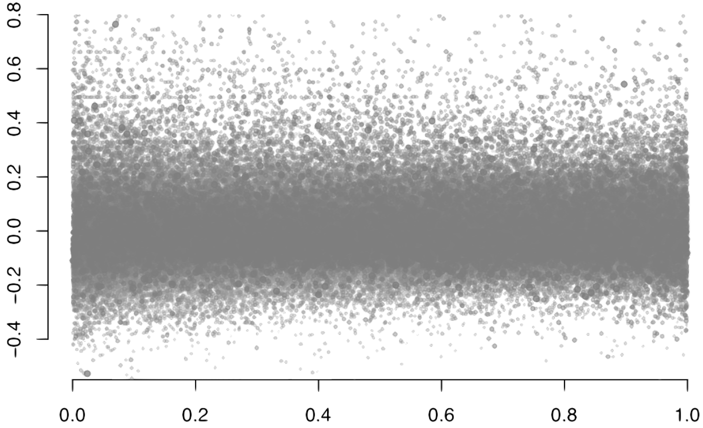
A Flexible Model for Returns

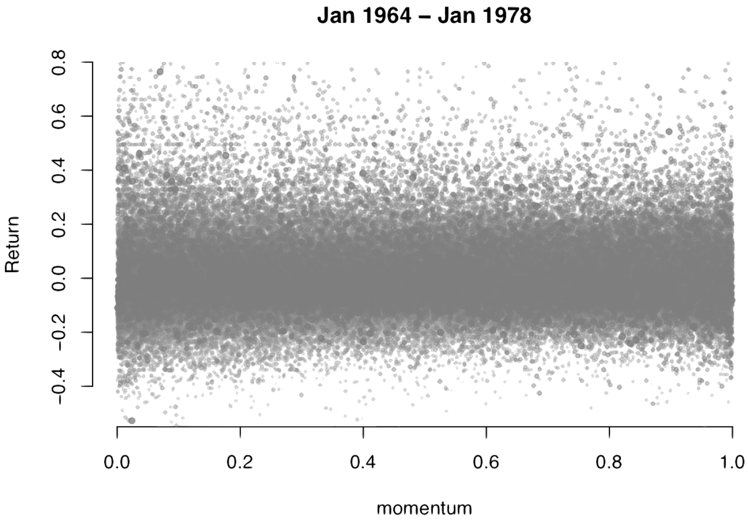
David Puelz
Jared Fisher (University of Texas)
Carlos Carvalho (University of Texas)

December 10, 2018

The data



The data: Monthly returns



Given this observed data, our motivating question is:

What firm characteristics are predictive of its return?

There are two parts to the answer:

1. Building an appropriate (Bayesian) model.

2. Selecting characteristics (given statistical uncertainty).

The object of interest

The conditional expectation of returns given observed characteristics

$$\mathbb{E}[R_{it} \mid \mathsf{X}_{it-1}] = f\left(\mathsf{X}_{it-1}\right)$$

R_{it}: excess return of firm i at time t
X_{it-1}: vector of characteristics of firm i at time t

The object of interest

The conditional expectation of returns given observed characteristics

$$\mathbb{E}[R_{it} \mid \mathrm{X}_{it-1}] = f\left(\mathrm{X}_{it-1}\right)$$

R_{it}: excess return of firm i at time t
X_{it-1}: vector of characteristics of firm i at time t

We would like to learn f

The object of interest

The conditional expectation of returns given observed characteristics

$$\mathbb{E}[R_{it} \mid \mathsf{X}_{it-1}] = f\left(\mathsf{X}_{it-1}\right)$$

R_{it}: excess return of firm i at time t
X_{it-1}: vector of characteristics of firm i at time t

We would like to learn f
and which X^{k}_{it-1}'s matter!

Step 1: A model for f

Portfolio sorts are one way ...

Jegadeesh and Titman (2001)

