Asset Pricing with Panel Trees Under Global Split Criteria*

Xin He Lin William Cong Guanhao Feng Jingyu He

October 2021

Abstract

We introduce a class of interpretable tree-based models (P-Trees) for analyzing panel data, with iterative and global (instead of recursive and local) splitting criteria to avoid overfitting and improve model performance. We apply P-Tree to generate a stochastic discount factor model and test assets for cross-sectional asset pricing. Unlike other tree algorithms, P-Trees accommodate imbalanced panels of asset returns and grow under the no-arbitrage condition. P-Trees also graphically capture nonlinearity and interaction effects and accommodate regime-switching and interactions between macroeconomic states and firm characteristics. For example, P-Tree identifies inflation as the most important macro predictor with regime-switching in U.S. equity data. Based on multiple pricing, prediction, and investment metrics, we find that (boosted or time-series) P-Trees outperform standard factor models and PCA latent factor models. An equally-weighted portfolio for five factors generated by P-Trees delivers an excess alpha of 1.09% against the Fama-French 3-factor benchmark, producing an annualized Sharpe ratio of 1.98 out-of-sample. Data-driven cutpoints in P-Trees reveal that long-run reversal, volume volatility, and industry-adjusted market equity drive cross-sectional return variations, consistent with variable importance analysis using random forests.

Key Words: CART, Cross-Sectional Returns, Latent Factor Model, Machine Learning, Panel Data, Stochastic Discount Factor, Tree Ensembles.

*We thank Richard Hahn, Stefan Nagel, Nick Polson, and Dacheng Xiu for invaluable discussions. We are grateful for helpful comments from seminar and conference participants at 2021 INFORMS Annual Meeting, Shanghai Jiao Tong University, Shanghai University of Finance and Economics, and the HKAIFT-Columbia joint seminar. Cong (E-mail: will.cong@cornell.edu) is at Cornell University SC Johnson College of Business; He (E-mail: xin.he@my.cityu.edu.hk), Feng (E-mail: gavin.feng@cityu.edu.hk), and He (E-mail: jingyuhe@cityu.edu.hk) are at the City University of Hong Kong.

1 Introduction

Decision tree algorithms such as the popular Classification and Regression Trees (CART, introduced in Breiman et al., 1984) are highly adaptive, nonlinear machine learning methods that have received wide and successful applications. They help split or cluster observations based on data features and predictors, effectively capture nonlinear and interaction effects, and importantly, admit their graphical representation and interpretations. Meanwhile, despite the popularity of linear factor models (e.g., Fama and French, 1993; Hou et al., 2015; Fama and French, 2015), researchers in empirical asset pricing recognize the need to apply nonlinear machine learning models to explain the cross-section variations in returns (e.g., Rossi and Timmermann, 2015; Freyberger et al., 2019; Nagel, 2021). Yet methods such as deep learning (e.g., Feng et al., 2020) appear to economists and practitioners as black boxes despite the superior pricing or prediction performances.

This study fills the gap by balancing model performance and interpretability using a new decision tree approach to split the cross-section of individual assets and address the high-dimensionality, nonlinearity, and interactions in financial data. Specifically, we develop a unified P-Tree framework ("P" stands for panel) for analyzing potentially imbalanced panel data, which allows us to generate and estimate a latent factor model. With the graphical visualization of a decision tree, P-Tree offers an alternative top-down solution to security sorting for splitting asset returns, transparently revealing feature interactions in both the time series and the cross section. Besides elucidating the stock return variation, P-Tree generates test assets and profitable trading strategies by grouping similar assets into leaf portfolios using their lagged characteristics and nonlinear interactions. With a global asset pricing criterion (no-arbitrage), P-Tree recovers the stochastic discount factor and yields a time-varying beta factor model that accurately prices individual asset returns. The methodology is broadly applicable for building trees for panel data with multi-period leaves and global split criteria as economic restrictions to guard against overfitting and improve model performance.

Our asset pricing application illustrates two methodology innovations in P-Trees. First, the nodes allow multi-period returns and more effectively extract information from panel data. An asset pricing P-Tree model resembles the conditional sorting scheme instead of the simultaneous sorting scheme (i.e., Fama-French ME - B/M 5×5 portfolios). While most security sorting procedures

do not consider cutpoints but sort assets into quintile or decile portfolios, P-Trees sequentially decide on cutpoints in a data-driven way. The leaves are a time series (in vector form) of (weighted) averages of individual asset returns within each period, constituting basis portfolios. Extant tree algorithms such as CART search the optimal cutpoints and order of splitting characteristics by minimizing the squared loss when building the decision tree, assuming independent and identically distributed (i.i.d.) observations conditional on the covariates. The conditional security sorting procedure can be viewed as a single-period tree building process since the CART algorithm predicts everything with a constant. Our innovation provides the first multi-period tree to approximate conditional security sorting in empirical asset pricing under an imbalanced panel data structure.

Second, a P-Tree grows based on a global splitting criterion guided by asset pricing theory. The split criterion in traditional CART is designed to fit the average returns rather than a factor pricing model. The CART recursive algorithms optimize at each split for the corresponding parent node without considering other sibling nodes. For example, the splitting criterion in Breiman et al. (1984) is simply the aggregated sum of squared error, which is equivalent to minimizing pricing errors when applied to asset pricing. This greedy strategy focuses on local optimization and usually leads to overfitting unless one considers an ensemble of trees. In contrast, P-Tree splits iteratively based on the asset pricing improvement estimated through the loss from a linear factor model. All assets in the cross-section are considered, including those not in the splitting parent node.

Empirically, we apply P-Trees to individual stock returns in the U.S. equity market from 1981 to 2020, with twenty years of data for training and testing. The P-Tree and market-adjusted P-Tree outperform most other models, including standard observable and latent factor models, in terms of asset pricing, return prediction, and factor investment performance for both in-sample and out-of-sample analysis. In particular, P-Tree generates a latent factor model with performance superior to most machine learning methods and are comparable to recent PCA or deep learning models (Lettau and Pelger, 2020; Kelly et al., 2019; Kim et al., 2021; Chen et al., 2020; Feng et al., 2020; Gu et al., 2021). An equally-weighted portfolio for five P-Tree factors delivers a significant alpha of 1.09% against the Fama-French 3 factors plus an annualized Sharpe ratio of 1.98 in the out-of-sample analysis. Finally, leaf basis portfolios generated by P-Tree and market-adjusted P-Tree portfolios show consistently significant performances as test assets in out-of-sample studies.

The evaluation of the out-of-bag variable importance shows only a few significant characteristics that drive cross-sectional return variation, such as volume volatility and industry-adjusted market equity. Moreover, we apply P-Tree to split the panel of return data over time series and find inflation as the most important macro predictor for regime-switching. The long-term reversal (MOM60M) and 1-year seasonality (SEAS1A) are more important for the high inflation periods, while trading volume (DOLVOL) and volume volatility (STD_DOLVOL) are more important for the low inflation periods. In addition, we apply P-Tree to visualize and investigate nonlinear and interaction factors, which helps strengthen and resurrect several anomalies.

In general, P-Tree has several advantages over other tree-based models. First, single decision trees have explicit and interpretable structures, graphically displying the feature interactions and sequential splitting. Second, they incorporate bespoke global splitting criteria such as no-arbitrage in the asset pricing example, reducing overfitting in noisy data and incorporating theoretical insights. Third, they extract greater information from panel data, while inheriting the advantage of the tree-based method that they are highly adaptive to the low signal-noise data environment and relatively short observation history. Finally, specific to the asset pricing application, the value-weighted leaf basis portfolios constitute natural test assets and imply lower transaction costs of factor construction than in PCA and deep learning models.

Literature. Our paper adds to the fast-emerging literature on machine learning in finance.[1] Our P-Tree not only splits the cross-section for creating test assets, but also provides regularized estimation to recover the stochastic discount factor. Prior attempts applying regression trees (including random forests and boosted trees) to financial forecasting asset pricing (e.g., Gerakos and Gramacy, 2012; Moritz and Zimmermann, 2016; Rossi and Timmermann, 2015; Rossi, 2018; Gu et al., 2020; Bianchi et al., 2021; Creal and Kim, 2021) typically do not tailor the tree structure for panel data with distinctions between time series and cross section. P-Trees reveal the importance of incorporating dynamic patterns in the macroeconomic time series and the interactions of macroeconomic

[1] For example, Freyberger et al. (2019), Gu et al. (2020), Cong et al. (2021) find equity returns predictable using machine learning and AI models. Bianchi et al. (2021) and Feng et al. (2020) find positive results on treasury bond returns, while Bali et al. (2021) and He et al. (2021) find positive results on corporate bond returns. For direct portfolio construction, Cong et al. (2020) provide a reinforcement learning approach, while Feng and He (2019) and DeMiguel et al. (2020) provide regularization methods by optimization or Bayesian modeling. Feng et al. (2020), Dong et al. (2021) and Bryzgalova et al. (2021) apply variable selection methods to the zoo of factors.

time series and asset characteristics.

The standard CART or its ensemble methods do not impose global criteria for growing the tree either.[2] We show that global criteria for splitting (which corresponds to no-arbitrage constraint) produce betters results for asset pricing than other tree-based ML approaches focusing on prediction. Consequently, we add to the emerging studies imposing economic restrictions for estimating or evaluating machine learning (e.g. Feng et al., 2020; Chen et al., 2020; Avramov et al., 2021). In particular, while no-arbitrage conditions have been used to estimate simple parametric models, it is equally important to require such asset pricing criteria (instead of the popular prediction objectives) when estimating large and flexible machine learning models to avoid overfitting under low signal-to-noise financial data and to improve model performance.

One recent attempt along this line is Creal and Kim (2021) which analyzes all characteristics at once and identify variables that best explain assets' betas. The authors split currency-return observations and implements a Bayesian factor model for currencies with regression tree priors, cleverly using asset pricing objectives in the likelihood estimation and resampling process to produce stochastic posterior trees. P-Trees similarly use asset pricing criteria (no-arbitrage) to guide tree grows, but we differ by (i) considering multi-period returns in the nodes to further extract information from panel data, (ii) growing more interpretable, deterministic trees that explicitly depict feature interactions in addition to assisting variable selection, with prescriptions for managing portfolios and developing new trading strategies in practice, (iii) estimating a latent factor model rather than using MCMC sampling over observable factors, and (iv) focusing the empirical analysis on equity markets and relating to previous tree-based studies for the same markets.

We also contribute to the growing literature in empirical asset pricing that develops latent factor models for pricing cross-sectional returns and generating risk-adjusted returns. Lettau and Pelger (2020), Kelly et al. (2019) and Kim et al. (2021) show different principal component methods to produce latent factors, while Chen et al. (2020), Feng et al. (2020) and Gu et al. (2021) generate the SDF through various deep learning models. In our empirical analysis, we show that several comparative advantages lead to the superior performance: The P-Tree provides a characteristics-driven partition for the cross-section of the stock universe. The generated leaf basis portfolios are

[2]Ensemble methods include random forests (Breiman, 2001), boosting trees, XGBoost (Chen and Guestrin, 2016), Bayesian additive regression trees (Chipman et al., 2010), and the recent XBART (He et al., 2019; He and Hahn, 2021).

value-weighted, and so the transaction cost of P-Tree factors is much lower than principal component portfolios. Section 3.5 further demonstrates how the interpretability of P-Trees allows trading strategies based on nonlinearity and sequential interactions in the feature space.

Finally, this paper is related to conditional asset pricing in general and the construction of basis assets. We follow Avramov (2004), Gagliardini et al. (2016), and Feng and He (2019) to allow time-varying, characteristic-driven factor loadings for individual stock returns. Ahn et al. (2009) use cluster analysis to create basis assets; we add by documenting clustering patterns and investment performance for leaf basis portfolios from both shallow and deep trees. Bryzgalova et al. (2020) generate kernel-spanning test assets by pruning an exogenously given tree with leaves from deterministic sorting; we focus on data-driven top-down tree growth and offer interpretability through explicitly showing both leaves and branches — something more than just the posterior tree distribution in Creal and Kim (2021) and the collection of leaves after pruning in Bryzgalova et al. (2020)). The literature, including regularized linear models of Kozak et al. (2020) and its generalization in Bryzgalova et al. (2020) and the RP-PCA rotation in Lettau and Pelger (2020), largely takes pre-specified basis functions (sorted portfolios) as given. We add by generating the basis functions, test portfolios, and the non-linear transformation of explanatory variables in a data-driven way using trees, complementing the approach in Chen et al. (2020) using adversarial networks, among others.

2 P-Tree Factor Model for Asset Pricing

2.1 Classification and Regression Trees

A tree is a set of split rules (or cutpoints) defining a rectangular partition of the covariate space. A cutpoint $\tilde{c} = (x_i, c_i)$ indicates splitting the data along variable x_i using a threshold c_i. The essence for decision trees is thus choosing cutpoints to grow the tree. Tree-based models are easy to understand, handles both numerical and categorical variables, requires relatively less data for training, and performs well for big data problems. CART (classification and regression trees Breiman et al., 1984) and its variants are among the popular algorithms in various tree-based applications such as clinical study (e.g., Arnett et al., 1988), ecology (e.g., Guisan and Zimmermann, 2000), medical diagnosis (e.g., Lemon et al., 2003), and more recently, asset pricing (e.g,. Rossi and Timmermann,

2015; De Prado, 2018; Rossi, 2018; Li and Rossi, 2020).

Typical tree models such as CART grow by recursive partition, optimizing over all cutpoint candidates using a split criterion. When considering split a node, the recursive algorithm only processes with data within that specific node — a greedy approach preferred mainly for easy coding and efficient computation. Take an application of CART to asset pricing for example. Let $r_{i,t}$ denote the return of asset i at time period t. With a squared loss split criterion, which treats the cross-sectional data as pool data, a cutpoint candidate is evaluated based on:

$$\sum_{i,t\in\text{left node}} (r_{i,t} - \bar{r}_{\text{left}})^2 + \sum_{i,t\in\text{right node}} (r_{i,t} - \bar{r}_{\text{right}})^2$$

where $\bar{r}_{\text{left}} = \frac{1}{\#_{\text{left node}}} \sum_{i,t\in\text{left node}} r_{i,t}$ and $\bar{r}_{\text{right}} = \frac{1}{\#_{\text{right node}}} \sum_{i,t\in\text{right node}} r_{i,t}$ are average returns of data in the left or right leaves respectively. Both the left and right leaves are constants over not only the cross section but also the time-series (the constant does not have subscript for time nor stock), which serve as pricing kernels for the corresponding individual stocks in a one-period model for scalar outcomes.

Such tree models leave much to be desired because we want to form a *basis portfolio* — a vector representing returns over multiple periods — at each leaf node instead of a constant. Moreover, a tree-based factor pricing model aiming to recover the SDF from basis portfolios at the leaves should satisfy no-arbitrage conditions. We therefore need to specify a tree that can accommodate the panel data and adopt, rather than a local recusive split criterion, a global one evaluated at all leaf portfolios according to their joint asset pricing efficacy.

Note that the two properties, (i) leaves containing a factor model instead of a constant and (ii) having a global split criterion is not restricted to the applications in asset pricing, although we focus on asset pricing to illustrate the advantages they bring. In fact, most of the extant regression or classification trees do not allow us to fully extract the dynamic information in the panel data. With noisy data, trees also easily overfit, unless one resort to ensemble methods that sacrifice a single tree's interpretability. Our global split criteria for guiding tree growth provides an alternative to guard against overfitting while maintaining interpretability. We next develop a class of tree models for panel data with global split criteria, which we term P-Trees ("P" stands for "panel"),

and illustrate their effectiveness through an application in asset pricing with conditional pricing kernels.

2.2 Conditional SDF Model Using P-Tree

A conditional stochastic discount factor (SDF) model that explains cross-sectional asset price returns satisfies:

$$E_t\left[m_{t+1}r_{i,t+1}\right]=0 \Longleftrightarrow E_t\left[r_{i,t+1}\right]=\underbrace{\frac{\mathrm{Cov}_t\left(m_{t+1},r_{i,t+1}\right)}{\mathrm{Var}_t\left(m_{t+1}\right)}}_{\beta_{i,t}}\underbrace{\left(-\frac{\mathrm{Var}_t\left(m_{t+1}\right)}{E_t\left[m_{t+1}\right]}\right)}_{\lambda_t}, \tag{1}$$

where $r_{i,t+1}$ is the excess return of individual asset i at time $t+1$. $\beta_{i,t}$ and λ_t are the time-t market expected SDF exposure of stock i and risk price of the SDF. A natural solution for m_{t+1}, is its projection onto the return space of the assets, i.e.,

$$m_{t+1}=1-w_t^{\intercal}r_{t+1}, \tag{2}$$

where $r_{t+1}=(r_{1,t+1},\cdots,r_{N,t+1})^{\intercal}$ denotes returns of all assets at time period $t+1$, and w_t is a vector of portfolio weights. Substituting (2) into (1) yields the SDF portfolio weights:

$$w_t=E_t\left[r_{t+1}r_{t+1}^{\intercal}\right]^{-1}E_t\left[r_{t+1}\right]. \tag{3}$$

The extant literature does not estimate w_t from a large number of individual stock returns mainly due to the large dimension of the covariance term.

