, \sigma^2 v_{\gamma_{jk\tau}})]$,
   - $m_{\gamma_{jk\tau}} = v_{\gamma_{jk\tau}} \sum_{t=1}^{\tau}\omega_t \boldsymbol{e}'_{(jk)t}\boldsymbol{w}_{jkt}$,
   - $v_{\gamma_{jk\tau}} = (\sum_{t=1}^{\tau}\omega_t \boldsymbol{w}'_{jkt}\boldsymbol{w}_{jkt} + \frac{1}{c_k})^{-1}$,
   - $\boldsymbol{e}_{(jk)t} = \boldsymbol{r}_t - \alpha_\tau \mathbf{1}_{n_t} - \sum_{\ell\neq k} W_{\ell t}\boldsymbol{\gamma}\boldsymbol{\tau}_\ell - \sum_{\ell\neq j}\boldsymbol{w}_{\ell kt}\gamma_{\ell k\tau}$, the residual assuming $\gamma_{jk\tau} = 0$.

   (b) If $I_{jk\tau} = 1$ then draw $\gamma_{jk\tau} \sim N_+(m_{\gamma_{jk\tau}}, \sigma^2 v_{\gamma_{jk\tau}})$, else $\gamma_{jk\tau} = 0$.

APPENDIX C: MAIN RESULTS OVER DIFFERENT SUBPERIODS

Here, we present the results of Table 2 with two subperiods of 20 years, 1974–1994 and 1995–2014, instead of three subperiods.

TABLE 6

Out-of-sample model performance when forecasting monthly returns over over the January 1974 *to December* 1994, *January* 1995 *to May* 2014 *periods. Out-of-sample R^2 is calculated, as done in Gu, Kelly and Xiu (2020), and given in percentage points. Sharpe ratios use the returns from equal-weighted decile portfolios, long on the firms with forecasted in the top decile and shorting the bottom decile. Bolded values indicate the best two models for a given window length*

	OOS R^2				Sharpe ratio			
	Window size (months)				Window size (months)			
	All	120	60	36	All	120	60	36
				1974–1994				
OLS	0.64	0.48	0.39	0.11	3.42	3.33	3.05	2.88
Random Forest	0.78	0.65	0.58	0.29	4.33	4.25	3.58	3.06
BART	**1.53**	**1.30**	**1.10**	**0.75**	**4.49**	**4.41**	**4.29**	**4.03**
Splines-0	**1.00**	0.86	0.76	0.44	4.28	4.25	4.05	3.83
Splines-6	1.00	**0.86**	**0.78**	**0.45**	**4.34**	4.28	4.09	3.82
Splines-24	0.91	0.77	0.71	0.42	4.32	**4.35**	**4.13**	**3.85**
				1995–2014				
OLS	0.55	0.43	−0.33	−0.95	1.84	1.67	1.01	0.76
Random Forest	0.73	0.62	−0.10	−0.84	**2.33**	1.79	1.06	0.61
BART	**1.05**	**0.80**	−0.25	−1.33	**2.67**	**2.51**	**1.90**	1.39
Splines-0	0.80	0.59	−0.04	−0.49	2.22	2.19	1.73	1.42
Splines-6	**0.81**	0.59	**−0.04**	**−0.45**	2.24	2.19	1.75	**1.50**
Splines-24	0.76	**0.63**	**0.03**	**−0.36**	2.33	**2.44**	**1.83**	**1.51**

Acknowledgments. The authors would like to thank the anonymous referees, an Associate Editor and the Editor, Brendan Murphy, for their constructive comments that improved the quality of this paper. We would also like to thank Andreas Neuhierl for sharing data with us. The first author was supported by NSF Grant DMS-1745640. The third author was supported by the Salem Center for Policy at McCombs.

REFERENCES

ANG, A. and KRISTENSEN, D. (2012). Testing conditional factor models. *J. Financ. Econ.* **106** 132–156. https://doi.org/10.1016/j.jfineco.2012.04.008

ANG, A., HODRICK, R., XING, Y. and ZHANG, X. (2006). The cross-section of volatility and expected returns. *J. Finance* **61** 259–299. https://doi.org/10.1111/j.1540-6261.2006.00836.x

BALAKRISHNAN, K., BARTOV, E. and FAUREL, L. (2010). Post loss/profit announcement drift. *J. Account. Econ.* **50** 20–41. https://doi.org/10.1016/j.jacceco.2009.12.002

BALL, R., GERAKOS, J., LINNAINMAA, J. T. and NIKOLAEV, V. V. (2015). Deflating profitability. *J. Financ. Econ.* **117** 225–248. https://doi.org/10.1016/j.jfineco.2015.02.004

BASU, S. (1983). The relationship between earnings' yield, market value and return for NYSE common stocks: Further evidence. *J. Financ. Econ.* **12** 129–156. https://doi.org/10.1016/0304-405X(83)90031-4

BEKAERT, G., HARVEY, C. R. and LUMSDAINE, R. L. (2002). Dating the integration of world equity markets. *J. Financ. Econ.* **65** 203–247. https://doi.org/10.1016/S0304-405X(02)00139-3

BHANDARI, C. L. (1988). Debt/equity ratio and expected common stock returns: Empirical evidence. *J. Finance* **43** 507–528. https://doi.org/10.1111/j.1540-6261.1988.tb03952.x

BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.