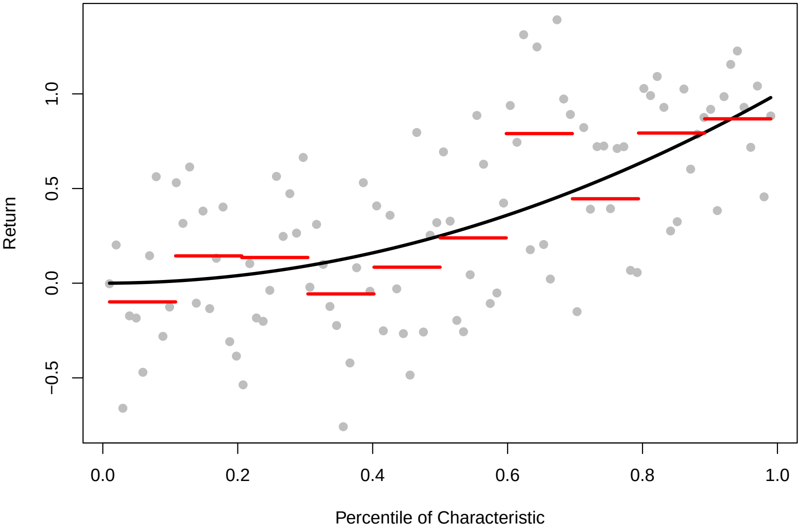
Table I

Momentum Portfolio Returns

This table reports the monthly returns for momentum portfolios formed based on past six-month returns and held for six months. P1 is the equal-weighted portfolio of 10 percent of the stocks with the highest returns over the previous six months, P2 is the equal-weighted portfolio of the 10 percent of the stocks with the next highest returns, and so on. The "All stocks" sample includes all stocks traded on the NYSE, AMEX, or Nasdaq excluding stocks priced less than $5 at the beginning of the holding period and stocks in the smallest market cap decile (NYSE size decile cutoff). The "Small Cap" and "Large Cap" subsamples comprise stocks in the "All Stocks" sample that are smaller and larger than the median market cap NYSE stock respectively. "EWI" is the returns on the equal-weighted index of stocks in each sample.

	All Stocks			Small Cap			Large Cap		
	1965–1998	1965–1989	1990–1998	1965–1998	1965–1989	1990–1998	1965–1998	1965–1989	1990–1998
P1 (Past winners)	1.65	1.63	1.69	1.70	1.69	1.73	1.56	1.52	1.66
P2	1.39	1.41	1.32	1.45	1.50	1.33	1.25	1.24	1.27
P3	1.28	1.30	1.21	1.37	1.42	1.23	1.12	1.10	1.19
P4	1.19	1.21	1.13	1.26	1.34	1.05	1.10	1.07	1.20
P5	1.17	1.18	1.12	1.26	1.33	1.06	1.05	1.00	1.19
P6	1.13	1.15	1.09	1.19	1.26	1.01	1.09	1.05	1.20
P7	1.11	1.12	1.09	1.14	1.20	0.99	1.09	1.04	1.23
P8	1.05	1.05	1.03	1.09	1.17	0.89	1.04	1.00	1.17
P9	0.90	0.94	0.77	0.84	0.95	0.54	1.00	0.96	1.09
P10 (Past losers)	0.42	0.46	0.30	0.28	0.35	0.08	0.70	0.68	0.78
P1–P10	1.23	1.17	1.39	1.42	1.34	1.65	0.86	0.85	0.88
t statistic	6.46	4.96	4.71	7.41	5.60	5.74	4.34	3.55	2.59
EWI	1.09	1.10	1.04	1.13	1.19	0.98	1.03	1.00	1.12

A generated example



Challenges and a solution

- X_{it-1} is multidimensional.
- No information sharing across the X space.
- Even if we had only 12 characteristics and sorted into quintiles along each dimension, that requires constructing $5^{12} =$ **244140625** portfolios!

We propose modeling the CEF using additive quadratic splines (with monotonicity constraints *and* time variation):

$$\mathbb{E}[R_{it} \mid X_{it-1}] = \alpha_t + \sum_{k=1}^{K} g_{kt}(x_{ki,t-1})$$

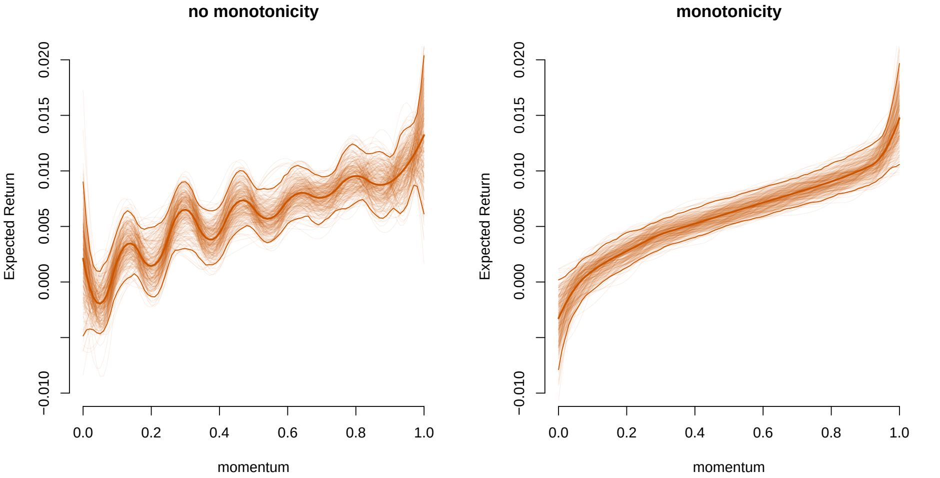
Features of our model

1. Monotonicity on partial effects are incorporated by constraint on the spline coefficients.
   $\rightarrow$ improvement over nonlinear models that fit to noise.

