Forest through the Trees: Building Cross-Sections of Stock Returns

Svetlana Bryzgalova, Markus Pelger, and Jason Zhu

Journal of Finance, ***2025***

解读人：赵伟皓

武汉大学金融系

2026.01.25

Research Question

How to construct the optimal test asset that can fully preserve the asset pricing information in the individual stock characteristics?

- Are test assets important? **Yes, and it's the core issue**.
- Does cross-section require tons of factors? No, true useful dimension is low;
- Is traditional sorting effective? No, it ignores nonlinearity and interaction;
- Is ML applicable for pricing? **Yes, but it must target SDF**;

AP Tree constructs cross-sections instead of a specific factor model.

Limitations of existing studies

Constructing test assets is a core issue that has not yet been fully explored.

1. **Limitations of traditional cross-sectional assets construction methods**
   - Single/double sorting, interaction between charateristics is cut off;
   - Sorting repeatedly uses same stock(information redundancy);
   - Sorting rules lack clear economic objectives(Spanning SDF).
2. **ML methods are not used for constructing pricing test assets**
   - RF, NN and other methods mainly dedicated to predicting stock returns;
   - Good predicting does not necessarily mean building an effective SDF;
3. **SDF is unobservable, high dimensional test assets may overfit severely**
   - Existing methods lack robust and operable assets construction mechanisms.

Importance of paper's research

1. **If the test assets are redundant or improperly constructed**
   - The pricing test results are unreliable;
   - Model comparison loses meaning; Factor significance may be unreliable.
2. **Most factors in 'factor zoo' are constructed based on traditional sorting**
   - These basis portfolios maybe biased and unable to span SDF;
   - Researches on sparse SDF may have biases due to problematic test assets; (Feng et al., 2020; Freyberger et al., 2020; Lettau et al., 2020; Kozak et al., 2020)
3. **Pricing focused machine learning contains economic interpretablity**
   - ML should not only be used for prediction, but should also serve clear economic goals.

Contribution

1. **Literature on Asset Pricing & Test Asset Construction**
   Prior: Directly estimating conditional SDF (not interpretable).(Chen et al., 2023)
   Extend: Building interpretable test assets that span SDF and capture interactions.

2. **Literature on Robust SDF Recovery & Shrinkage Estimation**
   Prior: Robust SDF recovery based on PCA and L2 regularization.(Kozak et al., 2020)
   Extend: Introduced **triple shrinkage**, Lasso, Ridge, Mean Shrinkage.

3. **Literature on return prediction and machine learning**
   Prior: ML models are prediction focused(accurate prediction means great pricing).
   Prior: Existing tree models use "local, greedy" pruning criteria.(Moritz et al., 2016)
   Extend: Proposing AP Trees + AP Pruning (Spanning SDF as objective).

Hypothesis

- H1: The Importance of Test Assets(Built by AP-trees);
  - Correct test assets lead to much higher Sharpe ratio, and maintain interpretability.
- H2: Pricing info of stock returns is ”low dimensional, nonlinear, interactive”;
- H3: Pricing focused ML method is superior to prediction focused methods.

Framework, Data, Sample

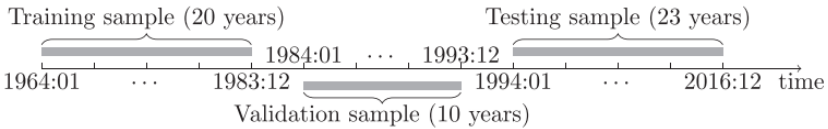
Tree Model

1. **AP Trees**: Using decision trees to recursively build candidate portfolios;
2. **AP Pruning (pruning based on SDF)**: SR maximization as objective.

Stock data (CRSP - Compustat)

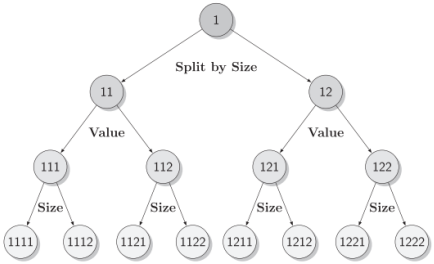
- 10 commonly used characteristics as sorting variables. (Size, B/M, etc.)