Following recent studies on latent factors (e.g., Lettau and Pelger, 2020; Kelly et al., 2019; Feng et al., 2020), we use basis portfolios from a P-Tree (as opposed to individual stocks) to overcome the estimation challenge. But unlike existing studies that pre-specify the basis portfolios, our P-Tree algorithm splits the cross section of asset returns into basis portfolios in a data-driven manner while imposing a global no-arbitrage criterion, in stark contrast with the standard CART criterion that is recursive and local to specific leaves. Consequently, our tree grows iteratively to generate both the leaf basis portfolio and the latent factor (SDF).

Our P-Tree model splits the stock universe into non-overlapping basis portfolios by cross-sectional quantiles of past firm characteristics (and possibly by time series macroeconomic variables). The number of basis portfolios increases one at a time when the tree splits a parent node into two child nodes. At the k-th split of the tree growing process, we denote the SDF $m_{t+1}^{(k)}$ generated by basis portfolios $R_{t+1}^{(k)}$ as:

$$m_{t+1}^{(k)} = 1 - W_t^{(k)} R_{t+1}^{(k)}, \tag{4}$$

$$W_t^{(k)} = E_t\left[R_{t+1}^{(k)} R_{t+1}^{(k)\intercal}\right]^{-1} E_t\left[R_{t+1}^{(k)}\right]. \tag{5}$$

For the conditional factor model, the time-varying factor exposure $\beta_{i,t}$ is given by:

$$\beta_{i,t}^{(k)} = \frac{\mathrm{Cov}_t\left(W_t^{(k)} R_{t+1}^{(k)}, r_{i,t+1}\right)}{\mathrm{Var}_t\left(W_t^{(k)} R_{t+1}^{(k)}\right)}. \tag{6}$$

After the k-th split of the tree growing process, we model the time-varying factor exposures in reduced-form in terms of past firm characteristic updates, which is common in the literature (e.g., Avramov, 2004; Kelly et al., 2019; Feng and He, 2019). In other words,

$$\beta_{i,t}^{(k)} = b_0^{(k)} + b_1^{(k)\intercal} z_{i,t}, \tag{7}$$

where $z_{i,t}$ are firm characteristics such as market equities or book-to-market ratios.

2.3 Growing a P-Tree Under No-Arbitrage

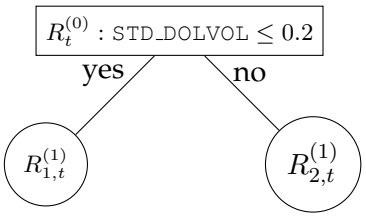
A P-Tree at birth corresponds to CAPM: a single leaf basis portfolio with returns of all assets. We gradually build the tree with additional splits by updating $\{R_{t+1}^{(k)}, W_t^{(k)}, \beta_{i,t}^{(k)}\}$ through an iterative scheme: First of all, leaf basis portfolios, $R_{t+1}^{(k)}$, are expanded by the iterative tree growth. Second, the latent factor weights, $W_t^{(k)}$, are estimated by the expanded leaf basis portfolios. Finally, time-varying factor exposures, $\beta_{i,t}^{(k)}$, are updated by the new characteristics data and updated SDF when fitting the cross-section.

Algorithm 1 summarizes the tree growth in pseudo-codes. We refer to the final nodes as terminal or leaf nodes and the intermediate nodes as internal nodes. Let $R_t^{(k)}$ denote the return of the

basis portfolios after the k-th leaf node of the tree, which can be equal or value weighted portfolio of the assets in that specific leaf. The number of basis portfolios increase by one after each iterative split. So, there are $k+1$ basis portfolios after the k-th split, and $R_t^{(k)} = [R_{1,t}^{(k)}, R_{1,t}^{(k)}, \cdots . R_{k+1,t}^{(k)}]^\intercal$.

Let us consider a baseline example using firm characteristics only. At the start, the entire cross-section of stock returns are in the top node (*root*) of the tree, $R_t^{(0)}$. We consider several cutpoints candidates (pairs of firm characteristic and threshold value to split the returns). All firm characteristics are normalized cross-sectionally in the range from -1 to 1. Instead of searching for hundreds of cutpoints, we only consider the cross-sectional quintiles such as -0.6, -0.2, 0.2, and 0.6. This choice follows the conventional security sorting and simplifies computation.

Figure 1: Demonstration of the first split.



Each cutpoint candidate partitions the root node to left and right child nodes comprised of a subset of assets and observation months of the data. Each potential leaf can form a leaf basis portfolio, denoted by $R_{1,t}^{(1)}$ and $R_{2,t}^{(1)}$. Since there are only two leaf nodes, the latent factor (SDF) is estimated as a mean-variance efficient portfolio of the two leaf basis portfolios,

$$f_t^{(1)} = w^{(1)} R_t^{(1)}, \quad w^{(1)} = \widehat{\Sigma}_1^{-1} \widehat{\mu}_1 \tag{8}$$

where $\widehat{\Sigma}_1^{-1}$ and $\widehat{\mu}_1$ are the covariance matrix and average returns of two leaf portfolios $R_t^{(1)} = [R_{1,t}^{(1)}, R_{2,t}^{(1)}]^\intercal$.[3] We use a latent factor model to define the loss function when evaluating cutpoint

[3]On a separate note, to deal with the estimation error in the sample mean and sample covariance matrix estimation for the mean-variance efficient portfolio, we add two small regularization parameters $\lambda_\Sigma = 10^{-4}$ and $\lambda_\mu = 10^{-4}$ in Equation 5. These two shrinkage parameters help to stabilize the portfolio weight estimation and avoid over-leveraging. The similar regularized portfolio optimization problem is also addressed in Kozak et al. (2020) Equation (18) and (22), and Bryzgalova et al. (2020).

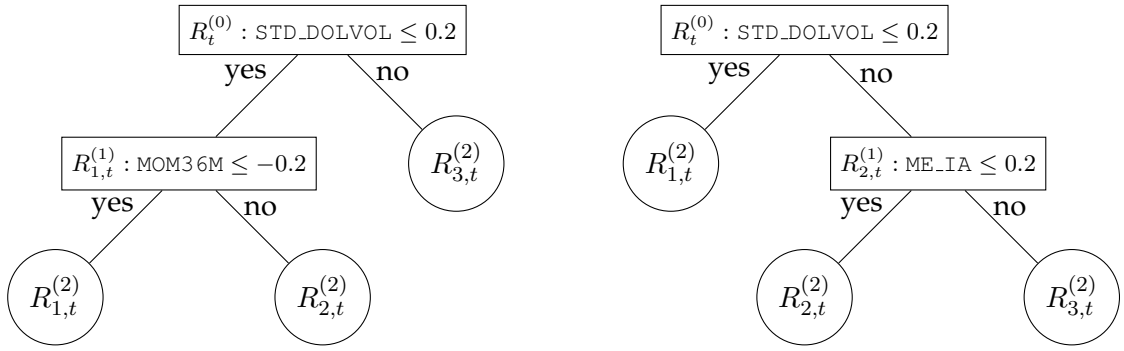
$$W_t^{(k)} = \left[E_t\left(R_{t+1}^{(k)} R_{t+1}^{(k)\intercal}\right) + \lambda_\Sigma I_{k+1}\right]^{-1} \left[E_t\left(R_{t+1}^{(k)}\right) + \lambda_\mu \mathbf{1}\right]. \tag{9}$$

candidate c_k:

$$\mathcal{L}(c_k) = \sum_{t=1}^{T} \sum_{i=1}^{N_t} \left(r_{i,t} - \beta(z_{i,t-1}) f_t\right)^2, \tag{10}$$

where $\beta(z_{i,t-1}) = b_0 + b^\intercal z_{i,t-1}$ are conditional factor loadings on the past firm characteristics. We estimate the regression coefficients by the ordinary least squares estimator, using a pooled regression model for individual stock returns on f_t and $f_t \times z_{i,t-1}$ without an intercept. (10) is evaluated for all cutpoint candidates $c_k \in \mathcal{C}$, and its minimizer is picked as the first cutpoint.

Figure 2: Demonstration of the cutpoint candidates for the second split.



(a) If split node $R_{1,t}^{(1)}$ at MOM3 6M.

(b) If split node $R_{2,t}^{(1)}$ at ME_IA.

Next, we proceed to the second cutpoint. Note that there exist two leaf nodes after the first split. The second split could happen at either the left or right child node of the root. We take an iterative approach to grow the tree. We evaluate the split criteria for all cutpoint candidates at two leaf nodes and pick the one minimizing the loss function. Figure 2 depicts the tree of the cutpoint candidates for the second split. In both cases, one leaf node is split, becomes an internal node, and creates two new leaf nodes. The SDF will be evaluated based on the three basis portfolios.

$$f_t^{(2)} = w^{(2)} R_t^{(2)}, \quad w^{(2)} = \widehat{\Sigma}_2^{-1} \widehat{\mu}_2 \tag{11}$$

where $R_t^{(2)} = [R_{1,t}^{(2)}, R_{2,t}^{(2)}, R_{3,t}^{(2)}]$. The graphical position of the three basis portfolios depends on the which node to split, as shown in Figure 2. Similary, $\widehat{\Sigma}_2^{-1}, \widehat{\mu}_2$ are covariance matrix and average return of the basis portfolios $R_{2,t}$.

The next cutpoint is chosen similarly. For the k-th split, each cutpoint candidate of the existing

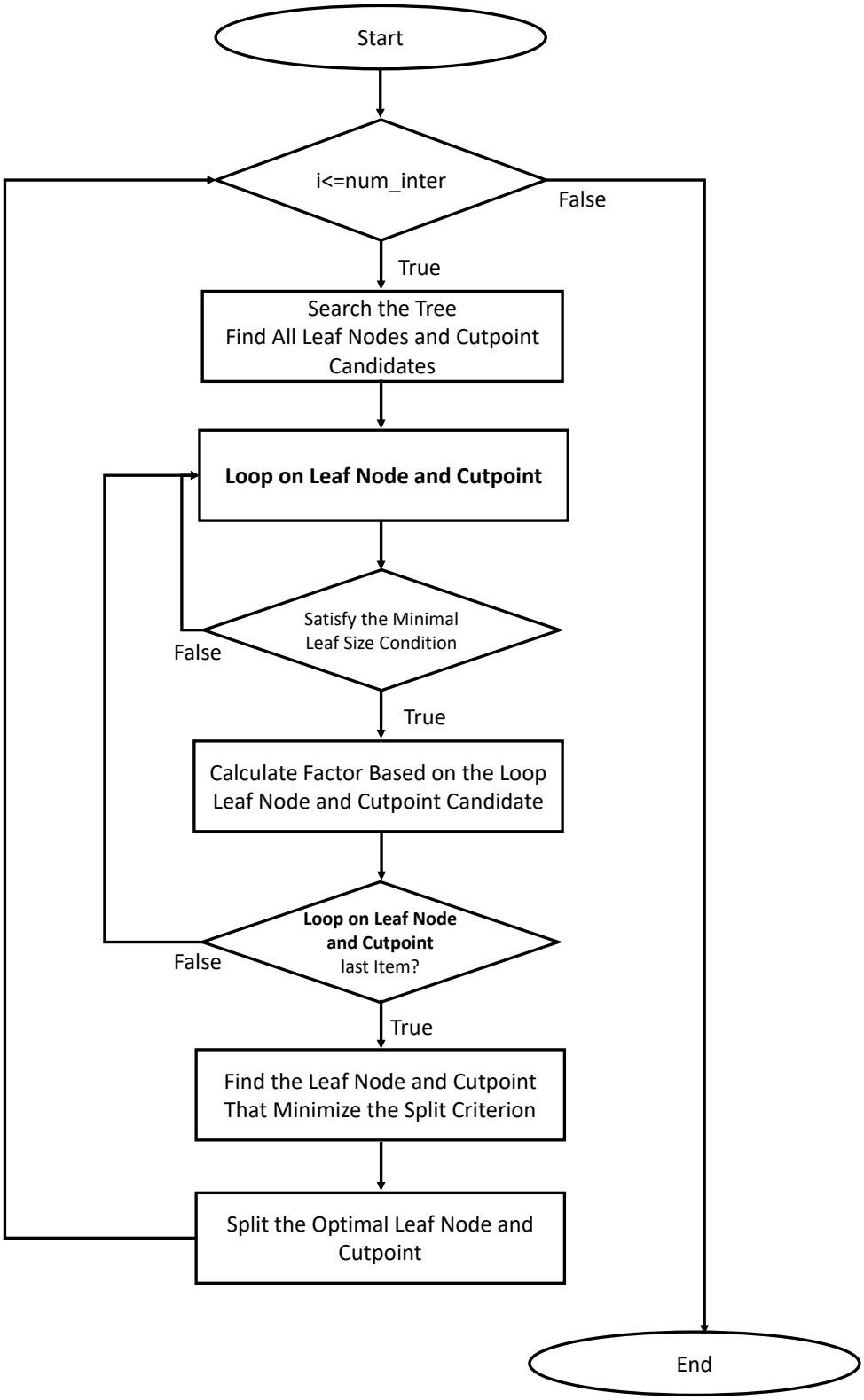
k leaf nodes creates $k+1$ leaf basis portfolio, which adds to the generation of the latent factor $f_t^{(k)}$, and hereby to evaluate the split criterion of Equation 10. The stopping conditions are pre-specified as the total number of iterations, max depth of the tree, or the minimum number of data observations in a leaf node, whichever is met first. Algorithm 1 and Figure 3 summarize the procedure.

The P-Tree model differs from the standard tree models such as CART in several important ways. First, each leaf of CART is associated with a constant (leaf parameter), which predicts which new data fall in that leaf. CART aims to approximate a function by step function represented by the tree. In contrast, the P-Tree has a clear economic objective. Each tree partitions the entire cross-section of stock returns to multiple leaf basis portfolios and creates the latent factor. Second, unlike CART for which splits are local and greedy, our P-Tree grows iteratively by examining *all* leaf basis portfolios. The split criterion is defined globally as the pricing loss of the generated latent factor for the returns of *all* assets, regardless of whether the data are in the current node or not.

Algorithm 1 The main algorithm that grows the asset pricing tree from the data.

procedure GROWTREE($\mathrm{y}, \mathbf{X}, \Phi, \Psi, d, T,$ `node`)
outcome Grow the tree T and find corresponding basis portfolios
  for i from 1 to num_iter **do** ▷ Loop over number of iterations
    if the stopping conditions are met **then**
      return.
    else
      Search the tree, find all leaf nodes $\mathcal{A}$
      for each leaf node A_j in $\mathcal{A}$ **do** ▷ Loop over all current leaf nodes
        for each cutpoint candidate c_k in $\mathcal{C}$ **do** ▷ Evaluate all cutpoint candidates
          Partition data in A_j according to c_k.
          if Either left or right child of A_j does not satisfy minimal leaf size **then**
            Ignore this cutpoint candidate.
          else
            Find the latent factor (SDF) based on all leaf portfolios as in equation (8).
            Calculate the split criterion in equation (10).
          end if
        end for
      end for
      Find the leaf node to split and cutpoint c_k that minimizes the split criterion (10).
      Split the node selected at cutpoint c_k. ▷ Split one node at one iteration
    end if
  end for
  return
end procedure

Figure 3: Flowchat of the P-Tree Algorithm



2.4 Nonlinearity, Interaction, and Interpretability

Observing interaction in P-Tree. Tree-based models are known to work well in capturing non-linearities and interactions in the data. Unlike deep learning models in which the nonlinearities and interactions concerning specific features are unobserved, the structure of a tree transparently displays which features interact and how. Despite this natural interpretability, a single tree tend to overfit noisy data, and ensembles such as random forests are often used to improve a model's out-of-sample performance at the expense of interpretability.

Imposing global asset pricing conditions help reduce the overfit, as we see in Bryzgalova et al. (2020). However, Bryzgalova et al. (2020) solve a regularized portfolio optimization problem and prune final leaves without explicitly deriving the tree growth. P-Tree displays the complete tree structure in a top-down manner while requiring global no-arbitrage. The output is a single tree that does not overfit the data and performs well for pricing assets, all while revealing the exact interactions of the input features in each layer of the tree growth.