2. Time dynamics modeled using a power-weighting likelihood approach.
   $\rightarrow$ improvement over rolling window models.

Why monotonicity?

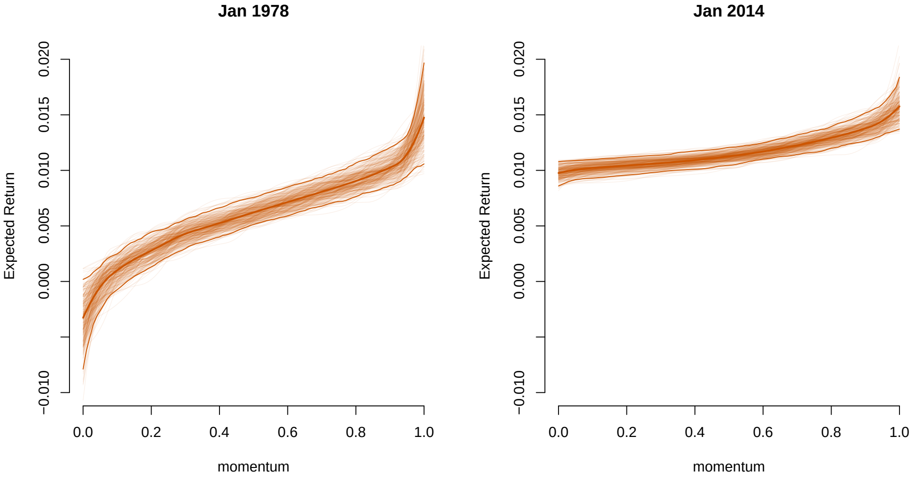
Estimated functions at January 1978



monotonicity is enforced by linear constraints on spline coefficients

Why time variation?

Estimated functions at January 1978 and 2014



dynamics are modeled by likelihood discounting, McCarthy and Jenson (2016)

Step 2: Characteristic selection

In light of all statistical uncertainty:

Which characteristics in X_{it-1}^k matter?

(Does this selected set vary over time?)

This problem (and many others like it!) can be studied using variable selection techniques from statistics to induce *sparsity*.

Sparsity

This problem (and many others like it!) can be studied using variable selection techniques from statistics to induce *sparsity*.

What's typically done? (broadly speaking)

- **Bayesian:** Shrinkage prior design.
- **Frequentist:** Penalized likelihood methods.

Common theme? Sparsity and inference go hand in hand.

Our view: Subset selection is a decision problem. We need a suitable loss function, **not** a more clever prior.

Separating priors from utilities

Our view: Subset selection is a decision problem. We need a suitable loss function, **not** a more clever prior.

This leads us to think of selection in a "post-inference world" by comparing models (sets of characteristics) based on utility.*

*sparsity and statistical uncertainty play a key role in this *post-inference* exercise.

Utility-based selection: Primitives

Let b_t be a model decision, λ_t be a complexity parameter, Θ_t be a vector of model parameters, and $\tilde{R}_t$ be future data.

1. Loss function $\mathcal{L}(b_t, \tilde{R}_t)$ – measures utility.
2. Complexity function $\Phi(\lambda_t, b_t)$ – measures sparsity.
3. Statistical model $\Pi(\Theta_t)$ – characterizes uncertainty.

Utility-based selection: Procedure

- Optimize $\mathbb{E}[\mathcal{L}(b_t, \tilde{R}_t) + \Phi(\lambda_t, b_t)]$, where the expectation is over $p(\tilde{R}_t, \Theta_t \mid \mathbf{R})$.

- Calculate "regret" versus a target b_t^* for decisions indexed by λ_t.

  $$\rightarrow \quad \Delta(b_{\lambda_t}, b_t^*, \tilde{R}_t) = \mathcal{L}(b_{\lambda_t}, \tilde{R}_t) - \mathcal{L}(b_t^*, \tilde{R}_t)$$

- Posterior summary: Look at graphical summaries of optimal models, i.e.:

  $$\rightarrow \quad \pi_{\lambda_t} = \mathbb{P}[\Delta(b_{\lambda_t}, b_t^*, \tilde{R}_t) < 0] \text{ (satisfaction probability)}$$

What does this look like for our problem?