BUSTAMANTE, C. M. and DONANGELO, A. (2017). Product market competition and industry returns. *Rev. Financ. Stud.* **30** 4216–4266. https://doi.org/10.1093/rfs/hhx033

CATTANEO, M. D., CRUMP, R. K., FARRELL, M. H. and SCHAUMBURG, E. (2020). Characteristic-sorted portfolios: Estimation and inference. *Rev. Econ. Stat.* **102** 531–551. https://doi.org/10.1162/rest_a_00883

CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4** 266–298. MR2758172 https://doi.org/10.1214/09-AOAS285

CHIPMAN, H. A., GEORGE, E. I., MCCULLOCH, R. E. and SHIVELY, T. S. (2019). High-dimensional nonparametric monotone function estimation using BART. Preprint. Available at arXiv:1612.01619.

CHUNG, K. H. and ZHANG, H. (2014). A simple approximation of intraday spreads using daily data. *J. Financ. Mark.* **17** 94–120. https://doi.org/10.1016/j.finmar.2013.02.004

COCHRANE, J. H. (2011). Presidential address: Discount rates. *J. Finance* **66** 1047–1108. https://doi.org/10.1111/j.1540-6261.2011.01671.x

COOPER, M., GULEN, H. and SCHILL, M. (2008). Asset growth and the cross-section of stock returns. *J. Finance* **63** 1609–1651. https://doi.org/10.1111/j.1540-6261.2008.01370.x

D'ACUNTO, F., LIU, R., PFLUEGER, C. and WEBER, M. (2018). Flexible prices and leverage. *J. Financ. Econ.* **129** 46–68. https://doi.org/10.1016/j.jfineco.2018.03.009

DATAR, V. T., NAIK, N. Y. and RADCLIFFE, R. (1998). Liquidity and stock returns: An alternative test. *J. Financ. Mark.* **1** 203–219. https://doi.org/10.1016/S1386-4181(97)00004-9

DE BONDT, W. and THALER, R. (1985). Does the stock market overreact? *J. Finance* **40** 793–805. https://doi.org/10.1111/j.1540-6261.1985.tb05004.x

FAMA, E. and FRENCH, K. (1992). The cross-section of expected stock returns. *J. Finance* **47** 427–465. https://doi.org/10.1111/j.1540-6261.1992.tb04398.x

FAMA, E. F. and FRENCH, K. R. (1993). Common risk factors in the returns on stocks and bonds. *J. Financ. Econ.* **33** 3–56. https://doi.org/10.1016/0304-405X(93)90023-5

FAMA, E. F. and FRENCH, K. R. (2008). Dissecting anomalies. *J. Finance* **63** 1653–1678. https://doi.org/10.1111/j.1540-6261.2008.01371.x

FAMA, E. F. and FRENCH, K. R. (2016). Dissecting anomalies with a five-factor model. *Rev. Financ. Stud.* **29** 69–103. https://doi.org/10.1093/rfs/hhv043

FAMA, E. F. and MACBETH, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *J. Polit. Econ.* **81** 607–636.

FREYBERGER, J., NEUHIERL, A. and WEBER, M. (2020). Dissecting characteristics nonparametrically. *Rev. Financ. Stud.* **33** 2326–2377. https://doi.org/10.1093/rfs/hhz123

GANDHI, P. and LUSTIG, H. (2015). Size anomalies in U.S. bank stock returns. *J. Finance* **70** 733–768. https://doi.org/10.1111/jofi.12235

GARFINKEL, J. (2009). Measuring investors' opinion divergence. *J. Acc. Res.* **47** 1317–1348. https://doi.org/10.1111/j.1475-679X.2009.00344.x

GEORGE, T. and HWANG, C. (2004). The 52-week high and momentum investing. *J. Finance* **59** 2145–2176. https://doi.org/10.1111/j.1540-6261.2004.00695.x

GORODNICHENKO, Y. and WEBER, M. (2016). Are sticky prices costly? Evidence from the stock market. *Am. Econ. Rev.* **106** 165–99. https://doi.org/10.1257/aer.20131513

GU, S., KELLY, B. and XIU, D. (2020). Empirical asset pricing via machine learning. *Rev. Financ. Stud.*. https://doi.org/10.1093/rfs/hhaa009

HAHN, P. R. and CARVALHO, C. M. (2015). Decoupling shrinkage and selection in Bayesian linear models: A posterior summary perspective. *J. Amer. Statist. Assoc.* **110** 435–448. MR3338514 https://doi.org/10.1080/01621459.2014.993077

HAN, Y., HE, A., RAPACH, D. and ZHOU, G. (2019). Firm characteristics and expected stock returns. Working paper.