As we show in our empirical analysis later, data-driven P-Trees allow us to observe feature interactions and use the information to significantly improve over trading strategies, which are based on uni-variate sorting without considering interactions or multiple sorts with exogenously-specified sequence and variables. They also provide information for recovering the SDF and building a factor pricing model that performs as well as, if not better, than the best competing models.

More generally, a global slit criterion can be used to guard against overfitting of a single tree and consequently presents directly empirical evidence for the nonlinearity and interaction in the data — a task other machine learning approaches or ensemble methods fall short of.

Factor and portfolio constructions. P-Trees provide an alternative solution to portfolio construction and strategy development that utilizes the information on feature interaction and non-linearity. For example, a two-layer P-Tree with two leaves splits the cross-section and creates two left and right leaf portfolios. Guided by economic theory, one can construct a univariate factor to long and short these two leaf basis portfolios. In practice, researchers typically use quintile or decile sorted portfolios to create long-short portfolios (top-bottom or bottom-top) to generate higher return spreads between long and short legs. Researchers might also apply bivariate or triple-way sort

to include the interaction of the characteristics when creating long-short portfolios.[4]

A P-Tree searches the optimal characteristics for interaction for the second and probably third splits when growing the tree with the asset pricing goal. First, growing a tree model with three layers makes it possible to have different interactions on the long and short legs from the first split. Second, by partitioning the stock universe, one can graphically display the nonlinear characteristics-return structures for the long and short legs. Third, the splitting order and cutpoints are data-driven and informs of the sequence of sorting and the interaction of characteristics.

Variable importance and P-Tree robustness. Rather than growing a single tree, ensemble methods grow multiple trees either independently or dependently. Random forest and boosting are two popular tree ensemble methods (Friedman et al., 2001). Plain-vanilla random forests (Breiman, 2001) grow a number of trees on bootstrapped training samples, which come from random samples of the observations as well as variables from the complete training set. The forecast of the entire ensemble is the average of all "decorrelated" trees, which usually has a smaller variance than that of a single tree. Although we do not use random forests to improve P-Tree performance, we can use it to identify important variables in creating latent factors as a robustness check for P-Tree models.

To this end, we apply the same bootstrap procedure in Random Forests to P-Trees, train various "decorrelated" P-Trees on bootstrapped subsamples. To bootstrap panel data of individual stock returns, we first take subsample in the time horizon, and for the selected period, the complete cross-section of the panel data is preserved. Second, we randomly draw ten characteristics out of the 61 for building each tree. The procedure repeats 500 times to form a forest (500 P-Trees).

In the empirical exercises, we focus on the out-of-bag (OOB) variable importance and define two importance metrics. The first measurement of importance entails a variable's frequency of being selected for splits. Intuitively, the more often a characteristic is selected as tree cutpoints, the more important it is. For each bootstrapped subsample, a subset of characteristics is randomly drawn to build the tree, and only a fraction of them are actually used to split. We count the number of times a particular characteristic z_i being used in the first K splits and the total number of

[4]Different characteristics-sorted univariate, bivariate, or triple-way portfolios are available on Ken French's website. Commonly used long-short factors, such as Fama-French five factors, are linear combinations for these portfolios restricted by the economic direction.

appearances in bootstrapped subsamples. The measurement of importance is defined as:

$$\text{Selection Probability}(z_i) \quad = \quad \frac{\#(z_i \text{ is selected at first K splits})}{\#(z_i \text{ appears in the bootstrapped subsamples})} \tag{12}$$

The second importance measurement aims to capture a "treatment effect." For the 500 trees grown on bootstrapped subsamples, we force half of them to use a characteristic but the other half not. Note that even if one characteristic is included in the subsample, it is not guaranteed to be selected as cutpoints when building the tree. The with- and without- sampling scheme for a particular characteristic creates the treatment effect evaluation and allows one to perform the significance test of its importance. We compute the asset pricing model fitting performance, the total R^2 (Equation 18), for this treatment effect evaluation.

$$\text{Char. Importance} \quad = \quad \left[\frac{E(\text{P-tree fit} \mid \text{with char}_i)}{E(\text{P-tree fit} \mid \text{without char}_i)} - 1\right] \times 100 \tag{13}$$

We can then compare the variables ranking high in importance in the random forest of P-Trees to those first selected as cutpoints in a single P-Tree model. If they are similar, the P-Tree model with global split criteria indeed behaves similarly as the less interpretable random forests that typically achieve better out-of-sample performance.

2.5 Splitting the Time-Series and Regime Switching

The previous sections focus on the cross-section splits, i.e., all nodes in the tree split at the firm characteristics $z_{i,t}$. As the tree grows from top to bottom, the panel of stocks is partitioned into many leaf nodes, and each leaf node maintains the entire time series from period 1 to T. Intuitively, the latent factors and/or factor loading functions may substantially differ under different macroeconomic states (e.g., high and low inflation or interest rate states). There is a long literature discussing regime changes in financial markets, involving, e.g., periods of high and low volatility, or booms and recessions (e.g., Maheu and McCurdy, 2000; Ang and Timmermann, 2012).

In a P-Tree with only cross-sectional splits, although the factor loadings are time-varying based on the firm characteristics, the latent factor generation is not. A natural solution is to create different factor models and estimate corresponding loadings under different macroeconomic states. With a

panel data of individual stock returns, the cutpoints of some nodes (for example, the root node) can obviously involve macroeconomic variables instead of the firm characteristics. In other words, P-Trees are flexible enough to incorporate time-series splits in addition to cross-section splits.

Given the short time-series observations of individual stock returns relative to the large cross-section, we only implement one time-series split of the data at the root node (i.e., the first cutpoint of the tree) for illustration. The P-Tree model will search over all possible cutpoint candidates of the macro variables and choose the one optimizes the factor pricing error. The split criterion of the first time-series split is defined as

$$\mathcal{L}(c_k) = \sum_{t=1}^{N_A} \sum_{i=1}^{N_t} (r_{i,t} - \beta_A(z_{i,t-1}) f_{A,t})^2 + \sum_{t=1}^{N_B} \sum_{i=1}^{N_t} (r_{i,t} - \beta_B(z_{i,t-1}) f_{B,t})^2 , \tag{14}$$

where the cutpoint candidate c_k partitions the time series of data to state A and B, for example, high or low inflation.[5] Each state has number of months N_A and N_B correspondingly. Comparing to the cross-section split criterion in equation (10), the time-series split criterion is the *total* pricing loss of two time periods, with two corresponding factors $f_{A,t}$ and $f_{B,t}$.

In this illustration, after searching for the optimal time series cutpoint, all subsequent growth only evaluates the cross-section cutpoints. Note that any further split on either child of the root node only depends on stock returns observations on one side. The split criterion is no longer iterative to the first split but still works for the entire time series on the one side. with both time-series and cross-section candidates for splits, the model can also inform how macroeconomic variables interact with firm characteristics.

2.6 Factor Pricing and Boosted P-Trees

Because of the ability to accommodate panel data and build conditional leaves, as well as the global criteria for greater performance and interpretability, P-Trees offer effective ways of building a (multi-)factor pricing model. In the baseline model, we already see that a P-Tree can generate the single-factor SDF. The multi-period nature of leaf returns allows P-Trees to better utilize panel data of asset returns for estimating conditional factor loadings, a step required when imposing the

[5] Our empirical analysis in Section 3 shows that Inflation is indeed the most important macroeconomic variable that differentiates factor models in two states.

no-arbitrage condition. Using no-arbitrage as a global split criterion in turn ensures good out-of-sample performance of the SDF derived from the P-Tree and the usefulness of the factor model. The two innovations in the P-Tree framework together makes it well-suited for asset pricing.

To strengthen the asset pricing performance, we can further use boosting to obtain a multi-factor P-Tree model. The idea is to combine a group of weak learners for better fitting and prediction (Freund and Schapire, 1997). The boosting ensemble method grows a list of trees iteratively where each tree fits the residual of all previous trees. They together form a strong learner, and the final prediction is the (weighted) sum of each tree outcome. For finance applications, Rossi and Timmermann (2015) use boosted regression trees to estimate conditional covariance for ICAPM.

We use the same strategy to build the sum of the P-Trees with each P-Tree giving a factor. Specifically, the boosting P-Tree iteratively creates additional factors to fit the unexplained pricing errors of all previous factors. Applying the boosting scheme to P-Tree establishes a sequence of factors to form the SDF, resulting in a multi-factor asset pricing model. The steps are as follows:

1. The first factor $f_{1,t}$ is generated by the standard P-Tree on excess returns $\{r_{i,t}\}$, as discussed in section 2.3. The residual of the first factor is $\epsilon_{i,t}^{(1)} = r_{i,t} - \beta(z_{i,t-1})f_{1,t}$, where the regression coefficients are the OLS estimator.

2. The second factor generated from fitting the residual $\epsilon_{i,t}^{(1)}$ instead of the original return $r_{i,t}$. The tree growing steps are the same as section 2.3 except replacing the split criterion of equation (10) by the following one

$$\mathcal{L}(c_k) = \sum_{t=1}^{T}\sum_{i=1}^{N_t}\left(\epsilon_{i,t}^{(1)} - \beta_2(z_{i,t-1})f_{2,t}\right)^2. \tag{15}$$

   Again, the new residual is defined as $\epsilon_{i,t}^{(2)} = \epsilon_{i,t}^{(1)} - \beta_2(z_{i,t-1})f_{2,t}$.

3. The third factor $f_{3,t}$ proceeds similarly to fit the residual of the first two trees $\{\epsilon_{i,t}^{(2)}\}$, with the corresponding split criterion

$$\mathcal{L}(c_k) = \sum_{t=1}^{T}\sum_{i=1}^{N_t}\left(\epsilon_{i,t}^{(2)} - \beta_3(z_{i,t-1})f_{3,t}\right)^2. \tag{16}$$

4. Repeating the procedures above K times, each tree fits the residual of all previous trees to generate all K factors.

In Section 3.3 when we apply the framework to U.S. equities, we set the number of factors $K = 3$. The three generated factors can be listed with decreasing importance order as $[f_{1,t}, f_{2,t}, f_{3,t}]$. Generating additional tree factors is similar to generating additional components in the principal component analysis. Below is the three-factor model with time-varying factor loadings. Any general application of this three-factor model requires re-estimating factor loadings.

$$E(r_{i,t}) = \widehat{\beta}_1(z_{i,t-1})f_{1,t} + \widehat{\beta}_2(z_{i,t-1})f_{2,t} + \widehat{\beta}_3(z_{i,t-1})f_{3,t} \tag{17}$$

Another natural application of the boosting is to create an augmented factor model. One can start with a common benchmark model such as CAPM or Fama-French 3-factor model. Then use the boosting P-Tree to fit the benchmark model's pricing errors directly. Tree factors developed beyond the benchmark model are supposed to explain "orthogonal" information from factor model residuals. We later report results of multi-factor P-Trees with and without the CAPM benchmark.

3 Empirical Application: A Study of U.S. Equities

3.1 Data

We apply the standard filters (e.g., used in Fama-French factor construction) to the universe of U.S. public equities: (1) We include only stocks listed on NYSE, AMEX, or NASDAQ for more than one year; (2) we use those observations for firms with a CRSP share code of 10 or 11; (3) we exclude stocks with negative book equity or negative lag market equity. For each stock, we use 61 firm characteristics with monthly observations, covering six major categories: momentum, value, investment, profitability, frictions (or size), and intangibles. Table A.1 lists these input variables whose computation follows Feng et al. (2020). To construct regression trees for multiple periods of data, we standardize the firm characteristics cross-sectionally in the range $[-1, 1]$.[6]

[6]For example, the market equity in 2019 December is uniformly standardized in the range of $[-1, 1]$. The firm with the lowest market equity is -1, and the firm with the highest market equity is 1, all others distributed uniformly in between. Therefore, this uniform standardization transforms the data onto $[-1, 1]$ every month. If a firm has missing values for some characteristics, the imputed values are 0, implying the firm is not important in security sorting.

In addition, we use ten macroeconomic variables as in Feng and He (2019) to demonstrate time-series splits. Table A.2 summarizes the macroeconomic variables, which includes market timing macro predictors, bond market predictors, and aggregate characteristics for S&P 500. We standardize these macro predictor data by the historical percentile numbers for the past ten years. For example, inflation greater than 0.7 implies that the current inflation level is higher than 70% of observations during the past decade. This rolling window data standardization is useful when we compare the predictor level to detect different macroeconomic regimes.

Our sample period goes from January 1981 to December 2020. We use the 20-year period from 1981 to 2000 for training and the 20-year period from 2001 to 2020 for testing. The average and median monthly numbers of stocks are 4,667 and 4,450 in the train sample, and 3,953 and 3,791 in the test sample. Our P-Tree is flexible to handle such an imbalanced data panel, while it is difficult to implement standard observable or latent factor models on individual stock return data.

3.2 Implementation of P-Trees

We apply P-Tree to generate leaf basis portfolios and the stochastic discount factor to fit cross-sectional individual stock returns. We plot our baseline P-Tree in Figure 4. The tree structure shows the splitting orders S#, selected splitting characteristics, and the cross-sectional cutpoints. Before the first split, the tree grows from the root node, representing the market portfolio for all stocks (N1). The leaf basis portfolios at each layer of the tree are numbered with N#. The numbers printed in the leaves are the median number of stock observations in the monthly updated leaf basis portfolios. We require the minimum number of stocks in a leaf basis portfolio, one natural tuning parameter to limit the tree size and avoid over-fitting, to be 50, which is reasonable for fitting a multi-period tree model. This implies that in our training sample, the tree stops growing after 22 splits and generates 23 leaf basis portfolios. The number of leaf basis portfolios is similar to the commonly used ME-B/M 25 portfolios, so we can compare performance values in the same variance scale.

As seen in Figure 4, endogenously the P-Tree first splits along the trading volume volatility (`STD_DOLVOL`) at 0.2 (60% quantile). After the split, 60% of the stocks go to the left leaf (labeled N2), and 40% go to the right (N3). The second split is along the industry-adjusted market equity (`ME_IA`) at 0.2 on the right leaf (N3), while the third split is on 3-year long-term reversal (`MOM36M`) on the left

one (N2). The P-Tree allows one to evaluate and visualize the interaction of characteristics. ME_IA is valuable for high STD_DOLVOL stocks, while MOM36M is valuable for low STD_DOLVOL stocks.

Relative to other machine learning methods, P-Tree provides a more precise mapping for generating these characteristic-managed portfolios. Figure 5 reports the corresponding partitions, together with the average returns and Sharpe ratios for leaf basis portfolios. These performance gaps show the usefulness of splitting the cross-section via the interaction of characteristics. The top right plot in Figure 5 further splits N2 by STD_DOLVOL and MOM36M. The bottom plot splits N6 by ATO.

We can also see the clustering patterns for these leaf basis portfolios. The market portfolio is partitioned into four portfolios (N4 - N7) at the tree depth of three, each group of individual stocks with similar characteristic exposures. The return-risk relationship of the four portfolios, as well as their mean-variance efficient portfolio (MVE3), is plotted in Figure 6. Further splitting these four portfolios, we have 23 leaf basis portfolios at the tree depth 6. We find those great-grandchild leaf basis portfolios (in a light color) labeled with (N4+ - N7+) cluster around their great grandparent nodes (in dark color). The mean-variance efficient portfolio from a 6-depth tree produces a higher Sharpe ratio than a 3-depth tree, while both are higher than the market portfolio. These performances are robust for the out-of-sample plot in Figure 7.

Complementing our main result P-Tree in Figure 4, we also provide a market-adjusted P-Tree shown in Figure A.1. Section 2.6 shows how we create a P-Tree by controlling a benchmark factor model, such as CAPM or Fama-French three factors. By controlling the market factor, we find the market-adjusted P-Tree structure is quite different from P-Tree and offers 18 basis portfolios. This set of market-adjusted test assets show quite different performance in Table 1 and 3.

3.3 Asset Pricing Performance

Performance metrics. We follow Feng et al. (2020) to include multiple performance metrics. Total R^2 are Pricing Error R^2 are designed for evaluating economic asset pricing performance for variation in both time-series and cross-section. In contrast, Predictive R^2 is designed for measuring statistical model fitness and the model forecasting power.

1.
$$\text{Total } R^2 = 1 - \frac{\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\left(r_{i,t} - \widehat{r}_{i,t}\right)^2}{\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} r_{i,t}^2}, \tag{18}$$

where $\widehat{r}_{i,t} = \hat{\beta}_{i,t}^{\intercal} f_t$. Total R^2 represents the fraction of realized return variation explained by the factor model-implied contemporaneous return, aggregated over all assets and all periods.