Primitives:

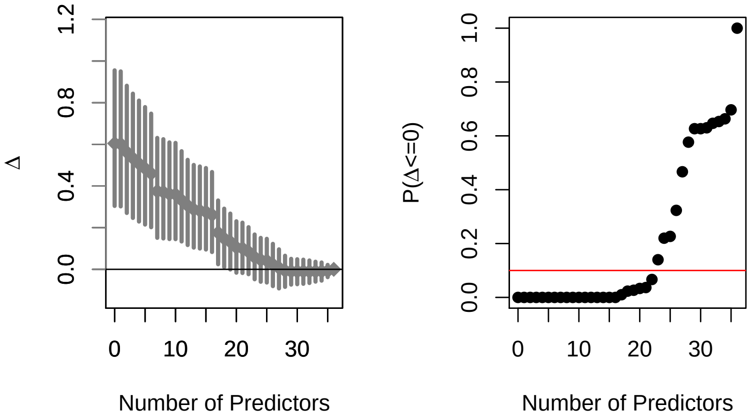
1. Loss: $\mathcal{L}(\tilde{\mathsf{R}}_t, \mathbf{A}_t, \Theta_t) = \frac{1}{2}(\tilde{\mathsf{R}}_t - \mathbb{X}_{t-1}\mathbf{A}_t)^T(\tilde{\mathsf{R}}_t - \mathbb{X}_{t-1}\mathbf{A}_t)$
2. Complexity: Group lasso penalty on the spline basis coefficients $\mathbf{A}_t$ defined as $\Phi(\lambda_t, \mathbf{A}_t)$
3. Model: Dynamic monotonic quadratic splines

Expected Loss:

Integrating over $p(\tilde{\mathsf{R}}_t, \Theta_t)$, we obtain:

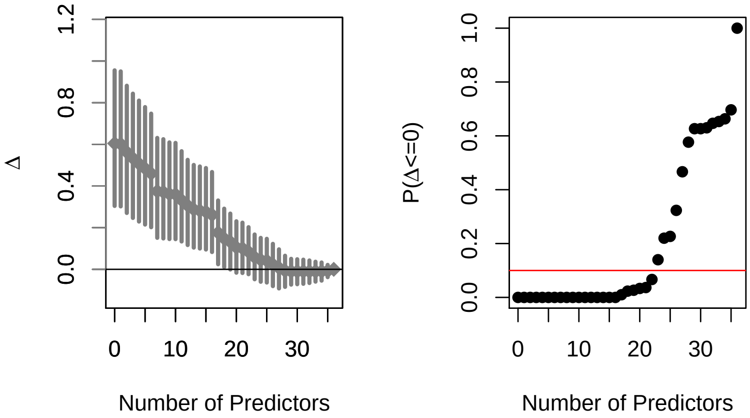
$$\mathcal{L}_{\lambda_t}(\mathbf{A}_t) = \left\|\mathbb{X}_{t-1}\mathbf{A}_t - \mathbb{X}_{t-1}\overline{\mathbf{B}}_t\right\|_2^2 + \Phi(\lambda_t, \mathbf{A}_t)$$

Procedure output: Posterior summary plots



- Like a LASSO solution path, but better!

Procedure output: Posterior summary plots



- Like a LASSO solution path, but better!
- Predictive uncertainty bands surround each expected utility optimal model.

Results

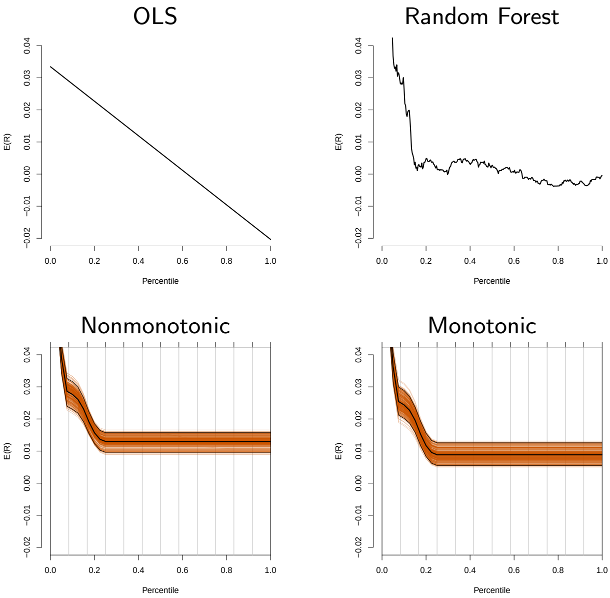
The data

Freyberger, Neuhierl, and Weber (2017)'s dataset:

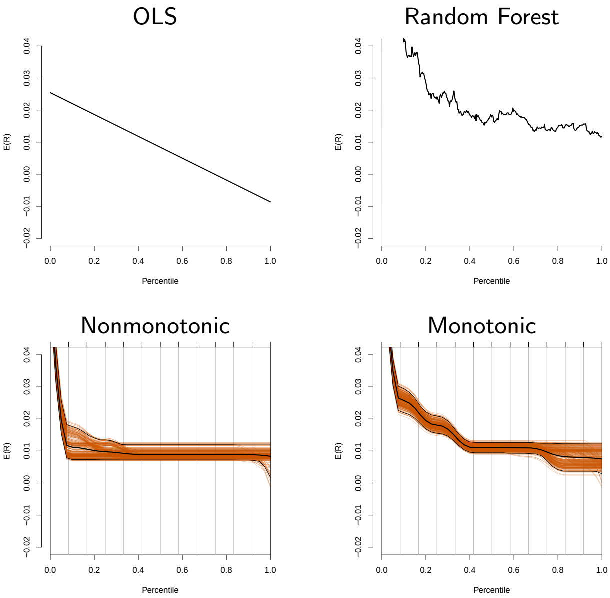
- CRSP monthly stock returns for most US traded firms
- 36 characteristics from Compustat and CRSP, including size, momentum, leverage, etc.
- July 1962 - June 2014

Presence and direction of monotonicity is determined by important papers in the literature.