Sample division



Step 1: Tree Construction

1. **Traditional method (unconditional sorting):**
   - Grouping stocks by charateristics.(e.g B/M, etc.) Groups are independent.
2. **AP-Trees (conditional sorting)**:
   - 2.1 Cut all stocks by "Market Cap" (big vs small);
   - 2.2 In the group of "S" and "B", make another cut based on "value", respectively;
   - It captures interactions (performance of value factors in small cap stocks).



Panel A. Example of a conditional tree based on size and value

Step 2: Definition of Pruning

Traditional Pruning (Local Prediction):

- Greedy Algorithm: If splitting can reduce local MSE, then retain it.
- Ignoring the covariance structure.

AP Pruning (Global Pricing):

- Global Optimization: Nodes are kept when they improve Sharpe ratio.
- Considering the full covariance matrix.

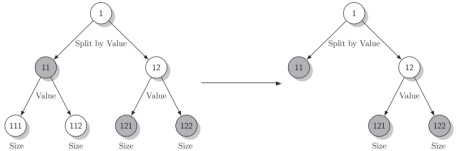


Figure 3. Illustration of pruning with multiple characteristics. The figure shows a sample subtree, original and pruned for portfolios of depth 2, and constructed based only on size and book-to-market characteristics. The fully pruned set of portfolios is based on the full tree with all possible combinations of splits. The right figure illustrates a potential outcome for one subtree

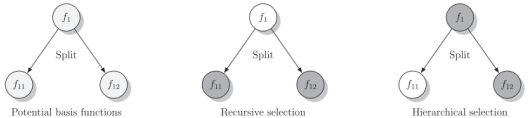


Figure 4. Equivalence of different representations of a tree. The figure shows two equivalent ways of selecting nodes in a split. Dark-colored nodes are selected, while white ones are not.

Step 3: Pruning Algorithm - Global Optimization

$$\text{Robust SDF:} \quad \min_w \frac{1}{2}(\hat{\mu}_{rob} - \hat{\Sigma}_{rob}w)'\hat{\Sigma}_{rob}^{-1}(\hat{\mu}_{rob} - \hat{\Sigma}_{rob}w) + \lambda_1||w||_1$$

$$\iff \text{MV Opitmization:} \quad \min_w \frac{1}{2}w^\top\hat{\Sigma}w + \underbrace{\lambda_1\|w\|_1}_{\text{LASSO}} + \underbrace{\frac{1}{2}\lambda_2\|w\|_2^2}_{\text{Ridge}}$$

$$\text{s.t.} \quad w^\top\mathbf{1} = 1, \qquad w^\top(\hat{\mu} + r_f\mathbf{1}) \geq \mu_0 + r_f$$

- L_1 norm: $\|w\|_1 = \sum_{i=1}^N |w_i|$; L_2 norm: $\|w\|_2^2 = \sum_{i=1}^N w_i^2$;
- $\lambda_0(\mu_0)$: **(Mean Shrinkage):** $(\hat{\mu}_{rob} \leftarrow \hat{\mu} + \lambda_0\mathbf{1})$.;
- λ_1(**LASSO**): Selecting dozens of core portfolios from thousands of nodes;
- λ_2 (**Ridge**)**:** Covariance Shrinkage ($\hat{\Sigma}_{rob} = \hat{\Sigma} + \lambda_2 I_N$), maintaining stablity.