2.
$$\text{Predictive } R^2 = 1 - \frac{\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\left(r_{i,t} - \widehat{r}_{i,t}\right)^2}{\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} r_{i,t}^2}, \tag{19}$$

where $\widehat{r}_{i,t} = \hat{\beta}_{i,t}^{\intercal}\lambda_f$. The risk premia λ_f can be the historical average of the tradable factor returns. Predictive R^2 summarizes the predictive performance by the factor model-implied return forecasts, aggregated over all assets and all periods.

3.
$$\text{Pricing Error } R^2 = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}(R_{i,t} - \widehat{R}_{i,t})\right)^2}{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T} R_{i,t}\right)^2}, , \tag{20}$$

where $\widehat{R}_{i,t} = \hat{\beta}_{i,t}^{\intercal} f_t$. The Pricing Error R^2 represents the fraction of the squared unconditional mean returns that is described by the common factor model, aggregated over all assets[7].

Comparing asset pricing models. First of all, we report the individual stock pricing performance, Total R^2 and Predictive R^2, in Table 1. Panel A shows P-Tree models with different number of factors, and Panel B shows the market-adjusted P-Tree model. Panels C and D provide results for Time-Series split factors, which are discussed in Section 3.4. In Panel E, we compare P-Tree models with standard factor models such as CAPM, Fama-French factor models of (Fama and French, 1993, 2015), and the Q-factor model of Hou et al. (2015), and recently published latent factor models, such as RP-PCA of Lettau and Pelger (2020) and IPCA of Kelly et al. (2019).

Total R^2 is about the cross-sectional return explanatory evaluation. Our P-Tree and the market-adjusted P-Tree show increasing positive performance for adding additional boosted factors. Second, the five-factor P-Tree and market-adjusted P-Tree outperform almost all models for pricing individual stock returns in terms of Total R^2. Predictive R^2 is equivalent to the out-of-sample R^2

[7]This measure is equivalent to the XS-R^2 used in Chen et al. (2020)

used in the return predictability literature, such as Gu et al. (2020), but demonstrates the return predictive power out of an asset pricing factor model. Both five-factor P-Tree and market-adjusted P-Tree produce positive numbers for in-sample and out-of-sample analysis. Notably, another latent factor model, IPCA of Kelly et al. (2019) shows strong performance for pricing and predicting individual stocks because of their direct objective function.

Second, we also investigate the asset pricing performance for the average returns of multiple sets of portfolios: Fama-French ME-B/M 5×5 portfolios, 49 industry portfolios, 23 leaf basis portfolios generated by P-Tree, and 18 leaf basis portfolios generated by market-adjusted P-Tree. The 23 leaf basis portfolios generated by P-Tree are displayed in Figure 4, and the 18 leaf basis portfolios generated by market-adjusted P-Tree are displayed in Figure A.1. The pricing error measure is the Pricing Error R^2 in Equation 20, which are calculated using forty years of data.

Consistent with prior literature, observable factor models (CAPM, FF3, FF5, and Q4) show great explanatory power for the cross-sectional average returns for the first three sets of test portfolios. Among the latent factor models, P-Tree, market-adjusted P-Tree, and RPPCA are almost as good. These factor model can explain more than 80% of the average returns, resulting in economically minor alpha scales. Note that though IPCA performs well for pricing individual stocks, it does not work well on pricing average returns for portfolios. However, our P-Tree model works well on pricing both individual stocks and basis portfolios.

The last column in Table 1 reveals that the 18 market-adjusted leaf basis portfolios are difficult to price using well-known observable or latent factor models, as evidenced by the large negative Pricing Error R^2. These 18 portfolios are generated by controlling the market factor, and therefore cannot be explained by CAPM. However, our market-adjusted P-Tree models explain their returns.

Investment performance, P-Tree factors, and test portfolios. We next compare the investment performances of these observable or latent factors in various models. We develop two investment strategies for each factor model: the 1/N equally-weighted portfolio and the mean-variance efficient portfolio.[8] Table 2 reports the average returns, Sharpe ratio, Jensen's alpha, and the t-statistic of alpha for both over twenty years.

[8]We follow Equation 9 to calculate portfolio weights and re-scale their sum to one.

Again, P-Tree and the market-adjusted P-Tree show increasing positive performance for adding additional boosted factors. For the in-sample analysis, the 1/N strategy of P-Tree5 in panel A generates a 2.42% monthly average return, a highly significant 1.84% alpha, and a 1.98 annualized Sharpe ratio, where the MVE strategy even provides higher results. For the out-of-sample analysis, this 1/N strategy of P-Tree5 delivers a 1.44% monthly average return, a highly significant 1.00% alpha, and a 1.25 annualized Sharpe ratio. These numbers are higher than the 1/N strategies of all other factor models, including Fama-French five factors and IPCA. These investment performance results are higher for the MVE strategy in both in-sample and out-of-sample analyses.

Finally, given that adding boosted factors is useful for pricing, prediction and investment, we investigate their additional investment information. Can these P-Tree factors be explained by observable factor models? Moreover, boosted factors are generated by shallow trees with a maximum 3 layers, and thus we can interpret these factors as interaction factors by first two splits.[9] The row names are the first two spitting characteristics of the corresponding tree. We regress each P-Tree factor on benchmark factor models to evaluate their alphas and summarize results in Table 3.

We find that almost all of our P-Tree factors show economically and statistically significant alphas against benchmark models. For example, the factor BM_IA-ME in Panel B is generated by the market-adjusted P-Tree and shows high alphas against the Fama-French 3 factors and the Q-Factor model. The MOM12M-ME factor in Panel A shows way higher alphas over the standard momentum factor of Carhart (1997). These boosted performances may come from the interaction of these characteristics, which we further investigate in Section 3.5.

Finally, note that the empirical literature primarily use portfolios (R_t in Equation 4) as test assets instead of individual stocks when evaluating asset pricing models. These test assets are usually non-overlapping portfolios that partition the the stock universe. However, standard test assets such as Fama-French ME-B/M 5×5 portfolios or other univariate/bivariate sorted portfolios may fail to span the efficient investment frontier. For data-driven test asset construction, Ahn et al. (2009) uses clustering to construct test assets by stock correlations, and Bryzgalova et al. (2020) prune the decision tree by reducing test assets created by conditional sorting. However, if one wants to apply decision trees to generate test assets, the same tree has to work for every period. This issue mo-

[9] We have a detailed discussion about interaction factors in Section 2.4.

tivates our P-Tree design that allows dissecting the cross-section relationship for multiple periods from a panel data of asset returns. The resulting leaf portfolios, for which the 18 market-adjusted leaf basis portfolios constitute one example, naturally serve as test assets for various models.

Test assets can be evaluated based on their pricing difficulty or investment performance. Given almost all models fail the asset pricing test of Gibbons et al. (1989), we consider the investment performance of test assets to see if they can span the efficient investment frontier. In Panel C of Table 3, We also report the 1/N and MVE strategies constructed by ME-B/M 5×5 portfolios, 49 Industry portfolios, 23 leaf basis portfolios, and 18 market-adjusted leaf basis portfolios. Only P-Tree and market-adjusted P-Tree portfolios show consistently significant performances as test assets in out-of-sample studies. These are strong evidence that the 23 and 18 leaf basis portfolios are valid and useful test assets that span the efficient frontier of the asset return space. Moreover, we also note that leaf portfolios from P-Trees are good test portfolios because the leaves are not near empty, as is the case in many other tree-based models.

3.4 Trees for Panel Data and Splitting the Time Series

Earlier studies have shown empirical evidence about time-varying betas and alphas during bull and bear markets (Fabozzi and Francis, 1977). P-Trees allow splitting in the time series dimension and provides an alternative investigation to factor-beta estimation with regime-switching.

We consider building a P-Tree combining time-series and cross-sectional splits. Given the short history of observations, we only consider time-series split for the first cutpoint. Without further cross-sectional splits, the first split determines the optimal macro predictor for regime-switching. After a complete search among all ten macro predictors and possible cutpoints, our model cuts inflation at 50% quantile, based on a sample from 1981 to 2000. The time-series P-Tree structure is displayed in Figure 8. We see the left and right branches hold different tree structures in high and low inflation periods, which implies the tree model adapts to different macro conditions.

Important characteristics for high and low inflation periods are reported in Figure A.2 The time-series split variable importance by selection probability is in Table 4. We find the high and low inflation periods have different characteristic importance. The long-term reversal (`MOM60M`) and 1-year seasonality (`SEAS1A`) are more important for the high inflation periods, while trading

volume (DOLVOL) and volume volatility (STD_DOLVOL) are more important for the low inflation periods. Inflation can be a lead regime-switching indicator, but these equity characteristics are valuable information to differentiate the cross-section.

Finally, the asset pricing results for the time-series P-Tree model are summarized in Panel C and D of Table 1, and their investment gains are reported in Table 2. In short, compared to standard P-Tree models, the 5-factor time-series P-Tree models in panels C and D offer higher Total and Predictive R^2 values. Also, this time-series P-Tree model provides higher and more stable Pricing Error R^2 to different test portfolios. For the recent twenty years, the 1/N equally-weighted portfolio for these five time-series tree factors provides a monthly average return of 2.78%, a highly significant alpha of 2.5%, and an annualized Sharpe ratio of 1.78. Apparently, by splitting models for high and low inflation periods, the investment performance improves tremendously. By incorporating regime-switching and time-series splits, our P-Tree model can guide market-timing in the equity market in practice, in addition to characteristic-based portfolio construction.

3.5 Nonlinearity and Interactions Revealed in P-Tree Models

One application for the P-Tree model is to exploit the nonlinearity and interaction of characteristics. CART might be the only nonlinear machine learning method that displays these interactions in the tree diagram and nonlinear patterns in the partition plot. Examples can be found on Figure 4 and 5. For the tree diagram, the long-short portfolio is to have a long position on one leaf and a short position on another one. For the partition plot, the long-short portfolio is to have a long position on one partition and a short position on another one. The economic theory and empirical literature determine the long and short leg directions for anomalies, such as small-minus-big or high-minus-low.

Trading strategies using nonlinearity and interaction. We collect the long-short direction of each characteristic from the literature and create four different long-short characteristic-based portfolios, as shown in Figure A.3. Specifications (a) and (b) are the standard univariate-sorted long-short portfolios. Sorting the stocks in four quarters rather than two halves increases the return spread of long-short factors. Specifications (c) and (d) are interaction portfolios generated by our P-Tree

and market-adjusted P-Tree models, which are essentially sequentially sorted long-short portfolios. We restrict all splitting cutpoints to be 0 (50%) for this example, and the training sample fits these interaction portfolios from 1981 to 2000.

We summarize the number of significantly positive cases of average returns and Jensen's alphas in Table 5 panel A and B for 5% and 10% significance levels. First of all, "Uni-Sort 4×1" has more significant factors than "Uni-Sort 2×1" in the training, test, and entire samples. Second, there are fewer significant factors in the test sample than the training sample due to the post-publication bias (McLean and Pontiff, 2016). Probably due to estimation error, the numbers of significant factors are higher for the entire sample.

Third, our interaction factors have more "average-significant" cases than "Uni-Sort 4×1" in the recent twenty years and the entire sample. Our market-adjusted interaction factors also have more "alpha-significant" factors than "Uni-Sort 4×1." Out of 61 factors for forty years, the interaction specification increases the significant factors from 27 to 36, and the market-adjusted interaction specifications increase from 36 to 48. These findings are consistent with the goals of P-Tree and market-adjusted P-Tree models.

In Table 5 panel C, we report the cross-sectional quantile numbers of the average returns and Jensen's alphas. We find those "Uni-Sort 4×1" factors produce high median (50%) average returns and Jensen's alphas in the first twenty years. However, the returns from these factors have decreased dramatically in the recent twenty years. Conversely, the interaction factors have higher median average returns and alphas in the recent twenty years and the entire sample. The market-adjusted interaction factors show consistently positive 25% values for alphas in every case. In summary, utilizing information about interactions might be a solution to the post-publication decay for risk premia of anomalies.

Further discussion of interaction portfolios. To further understand how these interaction portfolios behave, six examples are reported in Table 6 and Figure A.4. The "Uni-sort 4x1" portfolios show significant average returns and alphas, but the interaction portfolios increase these results. For example, the "Uni-sort 4x1" portfolio Standardized Unexpected Earnings (`SUE`) of Rendleman Jr et al. (1982) is persistently strong. Once we create an interaction factor with R&D to Market (`RMD`) and

Market Equity (ME), the results almost double. RMD is useful for the short leg of SUE, and ME is helpful for the long leg of SUE. We show the average returns for leaf portfolios during splitting in Figure A.4 and find decreasing and increasing numbers in the corresponding direction.

Many factors do not show significant risk premia or alphas and are deemed not replicable. However, once we include the "correct" control or interaction, some of these factors can be resurrected.[10] For example, we find "Uni-sort 4x1" factor Change in Profit Margin (CHPM) of Soliman (2008) has almost zero premium. Interacting with R&D to Sales (RD_SALE) on the short leg and Dollar Trading Volume (DOLVOL) on the long leg, this factor earns 47 basis points for monthly average returns and a 51 basis point for alpha. It is not surprising that the greedy algorithm of P-Tree helps to lower the short leg and raise the long leg for the in-sample analysis. Consistent results in the out-of-sample in Table 6 show strong evidence for our P-Tree model.

Interpretability and robustness. Observing feature interactions gives P-Tree models great interpretability. Nonlinear tree structure and interaction of characteristics are displayed when the decision tree continues the split to further layers. However, these patterns gleaned from individual P-Trees are only helpful if the P-Tree model can avoid overfitting. Following Section 2.4, we use Random Forest to check if a P-Tree model indeed splits using similar characteristics and is not reflections of noise in the data.

Figure 9 summarizes the feature importance in a random forest of P-Trees. According to (13), a negative value implies the model has an average high asset pricing performance value when we include this characteristic rather than dropping it. We also conduct a two-sample t-test using the bootstrap samples and display those 5% significant characteristics with a deeper color. Only four characteristics show statistical significance out of 61 of them. The in-sample results identify the significant characteristics as volume volatility (STD_DOLVOL), long-term reversal (MOM60M), market-adjusted market equity (ME_IA), and 1-year seasonality (SEA1A). These four characteristics also show asset pricing improvement for the recent twenty years.

Table 4 panel A summarizes the selection frequency of characteristics in top splits, based on (12). We find that volume volatility (STD_DOLVOL) has a 73% chance of being the first splitting

[10] Related is Asness et al. (2018) that resurrects the size factor by controlling for the junk factor.

characteristic once the bootstrap sample contains it. The other important characteristics are long-term reversal (MOM60M), trading volume (DOLVOL), and market-adjusted market equity (ME_IA).

Overall, the two different metrics reveal the same small set of important features for splitting the data. Given that volume volatility (STD_DOLVOL), long-term reversal (MOM60M), and market-adjusted market equity (ME_IA) are also the characteristics for the first split in the single-factor (Figure 4 and 8) or boosted multi-factor P-Trees, we know that P-Tree models behave similarly to a random forest ensemble and are not overfitted. The nonlinearities and interactions revealed in P-Tree models provide useful information but are difficult to see in ensembles.

4 Conclusion

We introduce P-Tree — a class of tree-based models for panel data analysis, which complements traditional classification and regression tree models by relaxing their common i.i.d. assumption and uses global criteria to guide cutpoint choices. When applied to a panel data of asset returns, P-Tree partitions the cross-section of the asset universe and generates a time-series factor model, effectively combining regression trees and latent factor models while allowing both time-series and cross-sectional splits. We further discuss boosting trees for multiple factors, random forest for feature importance, and nonlinear interactions for latent factor interpretation. Finally, we apply P-Tree to explain the cross-sectional variation of U.S. equity returns and achieve outstanding performance and profitable investments.

P-Tree models can be used beyond asset pricing and investment to analyze large panel data in general. Not only is the latent factor generation design easy to customize for different data-generating processes, the framework can be flexibly adopted as a supervised machine learning algorithm to have split criteria defined by the objective function in specific domain studies. P-Tree also admits many potential promising extensions. For example, future works can integrate the time-series split and cross-sectional split in the split criteria for every partition; one can conveniently model penalty terms or prior information a regularized version of P-Tree and Bayesian P-Tree; one can adjust the split criterion for portfolio optimization to create a minimum variance portfolio.[11]

[11] To facilitate more applications of P-Tree in economic, finance, and all other studies, we publish the R package TreeFactor for P-Trees on the authors' websites.

References

Ahn, D.-H., J. Conrad, and R. F. Dittmar (2009). Basis assets. *The Review of Financial Studies 22*(12), 5133–5174.

Ang, A. and A. Timmermann (2012). Regime changes and financial markets. *Annu. Rev. Financ. Econ. 4*(1), 313–337.