Comparison of estimated functions – Size in 1994

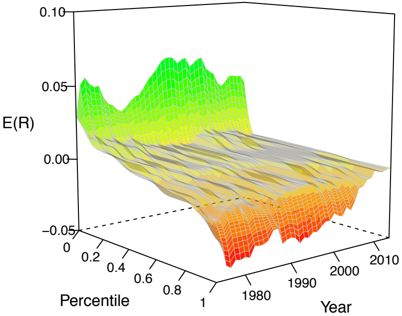


Comparison of estimated functions – Size in 2014

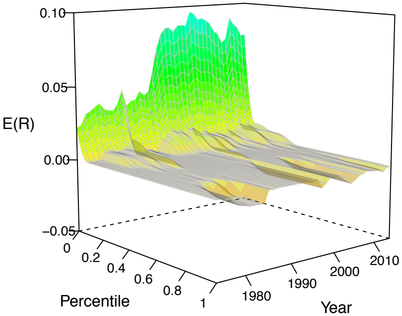


Dynamics of estimated functions

Short-term reversal

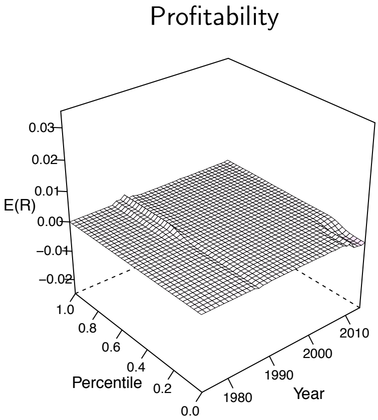


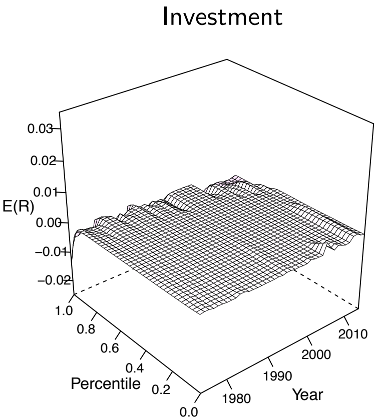
Size



Partial effects of characteristics change over time

Dynamics of estimated functions

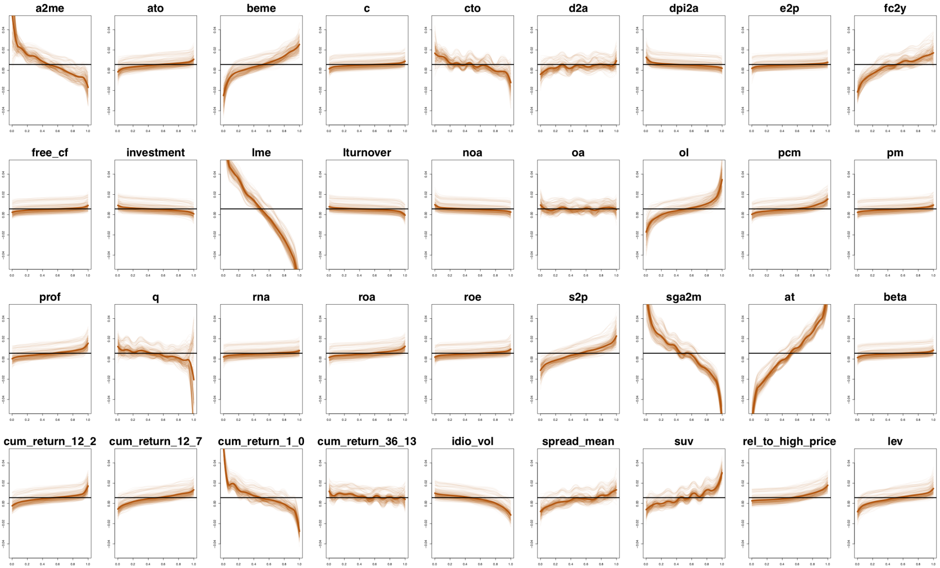




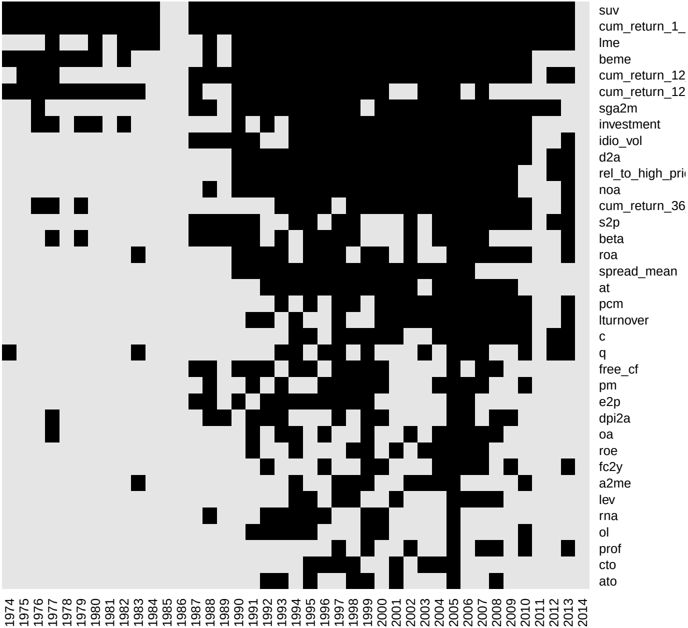
Not much evidence for the new Fama and French factors*

*conditional upon all other variables

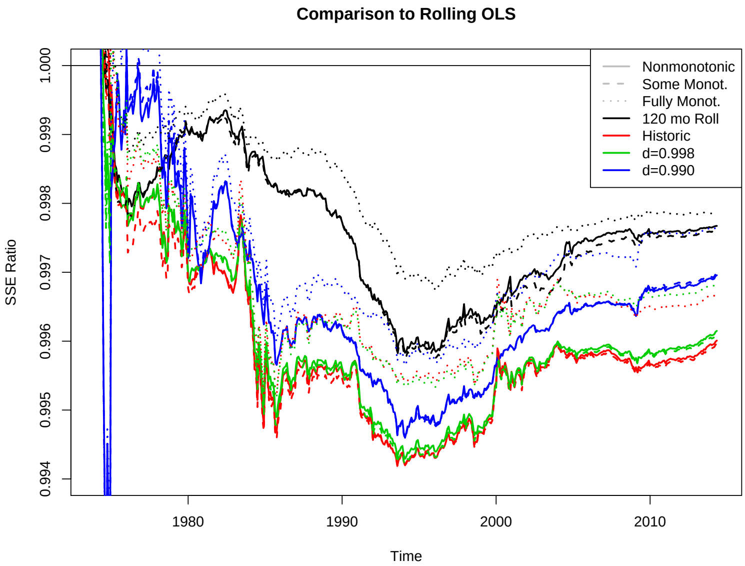
Which characteristic matter?