HARVEY, C. R., LIU, Y. and ZHU, H. (2016). ... and the cross-section of expected returns. *Rev. Financ. Stud.* **29** 5–68. https://doi.org/10.1093/rfs/hhv059

HAUGEN, R. A. and BAKER, N. L. (1996). Commonality in the determinants of expected stock returns. *J. Financ. Econ.* **41** 401–439. https://doi.org/10.1016/0304-405X(95)00868-F

HE, J., YALOV, S. and HAHN, P. R. (2019). XBART: Accelerated Bayesian additive regression trees. In *The 22nd International Conference on Artificial Intelligence and Statistics* (*AISTATS*).

HIRSHLEIFER, D., KEWEI, H., TEOH, S. H. and YINGLEI, Z. (2004). Do investors overvalue firms with bloated balance sheets? *J. Account. Econ.* **38** 297–331. https://doi.org/10.1016/j.jacceco.2004.10.002

HOU, K., KAROLYI, A. G. and KHO, B.-C. (2011). What Factors Drive Global Stock Returns?. *Rev. Financ. Stud.* **24** 2527–2574.

JEGADEESH, N. and TITMAN, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *J. Finance* **48** 65–91.

JEGADEESH, N. and TITMAN, S. (2001). Profitability of momentum strategies: An evaluation of alternative explanations. *J. Finance* **56** 699–720.

LETTAU, M. and NIEUWERBURGH, S. V. (2008). Reconciling the return predictability evidence. *Rev. Financ. Stud.* **21** 1607–1652.

LEWELLEN, J. (2015). The cross-section of expected stock returns. *Crit. Finance Rev.* **4** 1–44. https://doi.org/10.1561/104.00000024

LYANDRES, E., SUN, L. and ZHANG, L. (2007). The new issues puzzle: Testing the investment-based explanation. *Rev. Financ. Stud.* **21** 2825–2855. https://doi.org/0.1093/rfs/hhm058

MCCARTHY, D. and JENSEN, S. T. (2016). Power-weighted densities for time series data. *Ann. Appl. Stat.* **10** 305–334. MR3480498 https://doi.org/10.1214/15-AOAS893

NOVY-MARX, R. (2010). Operating leverage. *Rev. Finance* **15** 103–134. https://doi.org/10.1093/rof/rfq019

NOVY-MARX, R. (2012). Is momentum really momentum? *J. Financ. Econ.* **103** 429–453. https://doi.org/10.1016/j.jfineco.2011.05.003

PALAZZO, B. (2012). Cash holdings, risk, and expected returns. *J. Financ. Econ.* **104** 162–185. https://doi.org/10.1016/j.jfineco.2011.12.009

PASTOR, L. and STAMBAUGH, R. F. (2001). The equity premium and structural breaks. *J. Finance* **56** 1207–1239. https://doi.org/10.1111/0022-1082.00365

PATTON, A. J. and TIMMERMANN, A. (2010). Monotonicity in asset returns: New tests with applications to the term structure, the CAPM, and portfolio sorts. *J. Financ. Econ.* **98** 605–625.

PETTENUZZO, D. and TIMMERMANN, A. (2011). Predictability of stock returns and asset allocation under structural breaks. *J. Econometrics* **164** 60–78. MR2821794 https://doi.org/10.1016/j.jeconom.2011.02.019

PUELZ, D. W. (2018). Regularization in econometrics and finance. Ph.D. thesis, Univ. Texas at Austin, McCombs School of Business.

PUELZ, D., CARVALHO, C. M. and HAHN, P. R. (2015). Optimal ETF selection for passive investing.

PUELZ, D., HAHN, P. R. and CARVALHO, C. M. (2017). Variable selection in seemingly unrelated regressions with random predictors. *Bayesian Anal.* **12** 969–989. MR3724975 https://doi.org/10.1214/17-BA1053

PUELZ, D., HAHN, P. R. and CARVALHO, C. M. (2020). Portfolio selection for individual passive investing. *Appl. Stoch. Models Bus. Ind.* **36** 124–142.

SHIVELY, T. S., SAGER, T. W. and WALKER, S. G. (2009). A Bayesian approach to non-parametric monotone function estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 159–175. MR2655528 https://doi.org/10.1111/j.1467-9868.2008.00677.x

SLOAN, R. G. (1996). Do stock prices fully reflect information in accruals and cash flows about future earnings? *Account. Rev.* **71** 289–315.

SMITH, S. and TIMMERMANN, A. (2018). Break risk. Working paper.

SMITH, S. C., TIMMERMANN, A. and ZHU, Y. (2019). Variable selection in panel models with breaks. *J. Econometrics* **212** 323–344. MR3994020 https://doi.org/10.1016/j.jeconom.2019.04.033

SOLIMAN, M. T. (2008). The use of DuPont analysis by market participants. *Accoun. Rev.* **83** 823–853.
WRIGHT, M. N. and ZIEGLER, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Softw.* **77** Issue 1. https://doi.org/10.18637/jss.v077.i01
YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67. MR2212574 https://doi.org/10.1111/j.1467-9868.2005.00532.x