Arnett, F. C., S. M. Edworthy, D. A. Bloch, D. J. Mcshane, J. F. Fries, N. S. Cooper, L. A. Healey, S. R. Kaplan, M. H. Liang, H. S. Luthra, et al. (1988). The american rheumatism association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology 31*(3), 315–324.

Asness, C., A. Frazzini, R. Israel, T. J. Moskowitz, and L. H. Pedersen (2018). Size matters, if you control your junk. *Journal of Financial Economics 129*(3), 479–509.

Avramov, D. (2004). Stock return predictability and asset pricing models. *The Review of Financial Studies 17*(3), 699–738.

Avramov, D., S. Cheng, and L. Metzker (2021). Machine learning versus economic restrictions: Evidence from stock return predictability. *Management Science, accepted*.

Bali, T. G., D. Huang, F. Jiang, and Q. Wen (2021). Different strokes: Return predictability across stocks and bonds with machine learning and big data. Technical report, Georgetown University.

Bianchi, D., M. Büchner, and A. Tamoni (2021). Bond risk premiums with machine learning. *The Review of Financial Studies 34*(2), 1046–1089.

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984). *Classification and regression trees*. Routledge.

Bryzgalova, S., J. Huang, and C. Julliard (2021). Bayesian solutions for the factor zoo: We just ran two quadrillion models. Technical report, London Business School.

Bryzgalova, S., M. Pelger, and J. Zhu (2020). Forest through the trees: Building cross-sections of stock returns. Technical report, London Business School.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of Finance 52*(1), 57–82.

Chen, L., M. Pelger, and J. Zhu (2020). Deep learning in asset pricing. Technical report, Stanford University.

Chen, T. and C. Guestrin (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Chipman, H. A., E. I. George, R. E. McCulloch, et al. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics 4*(1), 266–298.

Cong, L. W., K. Tang, J. Wang, and Y. Zhang (2020). Alphaportfolio: Direct construction through deep reinforcement learning and interpretable ai. *Working Paper*.

Cong, L. W., K. Tang, J. Wang, and Y. Zhang (2021). Deep sequence modeling: Development and applications in asset pricing. *The Journal of Financial Data Science 3*(1), 28–42.

Creal, D. and J. Kim (2021). Empirical asset pricing with bayesian regression trees. Technical report, University of Oklahoma.

De Prado, M. L. (2018). *Advances in financial machine learning*. John Wiley & Sons.

DeMiguel, V., A. Martin-Utrera, F. J. Nogales, and R. Uppal (2020). A transaction-cost perspective on the multitude of firm characteristics. *The Review of Financial Studies 33*(5), 2180–2222.

Dong, X., Y. Li, D. Rapach, and G. Zhou (2021). Anomalies and the expected market return. *Journal of Finance, Forthcoming*.

Fabozzi, F. J. and J. C. Francis (1977). Stability tests for alphas and betas over bull and bear market conditions. *The Journal of Finance 32*(4), 1093–1099.

Fama, E. F. and K. R. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics 33*(1), 3–56.

Fama, E. F. and K. R. French (2015). A five-factor asset pricing model. *Journal of Financial Economics 116*(1), 1–22.

Feng, G., A. Fulap, and J. Li (2020). Real-time macro information and bond return predictability: Does deep learning help?

Feng, G., S. Giglio, and D. Xiu (2020). Taming the factor zoo: A test of new factors. *The Journal of Finance 75*(3), 1327–1370.

Feng, G. and J. He (2019). Factor investing: Hierarchical ensemble learning. *Journal of Econometrics, forthcoming*.

Feng, G., N. Polson, and J. Xu (2020). Deep learning of characteristics-sorted factor models. Technical report, City University of Hong Kong.

Freund, Y. and R. E. Schapire (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences 55*(1), 119–139.

Freyberger, J., A. Neuhierl, and M. Weber (2019). Dissecting characteristics nonparametrically. *The Review of Financial Studies, Forthcoming*.

Friedman, J., T. Hastie, R. Tibshirani, et al. (2001). *The elements of statistical learning*, Volume 1. Springer series in statistics New York.

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

Gerakos, J. J. and R. B. Gramacy (2012). Regression-Based Earnings Forecasts. Technical report, Dartmouth College.

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

Gu, S., B. Kelly, and D. Xiu (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies 33*(5), 2223–2273.

Gu, S., B. Kelly, and D. Xiu (2021). Autoencoder asset pricing models. *Journal of Econometrics 222*(1), 429–450.

Guisan, A. and N. E. Zimmermann (2000). Predictive habitat distribution models in ecology. *Ecological modelling 135*(2-3), 147–186.

He, J. and P. R. Hahn (2021). Stochastic tree ensembles for regularized nonlinear regression. *Journal of the American Statistical Association* (just-accepted), 1–61.

He, J., S. Yalov, and P. R. Hahn (2019). Xbart: Accelerated bayesian additive regression trees. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1130–1138.

He, X., G. Feng, J. Wang, and C. Wu (2021). Predicting individual corporate bond returns. Technical report, City University of Hong Kong.

Hou, K., C. Xue, and L. Zhang (2015). Digesting anomalies: An investment approach. *The Review of Financial Studies 28*(3), 650–705.

Kelly, B. T., S. Pruitt, and Y. Su (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics 134*(3), 501–524.

Kim, S., R. A. Korajczyk, and A. Neuhierl (2021). Arbitrage portfolios. *The Review of Financial Studies 34*(6), 2813–2856.

Kozak, S., S. Nagel, and S. Santosh (2020). Shrinking the cross-section. *Journal of Financial Economics 135*(2), 271–292.

Lemon, S. C., J. Roy, M. A. Clark, P. D. Friedmann, and W. Rakowski (2003). Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Annals of behavioral medicine 26*(3), 172–181.

Lettau, M. and M. Pelger (2020). Estimating latent asset-pricing factors. *Journal of Econometrics 218*(1), 1–31.

Li, B. and A. G. Rossi (2020). Selecting mutual funds from the stocks they hold: A machine learning approach. Technical report, Wuhan University.

Maheu, J. M. and T. H. McCurdy (2000). Identifying bull and bear markets in stock returns. *Journal of Business & Economic Statistics 18*(1), 100–112.

McLean, R. D. and J. Pontiff (2016). Does academic research destroy stock return predictability? *The Journal of Finance 71*(1), 5–32.

Moritz, B. and T. Zimmermann (2016). Tree-based conditional portfolio sorts: The relation between past and future stock returns. Technical report, Ludwig Maximilian University Munich.

Nagel, S. (2021). *Machine Learning in Asset Pricing*. Princeton University Press.

Rendleman Jr, R. J., C. P. Jones, and H. A. Latane (1982). Empirical anomalies based on unexpected earnings and the importance of risk adjustments. *Journal of Financial Economics 10*(3), 269–287.

Rossi, A. G. (2018). Predicting stock market returns with machine learning. Technical report, Georgetown University.

Rossi, A. G. and A. Timmermann (2015). Modeling covariance risk in merton's icapm. *The Review of Financial Studies 28*(5), 1428–1461.

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

Figure 4: Panel Tree for the period from 1981 to 2000

The P-Tree trained from the period from 1981 to 2000 is displayed in this figure. We show splitting characteristics and cutpoint values for each parent nodes. The node numbers (N#) and splitting order numbers (S#) are also printed on each parent node. We have included the median monthly number of assets in the final leaves basis portfolios. The description of characteristics are listed in Table A.1.

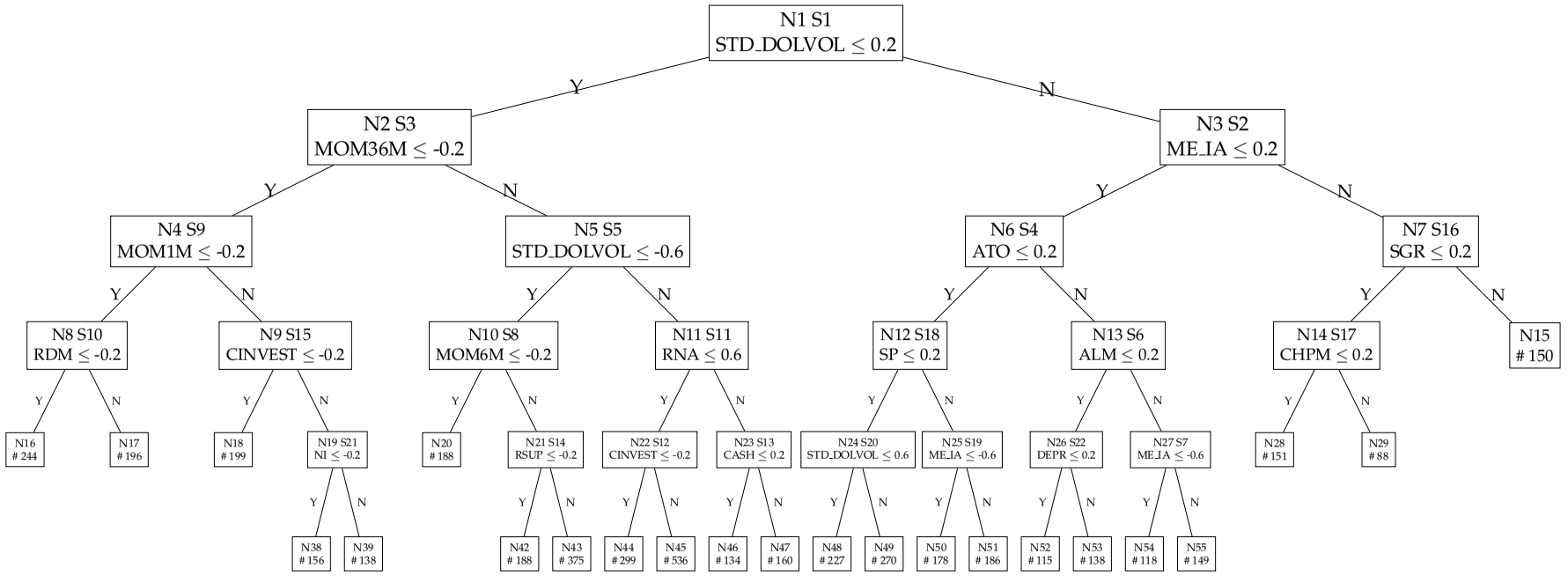


Figure 5: Partition Plots for Figure 4

This diagram visualizes the partition for the first five splits of the tree structure in Figure 4. For example, the first split (S1) is implemented with `STD_DOLVOL` on the entire stock universe, and the second split is implemented with `ME_IA` on the high `STD_DOLVOL` portfolios. The space area for each partition represents the corresponding proportion of the stock universe. We also provide the monthly average return and annualized Sharpe ratio for each leaf. The overlayed arrows represent the next split is implemented on the partitioned area from the previous partition.

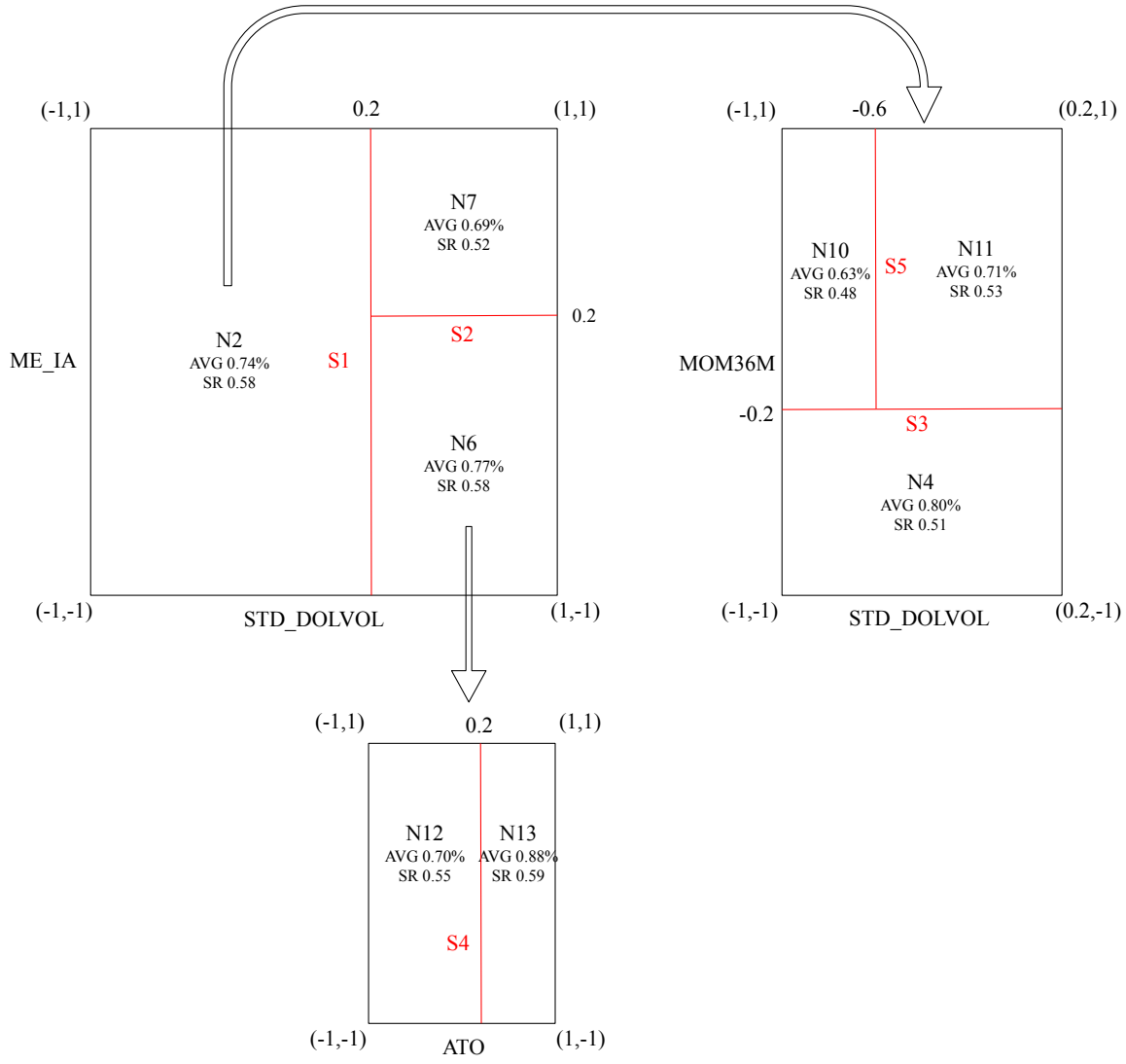


Figure 6: Return-Risk Clustering for Leaf Basis Portfolios

Following the tree structure in Figure 4, this figure plots the monthly average return and standard deviation for leaf basis portfolios. Labels N4, N5, N6, and N7 represent those four leaf basis portfolios at depth 3. Labels N4+, N5+, N6+, and N7+ represent corresponding leaf basis portfolios at depth 6. We also show our generated SDF portfolios at depth 3 and depth 6 labeled with MVE-3 and MVE-6. For the reference lines, we have included the annualized Sharpe ratio 0.3, 0.6, and 0.9.