Volatility and momentum strategies selected often



Structure through monotonicity helps for prediction



Concluding thoughts, and thanks!

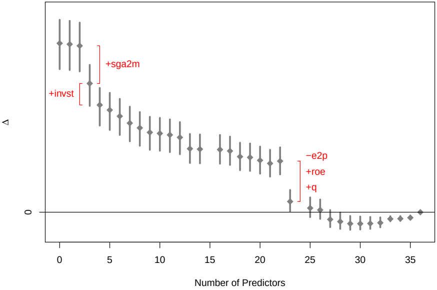
- Be wary of machine learning methods, especially when modeling finance data.

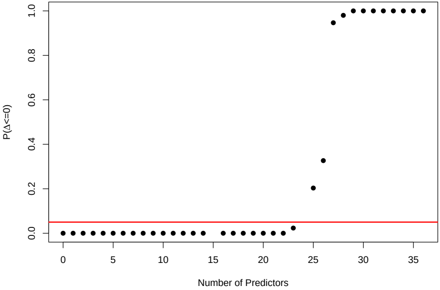
- **Utility functions can enforce inferential preferences that are not prior beliefs.**

- Statistical uncertainty should be used as a guide to avoid overfitting.

Extra slides

What characteristics matter over the entire period?

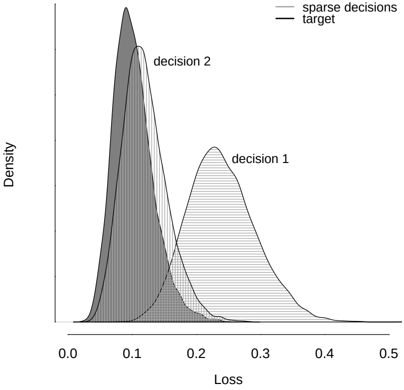


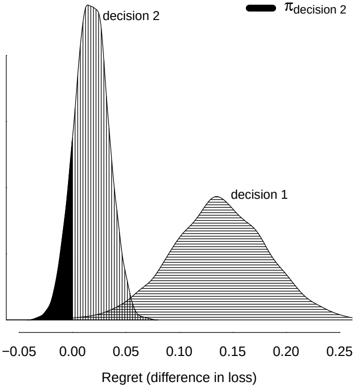


Regret-based selection: Illustration

d_λ : sparse decisions, d^* : target decision.

$\pi_\lambda = \mathbb{P}[\rho(d_\lambda, d^*, \tilde{Y}) < 0]$: probability of not regretting λ-decision.





UBS for Monotonic function estimation

The regression model is:

$$R_{it} = \alpha_t + \sum_{k=1}^{K} f_{kt}(x_{ki,t-1}) + \epsilon_{it}, \quad \epsilon_{it} \sim N(0, \sigma^2)$$

Insight – with quadratic splines for all f_{kt}, this can be written as a predictive regression:

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

where

$$\mathbb{X}_{t-1} = \begin{bmatrix} \mathbf{1}_{n_t} & \mathbf{X}_{t-1} \end{bmatrix}, \quad \mathbf{B}_t = \begin{bmatrix} \alpha_t & \boldsymbol{\beta}_t \end{bmatrix}$$

$\mathbf{X}_{t-1}$ is matrix of size $n_t \times K(m+2)$, $\boldsymbol{\beta}_t$ is vector of size $K(m+2)$. Therefore, each firm is given a row in $\mathbf{X}_{t-1}$, and each $m+2$ block of $\boldsymbol{\beta}_t$ corresponds to the coefficients on the spline basis for a particular characteristic, k.

We can now proceed as Hahn and Carvalho (2015). The loss function is the negative log density of the regression plus a penalty function Φ with parameter λ_t. Also, let the "sparsified action" for the coefficient matrix $\mathbf{A}_t$.

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

After integrating over $p(\tilde{\mathrm{R}}_t, \Theta_t)$, we obtain:

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

- Power-weighted likelihoods let information decay over time
- To estimate parameters at time τ, let $\delta_t = 0.99^{\tau - t}$, such that $\delta_1 \leq \delta_2 \leq ... \leq \delta_\tau = 1$, the likelihood at time $\tau \in \{1, ..., T\}$ is

$$p(R_1, ..., R_\tau | \Theta_\tau) = \prod_{t=1}^{\tau} p(R_t | \Theta_\tau)^{\delta_t}.$$

Model Summary

$$R_t|\cdot \sim N\left(\alpha_t 1_{n_t} + \sum_{k=1}^{K} f_{kt}(x_{k,t-1}),\ \sigma_t^2 I_n\right)^{\delta_t}$$

$$
\begin{aligned}
f_{kt}(x_{k,t-1}) &= X_{k,t-1}\beta_{kt} = X_{k,t-1}L^{-1}L\beta_{kt} = W_{kt}\gamma_{kt} \\
\alpha_t &\sim N(0, 10^{-2}) \\
\sigma_t^2 &\sim U(0, 10^3) \\
(\gamma_{jkt}|I_{jkt} = 1, \sigma_t^2) &\sim N_+(0, c_k\sigma_t^2) \\
(\gamma_{jkt}|I_{jkt} = 0) &= 0 \\
I_{jkt} &\sim Bn(p_{jk} = 0.2).
\end{aligned}
$$

Our Contribution

If we are serious about understanding the functional form of these partial relationships, then we should have

1. Additive splines: flexible and can separate to marginal effects
2. Monotonicity: complement the flexibility of the splines with a priori known structure
3. A single intercept: identifiable and intuitive
4. Time-dynamics modeled, not just a rolling window
5. Separation between the shrinkage of coefficients and selection of characteristics

1 - Additive Model

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

- $x_{k,i,t-1} \in (0,1)$ is the empirical percentile of characteristic k for firm i at time $t-1$, ranked over all firms
- Note that there are no interactions built into the model, as the intention is to see the partial effect

- For m known knots, $\tilde{x}_1, ..., \tilde{x}_m$,

$$f(x) = \beta_1 x + \beta_2 x^2 + \beta_3 (x - \tilde{x}_1)_+^2 + ... + \beta_{m+2} (x - \tilde{x}_m)_+^2$$

- For m known knots, $\tilde{x}_1, ..., \tilde{x}_m$,

$$f(x) = \beta_1 x + \beta_2 x^2 + \beta_3 (x - \tilde{x}_1)_+^2 + ... + \beta_{m+2} (x - \tilde{x}_m)_+^2$$

- Nondecreasing if all first derivatives are nonnegative

- For m known knots, $\tilde{x}_1, ..., \tilde{x}_m$,

$$f(x) = \beta_1 x + \beta_2 x^2 + \beta_3 (x - \tilde{x}_1)_+^2 + ... + \beta_{m+2}(x - \tilde{x}_m)_+^2$$