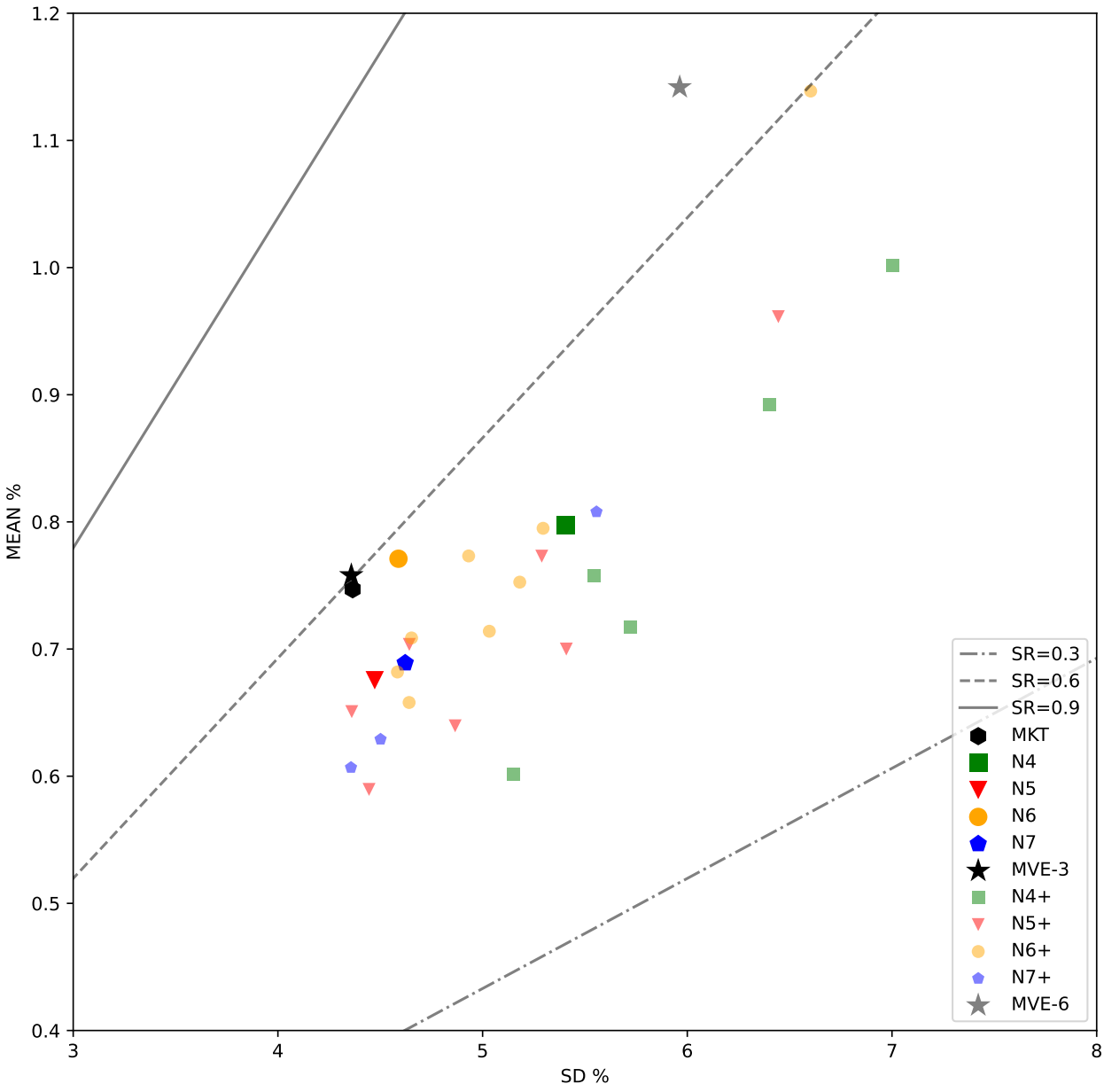


Figure 7: Out-of-Sample: Return-Risk Clustering for Leaf Basis Portfolios

This figure shows the out-of-sample performance for all those portfolios plotted in Figure 6. The figure format follows Figure 6.

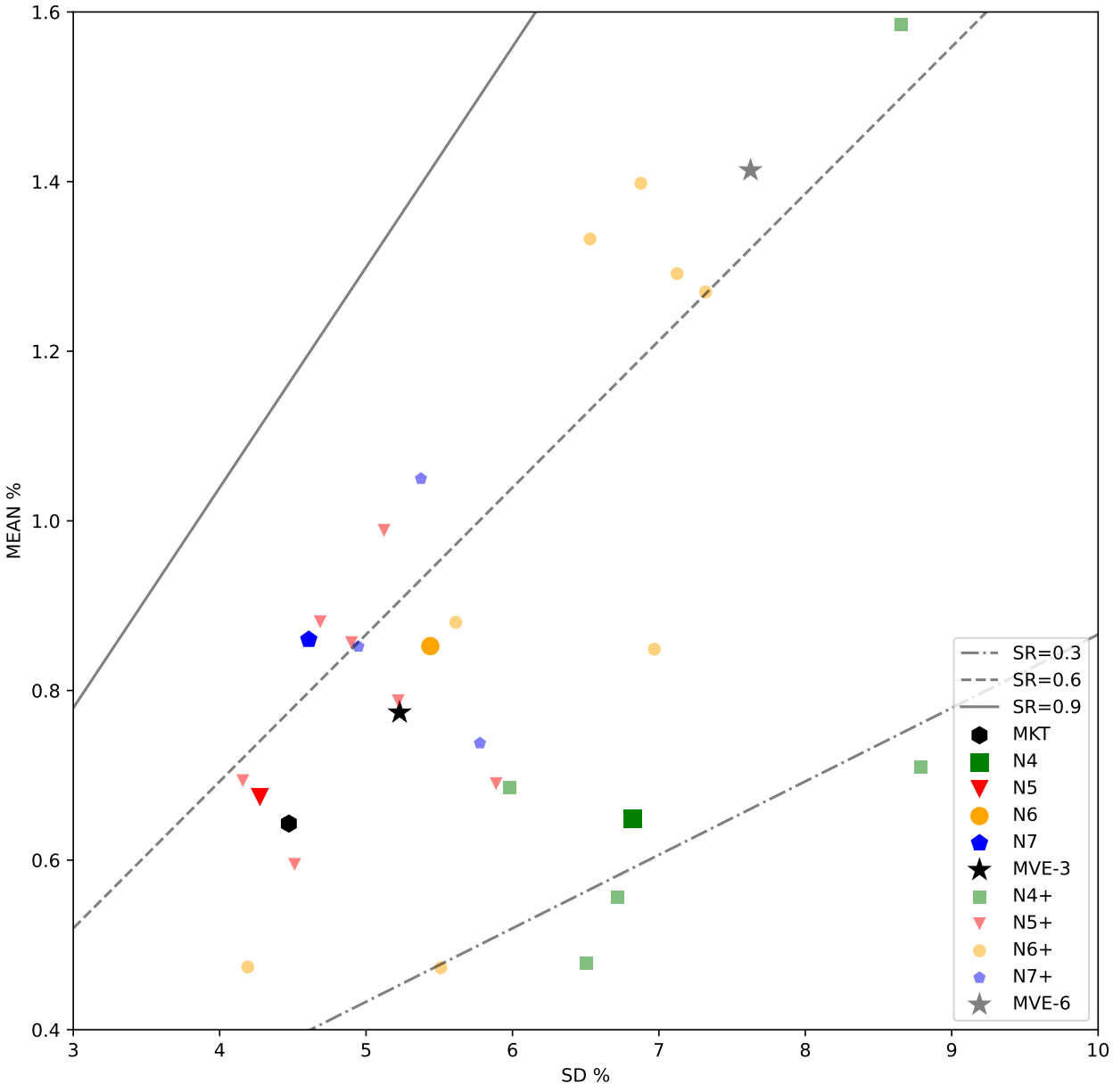


Figure 8: Panel Tree for the period from 1981 to 2000: High/Low Inflation

This figure shows the Panel Tree structure by considering both time-series and cross-sectional variation. The most important macro predictor is Inflation, and the first split is implemented when the current inflation level is lower than the median of the past decade. For high and low inflation periods, two different tree models are provided as two child leaves. The figure format follows Figure 4.

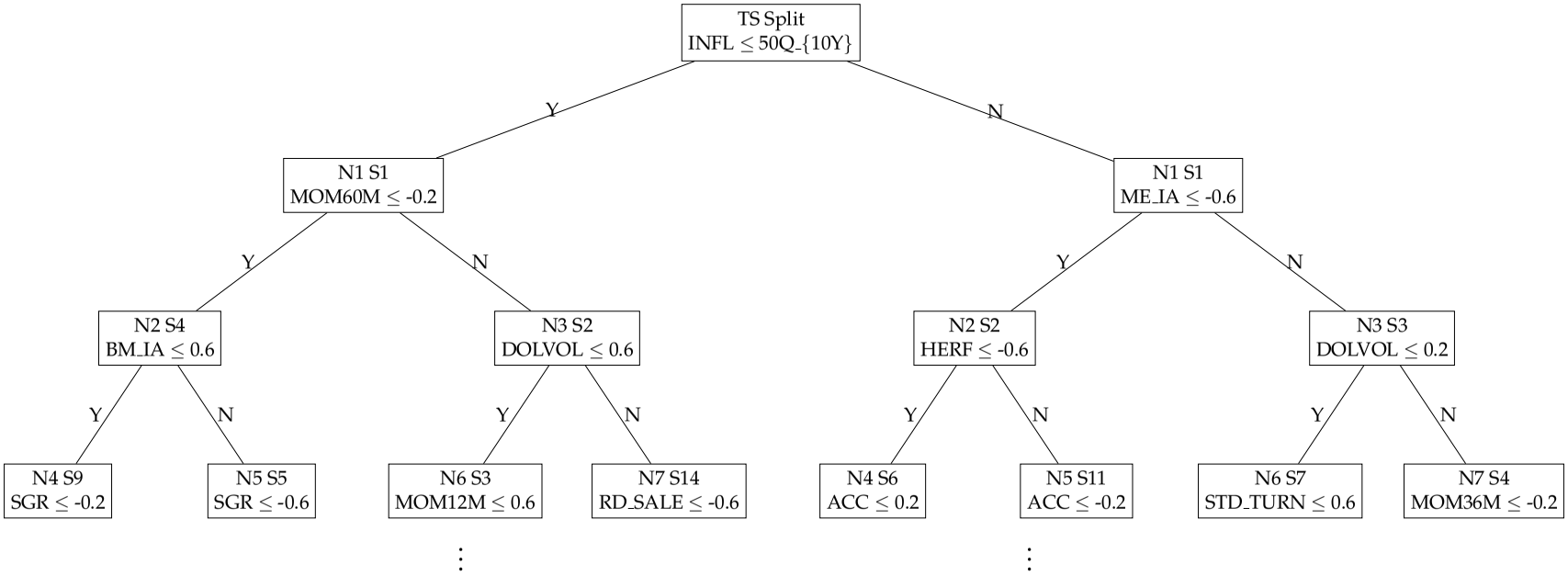
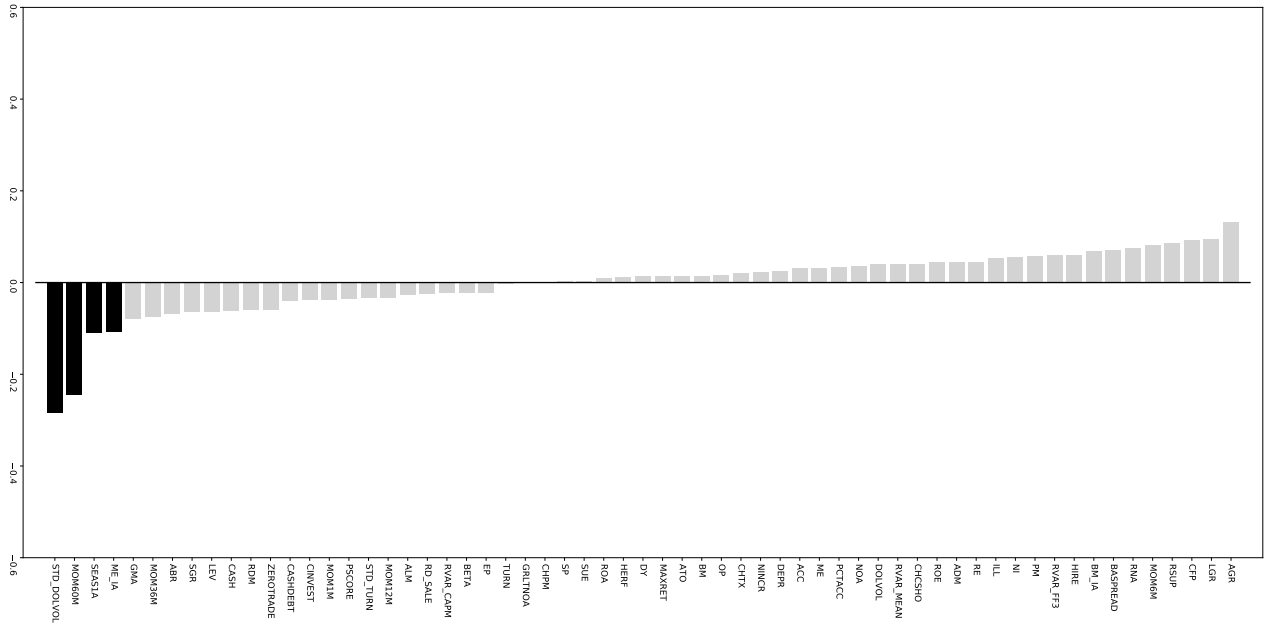
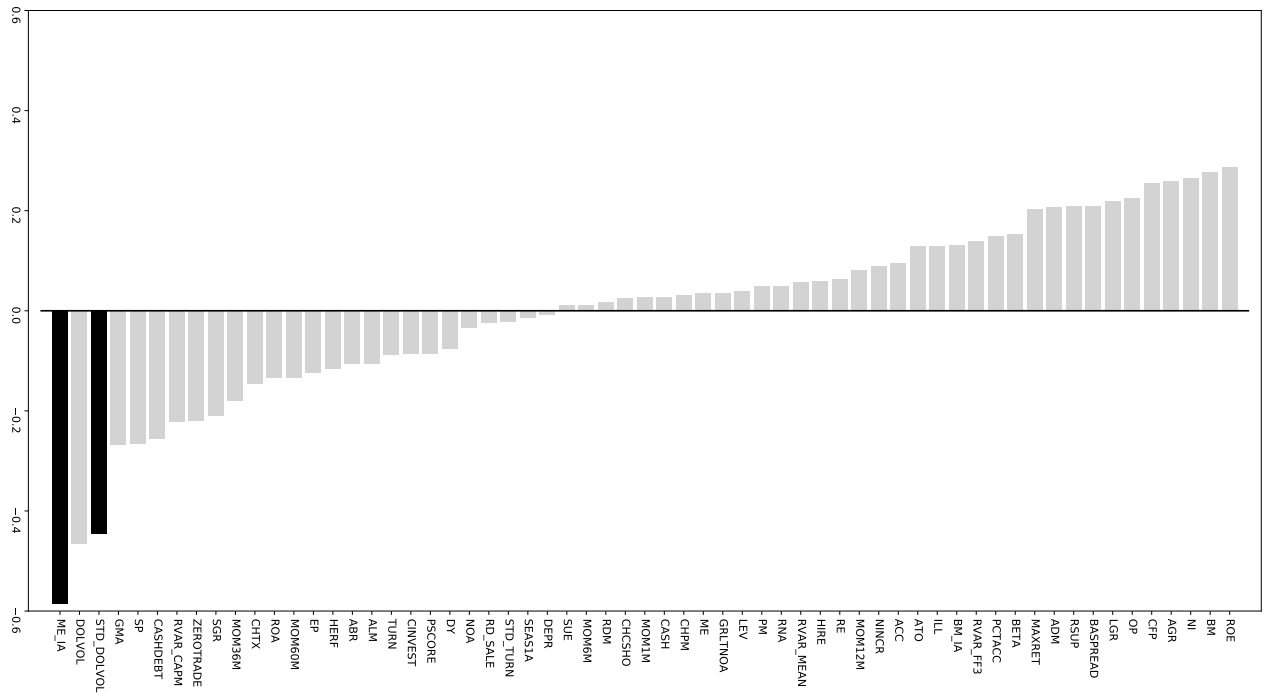


Figure 9: Out-of-Bag Characteristics Significance

This figure reports the characteristic significance by the out-of-bag ensembles from the random forest of 500 trees. The train sample period is 1981-2000, and the test sample period is 2001-2020. The variable importance measure is defined as the average increase percentage of loss function by including a characteristic in a tree model. A negative value implies that including this characteristic reduces loss and is useful. The dark color bars on the left are significant characteristics at the 5% level by the two-sample t-test. The description of characteristics are listed in Table A.1.



1981-2000



2001-2020

Table 1: Asset Pricing Performance

This table reports the performances of asset pricing models. The 'Tot' (total R^2 %) in Equation 18 and 'Pred' (predictive R^2 %) in Equation 19 are goodness of fitness measures for individual stock returns. The in-sample period is from 1981 to 2000, and the out-of-sample period is from 2001 to 2020. Panel A shows results for the Panel Tree model with #factors. Panel B shows the market-adjusted P-Tree, with the Market factor as the first factor and others generated by boosting P-Trees. Panel C and D demonstrate the same performance when applying the time-series split in Section 2.5, while Panel E provides comparisons for commonly used benchmark models. We also report the Pricing Error R^2 % in Equation 20, using the factor models in the rows to price the test asset portfolios in the columns. Pricing Error R^2 values are estimated using the entire sample from 1981 to 2020. The "P-Tree23" indicates the 23 basis portfolios generated by the P-Tree in Figure 4, and the "Ma-PT18" indicates the 18 basis portfolios generated by the market-adjusted P-Tree in Figure A.1.