- Nondecreasing if all first derivatives are nonnegative
- Shively, Sager and Walker (2009) claim this yields $m + 2$ linear constraints:

$$L\beta \geq 0$$

2 - Monotonicity

For m knots, there are $m + 2$ conditions to satisfy:

$$
\begin{aligned}
&0 \leq f'_{kt}(0) = \beta_{1kt} \\
&0 \leq f'_{kt}(\tilde{x}_{1k}) = \beta_{1kt} + 2\beta_{2kt}\tilde{x}_{1k} \\
&0 \leq f'_{kt}(\tilde{x}_{2k}) = \beta_{1kt} + 2\beta_{2kt}\tilde{x}_{2k} + 2\beta_{3kt}(\tilde{x}_{2k} - \tilde{x}_{1k}) \\
&\ \vdots \\
&0 \leq f'_{kt}(1) = \beta_{1kt} + 2\beta_{2kt} + 2\beta_{3kt}(1 - \tilde{x}_{1k}) + \ldots + 2\beta_{m+2,kt}(1 - \tilde{x}_{mk})
\end{aligned}
$$

2 - Monotonicity

This can be vectorized as

$$0 \leq \begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 1 & 2\tilde{x}_{\mathbf{1}k} & 0 & \dots & 0 & 0 \\ 1 & 2\tilde{x}_{\mathbf{2}k} & 2(\tilde{x}_{\mathbf{2}k} - \tilde{x}_{\mathbf{1}k}) & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 2 & 2(1 - \tilde{x}_{\mathbf{1}k}) & \dots & 2(1 - \tilde{x}_{m-\mathbf{1},k}) & 2\beta_{m+\mathbf{2},kt}(1 - \tilde{x}_{mk}) \end{bmatrix} \beta_{kt}$$

- For m known knots, $\tilde{x}_1, ..., \tilde{x}_m$,

$$f(x) = \beta_1 x + \beta_2 x^2 + \beta_3 (x - \tilde{x}_1)_+^2 + ... + \beta_{m+2}(x - \tilde{x}_m)_+^2$$

- Nondecreasing if all first derivatives are nonnegative
- Shively, Sager and Walker (2009) show this yields $m + 2$ linear constraints:

$$L\beta \geq 0$$

  and the correct prior on $\gamma = L\beta$ will enforce monotonicity

- For m known knots, $\tilde{x}_1, ..., \tilde{x}_m$,

$$f(x) = \beta_1 x + \beta_2 x^2 + \beta_3 (x - \tilde{x}_1)_+^2 + ... + \beta_{m+2}(x - \tilde{x}_m)_+^2$$

- Nondecreasing if all first derivatives are nonnegative
- Shively, Sager and Walker (2009) show this yields $m + 2$ linear constraints:

$$L\beta \geq 0$$

  and the correct prior on $\gamma = L\beta$ will enforce monotonicity
- We use a modified version of their shrinkage prior:

$$\begin{aligned}(\gamma_j | I_j = 0) &\sim \delta_0 \\ (\gamma_j | I_j = 1) &\sim N_+(0, c\sigma^2) \\ I_j &\sim Bernoulli(0.2)\end{aligned}$$

3 - Intercept adjustment

Recall our additive model, with spline basis $X_{i,k,t-1}$ and a single intercept

$$\mathbb{E}(r_{it}|x_{i,t-1}) = \alpha_t + \sum_{k=1}^{K} X_{i,k,t-1}\beta_{kt}$$

$\Rightarrow$ α_t is the expected return for a firm with the minimum value for all characteristics, i.e. $X_{i,k,t-1} = 0, \forall k$.
Problems:

1. Computationally challenging due to few and volatile data points
2. Intuitively unfavorable as a baseline
3. Cannot see the lower tail effects change over time

Proposal: let the intercept be the expected return for a firm that has the median value for all characteristics

- Requires transforming the splines such that they equal 0 at the median $x = 0.5$ and not $x = 0$
- This then requires carefully expand spline basis and the monotonicity constraint matrix L

- Power-weighted likelihoods let information decay over time
- To estimate parameters at time τ, let $\delta_t = 0.99^{\tau - t}$, such that $\delta_1 \leq \delta_2 \leq ... \leq \delta_\tau = 1$, the likelihood at time $\tau \in \{1, ..., T\}$ is

$$p(r_1, ..., r_\tau | \Theta_\tau) = \prod_{t=1}^{\tau} p(r_t | \Theta_\tau)^{\delta_t}.$$

Empirics

- Freyberger, Neuhierl, and Weber (2017)'s dataset:
  - CRSP monthly stock returns for most US traded firms
  - 36 characteristics from Compustat and CRSP, including size, momentum, leverage, etc.
  - July 1962 - June 2014
- Model trained on 120 month rolling window
- Presence and direction of monotonicity is determined by important papers in the literature
  - "Fully" Monotonic model includes constraints on 24 of 36 predictors
  - "FF5" Monotonic model includes constraints on 6 predictors: size, book-to-market, investment, profitability, two horizons of momentum