	Individual Stocks				Portfolios			
	In-Sample		Out-of-Sample					
	Tot	Pred	Tot	Pred	FF25	Ind49	P-Tree23	Ma-PT18
Panel A: Panel Tree Factors								
PTree1	8.46	-0.04	12.38	0.29	88.0	82.2	97.1	-250.3
PTree5	11.63	0.37	15.78	0.38	68.1	15.5	77.0	-70.6
Panel B: MKT Adj. Panel Tree Factors								
MaPTree2	9.86	0.03	12.80	-0.25	79.2	92.2	59.2	79.2
MaPTree5	11.60	0.24	16.27	0.18	77.9	76.9	86.0	61.4
Panel C: Time-Series Split - Panel Tree Factors								
TS-PTree1	9.16	0.00	13.22	0.28	65.8	17.5	71.1	-113.9
TS-PTree5	12.09	0.57	16.21	0.09	76.1	35.2	81.1	6.5
Panel D: Time-Series Split - MKT Adj. Panel Tree Factors								
TS-MaPTree2	10.39	0.23	14.01	-0.20	71.6	84.6	49.6	56.9
TS-MaPTree5	12.12	0.49	16.53	0.02	86.8	58.4	83.1	51.2
Panel E: Others								
CAPM	6.86	0.05	11.18	0.29	91.2	87.7	93.5	-122.5
FF3	9.92	0.19	14.01	0.38	94.8	84.9	92.8	-110.1
FF5	10.34	0.24	14.57	0.38	96.0	77.9	91.6	-25.9
Q4	10.07	0.30	14.49	0.34	95.7	81.7	88.6	31.8
RPPCA	10.31	0.03	14.44	0.36	77.8	57.6	84.7	-258.7
IPCA	12.54	1.23	16.46	0.87	-11.7	-69.1	-28.9	-281.7

Table 2: Investment Performance

This table reports the investment performance of the factor models. The table format is similar to Table 1. We report the average returns, Sharpe ratio, and monthly Jensen's alpha % for the equally-weighted (1/N) and mean-variance efficient (MVE) portfolios. For t-statistics *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

	In-Sample (1981-2000)						Out-of-Sample (2001-2020)					
	1/N			MVE			1/N			MVE		
	AVG	SR	α	AVG	SR	α	AVG	SR	α	AVG	SR	α
Panel A: Panel Tree Factors												
PTree1	1.14	0.66	0.34	1.14	0.66	0.34	1.41	0.64	0.48	1.41	0.64	0.48
PTree5	2.42	1.98	1.84***	4.44	3.24	4.22***	1.44	1.25	1.00***	2.18	1.12	2.36***
Panel B: MKT Adj. Panel Tree Factors												
MaPTree2	3.24	1.52	2.55***	4.85	1.56	4.19***	0.89	0.49	0.48	1.05	0.39	0.79
MaPTree5	2.34	1.78	1.70***	4.97	3.74	4.74***	1.06	0.95	0.64***	2.63	1.71	2.46***
Panel C: Time-Series Split - Panel Tree Factors												
TS-PTree1	1.11	0.71	0.36*	1.11	0.71	0.36*	1.20	0.63	0.37*	1.20	0.63	0.37*
TS-PTree5	4.18	2.43	3.53***	7.60	3.59	7.22***	2.78	1.78	2.50***	4.25	2.02	4.15***
Panel D: Time-Series Split - MKT Adj. Panel Tree Factors												
TS-MaPTree2	3.25	1.14	2.56***	3.67	1.16	3.00***	1.14	0.54	0.98**	1.22	0.52	1.14**
TS-MaPTree5	2.67	1.38	1.98***	2.96	1.97	2.47***	1.15	0.77	0.78***	1.58	1.24	1.35***
Panel E: Others												
FF3	0.38	0.85	0.20***	0.55	1.17	0.41***	0.28	0.40	0.00	0.21	0.29	-0.06
FF5	0.38	1.34	0.33***	0.46	1.47	0.39***	0.25	0.59	0.12	0.26	0.62	0.12*
Q4	0.52	1.40	0.36***	0.58	1.95	0.52***	0.29	0.77	0.16***	0.22	0.69	0.18***
RP-PCA	0.48	0.21	-0.63***	2.09	1.64	1.72***	0.91	0.51	0.16	0.93	0.47	0.21
IPCA	1.85	1.96	1.40***	2.84	4.94	2.80***	1.40	1.22	0.91***	1.78	2.81	1.67***

Table 3: Factor Spanning Alpha Test

This table reports the monthly alphas in basis point and their significance for the factor spanning test. Panel A shows results for each factor of the Panel Tree model in Table 1, and Panel B shows the market-adjusted tree factors. We regress each of the factors in the rows against a factor model in the columns. We name the factors by the first two splitting characteristics in the tree structure for the P-Tree factor, which can also be viewed as interaction factors. We also show the mean-variance efficient (MVE) and 1/N portfolios for five factors. Panel C provides the same results for different test assets. Notably, "PT23" represents the 23 basis portfolios from our P-Tree model in Figure 4, and "Ma-PT18" is the 18 basis portfolios from our market-adjusted P-Tree model in Figure A.1. For t-statistics *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

	In-Sample		Out-of-Sample		Entire-Sample	
	FF3	Q4	FF3	Q4	FF3	Q4
			Panel A: Panel Tree Factors			
P-Tree1	45***	65***	39*	65***	38***	67***
OP-ME	170***	61	137***	55	156***	59**
BM_IA-DOLVOL	44*	61**	116***	124***	85***	101***
MOM12M-ME	470***	392***	261***	149***	364***	252***
ME-STD_DOLVOL	36*	9	-7	-21	13	-8
MVE(5-factor)	370***	317***	257***	179***	314***	239***
1/N(5-factor)	153***	118***	109***	74***	131***	94***
			Panel B: MKT Adj. Panel Tree Factors			
RVAR_FF3-STD_TURN	435***	300***	148**	18	275***	138***
BM_IA-ME	57***	71***	88***	112***	81***	104***
MOM12M-ME_IA	227***	157***	149***	65*	182***	87***
ME-CASH	102***	130***	-21	17	38***	70***
MVE(5-factor)	431***	376***	257***	216***	345***	291***
1/N(5-factor)	164***	131***	73***	43***	115***	80***
			Panel C: Other Test Assets			
MVE-FF25	281***	253***	129***	68*	192***	150***
MVE-IND49	329***	323***	47	77	70*	89**
MVE-PT23	45***	65***	39*	65***	38***	67***
MVE-Ma-PT18	435***	300***	148***	18	275***	138***
1/N-FF25	-5	-10*	0	3	-1	0
1/N-IND49	-20**	-39***	10	6	32*	40**
1/N-PT23	-1	1	12*	21***	7*	13***
1/N-MAPT18	79***	59***	40***	14*	58***	32***

Table 4: Characteristics Importance by Top Splits

This table reports the most frequently selected characteristics from the random forest of 500 trees, which can be used to assess the variable importance. The random forest ensemble design is discussed in Section 2.4. The "Top 1" rows only count the first split for 500 trees. The "Top 2" or "Top 3" rows only count the first two or three splits. The numbers reported are the selection frequency for these top characteristics selected out of the 500 ensembles. Panel A uses the entire train sample, and the other panels report the mostly characteristics chosen for high and low inflation periods from 1981 to 2000. The description of characteristics is listed in Table A.1.

Panel A: Entire Training Sample (1981-2000)

	1	2	3	4	5
Top1	STD_DOLVOL 0.73	MO60M 0.64	DOLVOL 0.61	ME_IA 0.33	LGR 0.29
Top2	STD_DOLVOL 0.78	DOLVOL 0.73	MMO60M 0.72	ME_IA 0.61	BETA 0.44
Top3	STD_DOLVOL 0.78	MMO60M 0.76	DOLVOL 0.73	ME_IA 0.64	ATO 0.52

Panel B: High Inflation

	1	2	3
Top 1	MOM60M 0.53	DOLVOL 0.44	ME_IA 0.42
Top 2	MOM60M 0.59	DOLVOL 0.56	ME_IA 0.53
Top 3	MOM60M 0.60	DOLVOL 0.56	ME_IA 0.53

Panel C: Low Inflation

	1	2	3
Top 1	ME_IA 0.56	DOLVOL 0.42	ME 0.41
Top 2	ME_IA 0.68	ME 0.55	DOLVOL 0.49
Top 3	ME_IA 0.63	MOM60M 0.50	ME 0.47

Table 5: Uni-Sort Factors v.s. Interaction Factors

This table summarizes the significant counts for long-short factors for average returns and Jensen's alphas. We count the number of significant average returns and alphas at the 5 % and 10% level in panels A and B. We report the cross-sectional quantile values of average returns and alphas in panel C among the 61 characteristics. The four specifications are (1) 4x1 long-short portfolio, (2) 2x1 long-short portfolio, (3) interaction factors, and (4) market-adjusted interaction factors. Our panel tree creates specifications (3) and (4), which follow the same models in Table 1.

Panel A: # of Significant Cases with 5% level

	Uni-Sort 4x1		Uni-Sort 2x1		Interaction		Mkt-Adj Interaction	
	# Mean	# Alpha	# Mean	# Alpha	# Mean	# Alpha	# Mean	# Alpha
81-00	17	30	6	22	11	27	10	30
01-20	4	22	5	11	24	28	11	21
81-20	22	33	12	28	32	31	16	42

Panel B: # of Significant Cases with 10% level

	Uni-Sort 4x1		Uni-Sort 2x1		Interaction		Mkt-Adj Interaction	
	# Mean	# Alpha	# Mean	# Alpha	# Mean	# Alpha	# Mean	# Alpha
81-00	26	34	10	31	18	30	19	39
01-20	8	27	7	18	29	30	14	33
81-20	27	36	18	34	36	41	21	48

Panel C: Cross-Sectional Quantiles for Average and Alpha

		Uni-Sort 4x1		Uni-Sort 2x1		Interaction		Mkt-Adj Interaction	
	q	Avg	Alpha	Avg	Alpha	Avg	Alpha	Avg	Alpha
	25	0.12	0.06	0.07	0.06	0.09	0.09	0.1	0.22
81-00	50	0.33	0.46	0.18	0.21	0.22	0.28	0.19	0.29
	75	0.61	0.69	0.23	0.31	0.33	0.43	0.27	0.41
	25	0.02	0.07	-0.02	0.05	0.06	0.08	-0.02	0.13
01-20	50	0.18	0.29	0.06	0.13	0.41	0.36	0.09	0.25
	75	0.36	0.58	0.2	0.31	0.56	0.65	0.25	0.41
	25	0.13	0.14	0.04	0.08	0.08	0.12	0.05	0.17
81-20	50	0.28	0.34	0.11	0.16	0.30	0.29	0.14	0.28
	75	0.41	0.61	0.19	0.26	0.43	0.48	0.25	0.39

Table 6: Examples for Interaction Factors

This table follows the 6 examples in Figure A.4. We report the monthly average returns (%) and Jensen's alpha (%) of the 4x1 long-short factors and the interaction factors. The interaction factors are created with the train sample period from 1981 to 2000, and we have provided the corresponding interaction characteristics. For t-statistics *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The description of characteristics are listed in Table A.1.

Panel A: Stand. Unexp. Earnings

	Uni-Sort 4x1		+RDM+ME	
	Mean	Alpha	Mean	Alpha
81-00	0.68***	0.63***	0.95***	0.91***
01-20	0.53***	0.67***	1.43***	1.36***
81-20	0.61***	0.66***	1.19***	1.13***

Panel B: Profit Margin

	Uni-Sort 4x1		+RDS+DOLVOL	
	Mean	Alpha	Mean	Alpha
81-00	0.10	0.06	0.39***	0.46***
01-20	0.18	0.28*	0.55***	0.56***
81-20	0.14	0.17	0.47***	0.51***

Panel C: Cash Holdings

	Uni-Sort 4x1		+NOA+BAS	
	Mean	Alpha	Mean	Alpha
81-00	0.39	0.05	1.12***	1.10***
01-20	0.43*	0.21	-0.04	0.02
81-20	0.41**	0.14	0.54***	0.56***

Panel D: Dollar Trading Volume

	Uni-Sort 4x1		+RVAR_FF3+ME	
	Mean	Alpha	Mean	Alpha
81-00	-0.20	0.02	0.11	0.33**
01-20	0.33*	0.42**	0.34*	0.45***
81-20	0.06	0.22	0.23*	0.39***

Panel E: Seasonality

	Uni-Sort 4x1		+RDM+DOLVOL	
	Mean	Alpha	Mean	Alpha
81-00	0.54**	0.37*	0.13	0.21
01-20	0.01	0.05	0.47**	0.46**
81-20	0.27	0.22	0.30*	0.33**

Panel F: R&D to Sales

	Uni-Sort 4x1		+RVAR_FF3+EP	
	Mean	Alpha	Mean	Alpha
81-00	0.43	0.08	1.11***	1.29***
01-20	-0.06	-0.28	0.04	0.49*
81-20	0.19	-0.10	0.58***	0.90***

Appendices

Table A.1: Equity Characteristics (61 in total)

This table lists the description for equity characteristics used in the empirical study.

No.	Characteristics	Description
1	ABR	Abnormal returns around earnings announcement
2	ACC	Operating Accruals
3	ADM	Advertising Expense-to-market
4	AGR	Asset growth
5	ALM	Quarterly Asset Liquidity
6	ATO	Asset Turnover
7	BASPREAD	Bid-ask spread (3 months)
8	BETA	Beta (3 months)
9	BM	Book-to-market equity
10	BM_IA	Industry-adjusted book to market
11	CASH	Cash holdings
12	CASHDEBT	Cash to debt
13	CFP	Cashflow-to-price
14	CHCSHO	Change in shares outstanding
15	CHPM	Industry-adjusted change in profit margin
16	CHTX	Change in tax expense
17	CINVEST	Corporate investment
18	DEPR	Depreciation / PP&E
19	DOLVOL	Dollar trading volume
20	DY	Dividend yield
21	EP	Earnings-to-price
22	GMA	Gross profitability
23	GRLTNOA	Growth in long-term net operating assets
24	HERF	Industry sales concentration
25	HIRE	Employee growth rate
26	ILL	Illiquidity rolling (3 months)
27	LEV	Leverage
28	LGR	Growth in long-term debt
29	MAXRET	Maximum daily returns (3 months)
30	ME	Market equity
31	ME_IA	Industry-adjusted size
32	MOM12M	Cumulative Returns in the past (2-12) months
33	MOM1M	Previous month return
34	MOM36M	Cumulative Returns in the past (13-35) months
35	MOM60M	Cumulative Returns in the past (13-60) months
36	MOM6M	Cumulative Returns in the past (2-6) months
37	NI	Net Equity Issue
38	NINCR	Number of earnings increases
39	NOA	Net Operating Assets
40	OP	Operating profitability

Continue: Equity Characteristics (61 in total)

No.	Characteristics	Description
41	PCTACC	Percent operating accruals
42	PM	profit margin
43	PS	Performance Score
44	RD_SALE	R&D to sales
45	RDM	R&D Expense-to-market
46	RE	Revisions in analysts' earnings forecasts
47	RNA	Return on Net Operating Assets
48	ROA	Return on Assets
49	ROE	Return on Equity
50	RSUP	Revenue surprise
51	RVAR_CAPM	Residual variance - CAPM (3 months)
52	RVAR_FF3	Res. var. - Fama-French 3 factors (3 months)
53	RVAR_MEAN	Return variance (3 months)
54	SEAS1A	1-Year Seasonality
55	SGR	Sales growth
56	SP	Sales-to-price
57	STD_DOLVOL	Std of dollar trading volume (3 months)
58	STD_TURN	Std. of Share turnover (3 months)
59	SUE	Unexpected quarterly earnings
60	TURN	Shares turnover
61	ZEROTRADE	Number of zero-trading days (3 months)

Table A.2: Macro Predictors for Market Timing

This table lists the description for macro predictors used in the empirical study.

No.	Variable Name	Description
1	EP	Earnings-to-price of S&P 500
2	DY	Dividend yield of S&P 500
3	LEV	Leverage of S&P 500
4	NI	Net equity issuance of S&P 500
5	SVAR	Stock Variance of S&P 500
6	ILL	Pastor-Stambaugh illiquidity
7	INFL	Inflation
8	TBL	Three-month treasure bill rate
9	DFY	Default yield
10	TMS	Term spread

Figure A.1: Market-Adjusted P-Tree for the period from 1981 to 2000

The market-adjusted P-Tree trained from the period from 1981 to 2000 is displayed in this figure. All format follows Figure 4 .

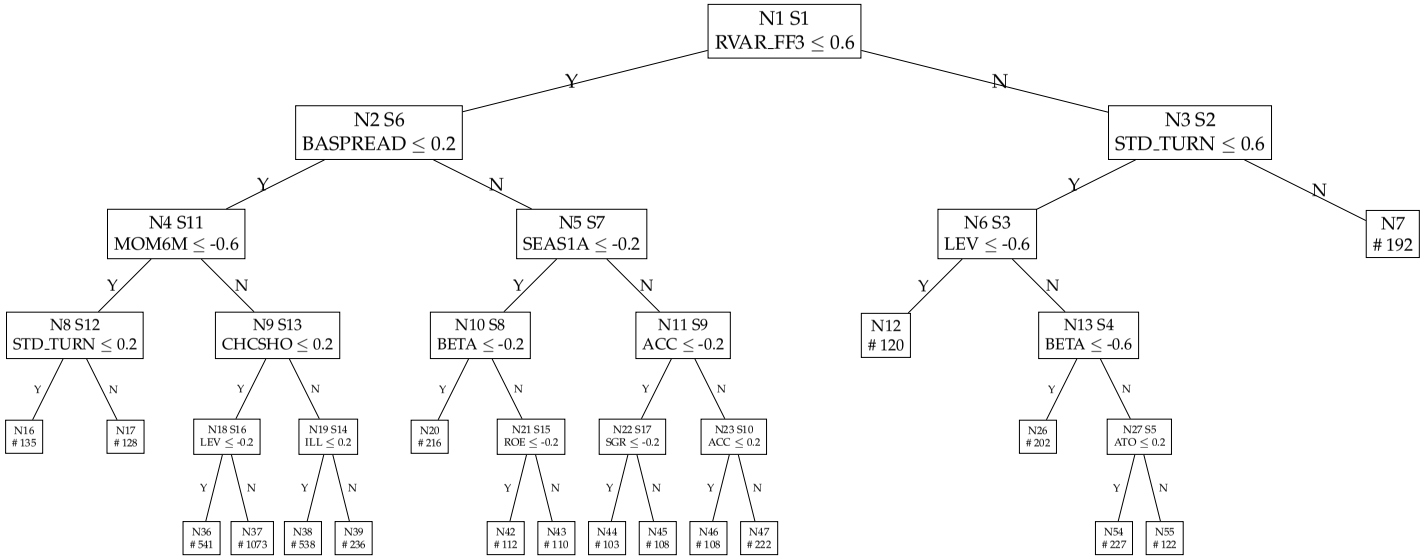
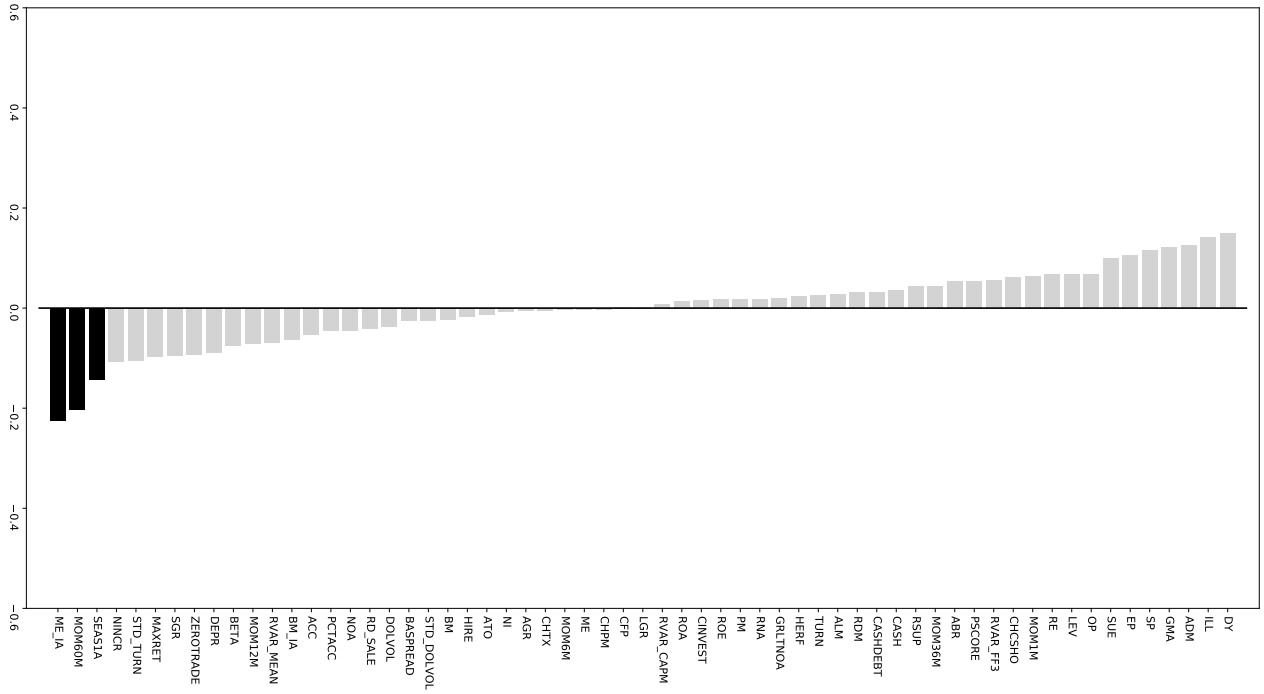
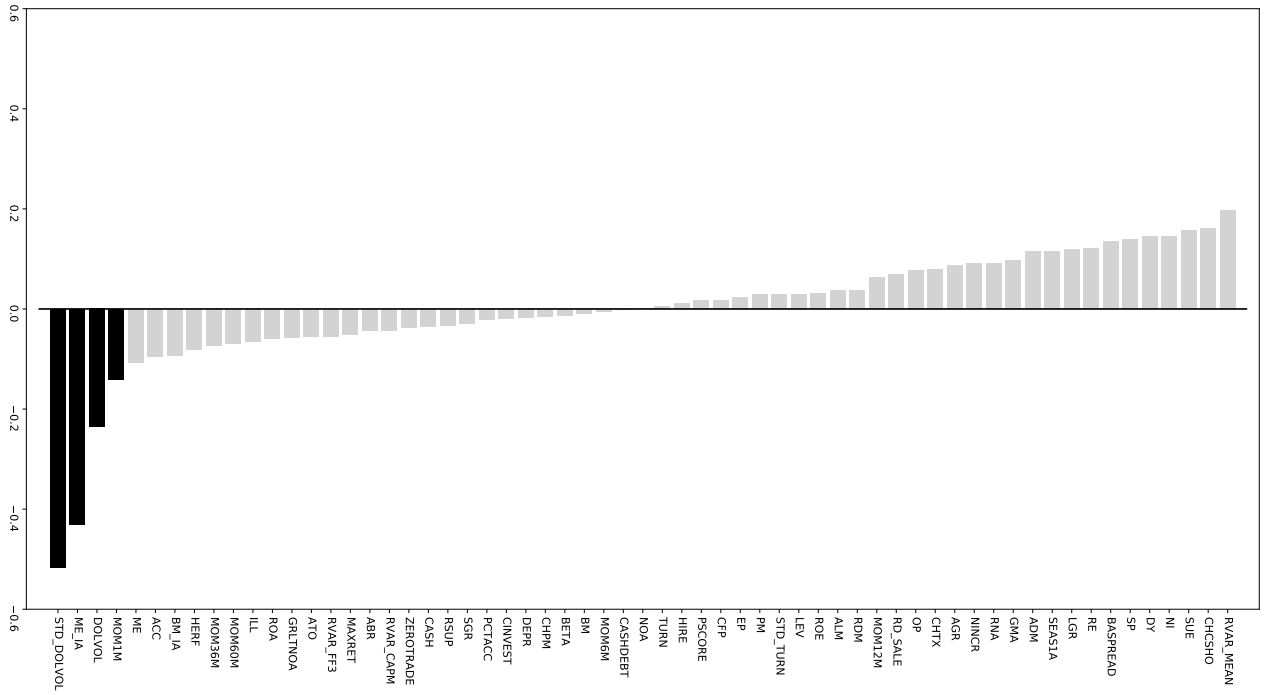


Figure A.2: Out-of-Bag Characteristics Significance: High/Low Inflation

This figure reports the characteristic significance by the out-of-bag ensembles from the random forest of 500 trees. We report results for high and low inflation periods in the train sample 1981-2000. This figure details follow Figure 9. The left dark columns indicate significant characteristics.



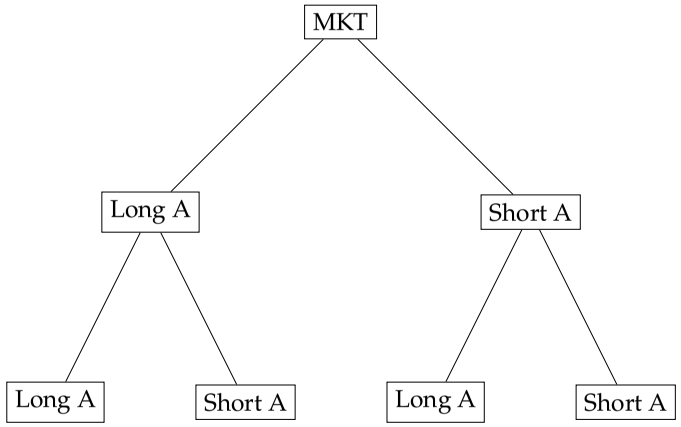
High INFL



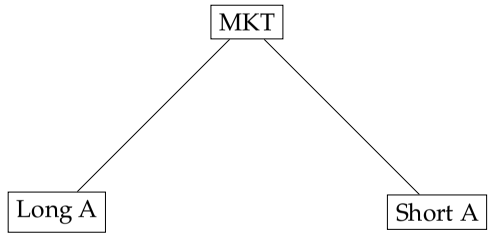
Low INFL

Figure A.3: Four Specifications of Characteristic-sorted Factors

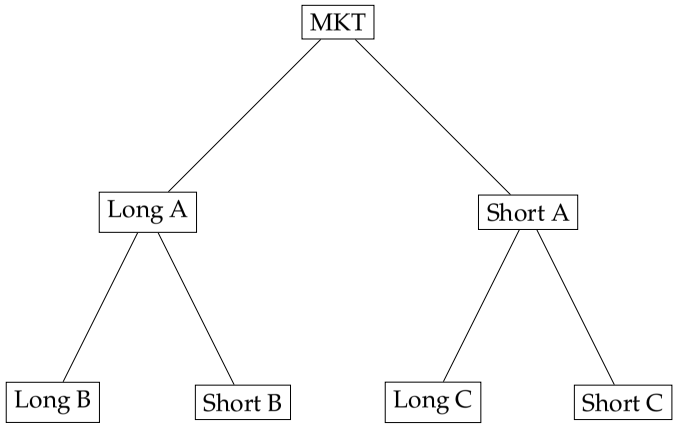
We show four different specifications for creating long-short factors by the tree model. Specification (a) is the classic sorting with a single characteristic and creates four 4x1 sorted portfolios. Specification (b) is similar and creates two 2x1 sorted portfolios. Specification (c) is trained by the Panel Tree model with interactions of characteristics. To each characteristic, we fit the Panel tree model to search optimal characteristics for its interaction by fixing the first split. Specification (d) is similar to (c) but fits with market-adjusted returns.



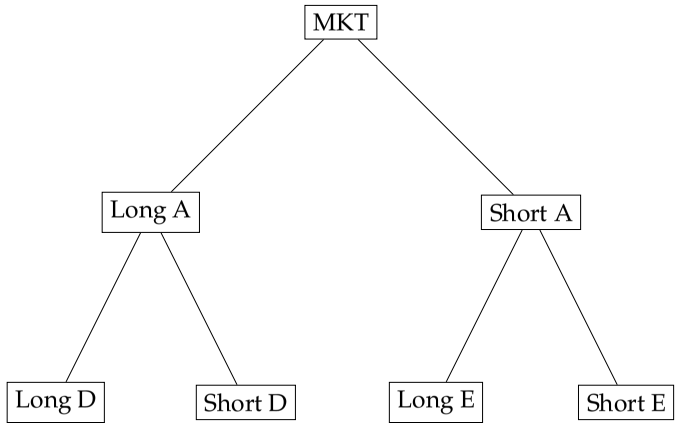
(a) Uni-Sort 4x1



(b) Uni-Sort 2x1



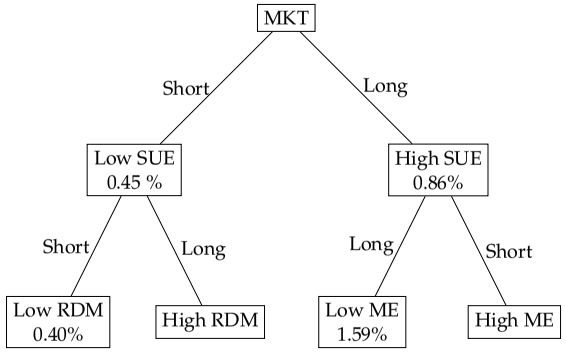
(c) Interaction



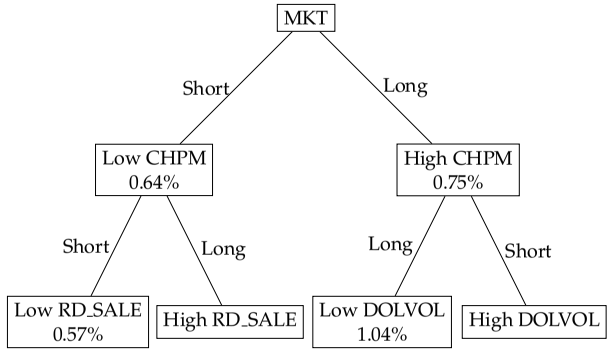
(d) Market-Adjusted Interaction

Figure A.4: Examples for Interaction Factors

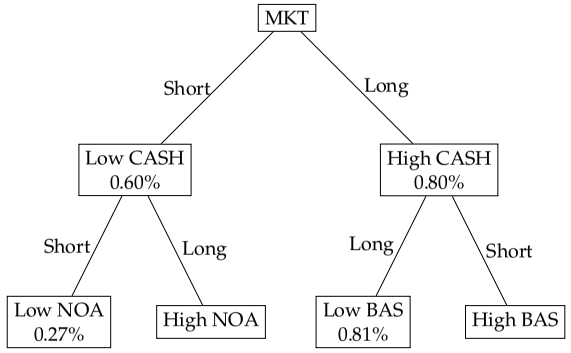
This figure shows six examples of the interaction factors. More details are reported in Table 6. In the second layer, the numbers report the average returns of the long portfolio and short portfolio. In the third layer, the numbers are the average returns of the long-long portfolio and short-short portfolio. One can see further splitting helps to create a higher return spread.



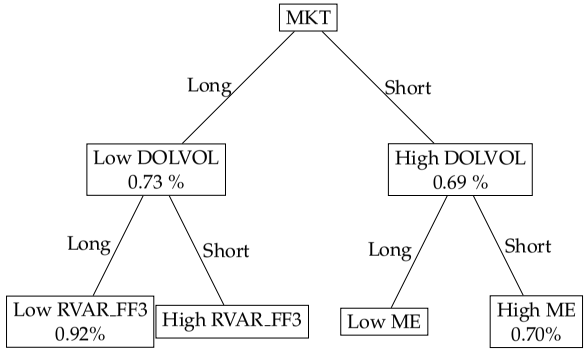
(a) Standardized Unexpected Earning



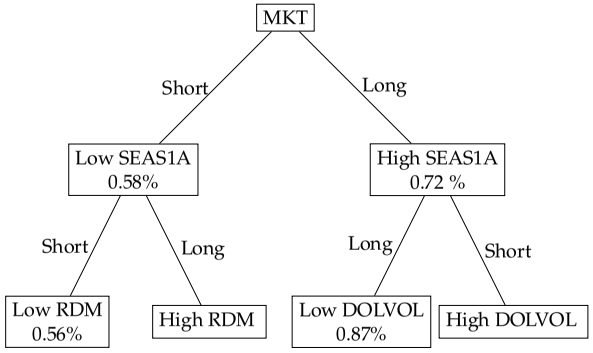
(b) Profit Margin



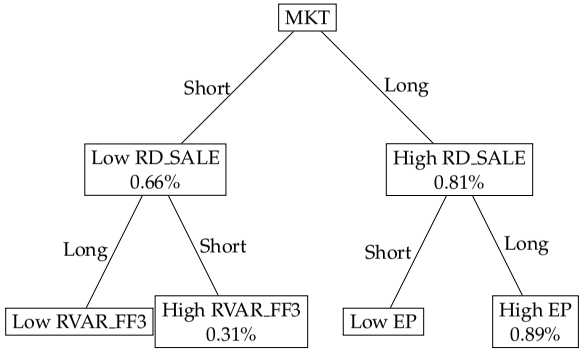
(c) Cash Holdings



(d) Dollar Trading Volume



(e) Seasonality



(f) R&D-to-Sales