Model Implementation

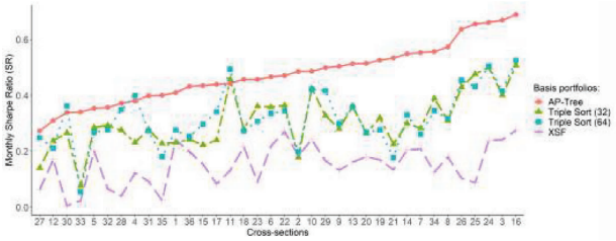
Train
- **(1). Building all potential AP Tree nodes (depth of 4)**;
- (2). Estimation of sample mean and covariance matrix;
- (3). Calculating portfolio weights through given $(\lambda_0, \lambda_1, \lambda_2)$;

Validation
- (1). Calculating SR using obtained weights;
- **(2). Selecting $(\lambda_0, \lambda_1, \lambda_2)$ that maximizes SR;**
- **(3). Tree structure and nodes(by Pruning);**

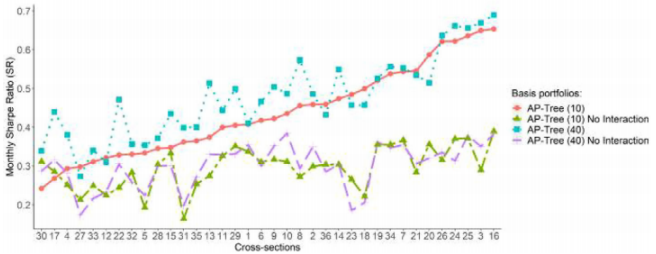
Test
- (1). Test using selected portfolios and parameters;
- (2). Evaluating out-of-sample performance (SR, α).

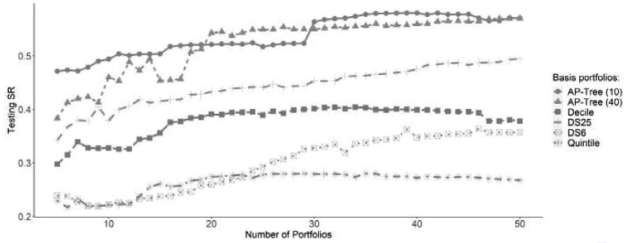
H1: AP Trees outperforms benchmark models

		Type of Cross-Section			
		AP Trees (10)	AP Trees (40)	TS (32)	TS(64)
SDF SR		0.65	0.69	0.51	0.53
α	FF3	0.94	0.90	0.75	0.84
		[10.11]	[11.03]	[7.40]	[8.13]
	FF5	0.81	0.76	0.47	0.61
		[8.76]	[9.60]	[5.57]	[6.73]
	XSF	0.81	0.76	0.46	0.61
		[8.77]	[9.46]	[5.39]	[6.69]
	FF11	0.89	0.80	0.37	0.65
		[9.12]	[9.60]	[4.29]	[6.91]
XS-R^2					
	FF3	18.0%	51.0%	82.0%	82.0%
	FF5	11.0%	64.0%	91.0%	90.0%
	XSF	28.0%	65.0%	91.0%	90.0%
	FF11	—	42.0%	92.0%	87.0%

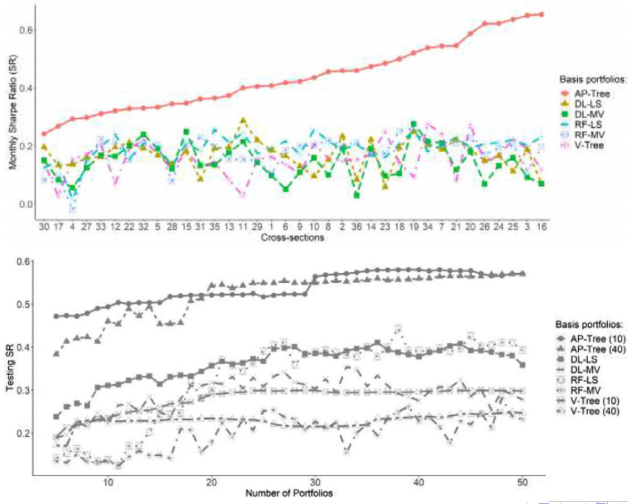


H2: Interaction, Nonlinearity, Low dimensionality





H3: Pricing focused ML outperforms prediction focused MLs



对 AP Tree 的目标函数进行拓展

1. 引入下行风险 (Downside Risk)，剪枝时，不再惩罚所有的波动，只惩罚“坏的波动”。
   - 那些波动大但主要是向上暴涨的组合（正偏度）将被保留，而经常暴跌的组合会被剔除。
2. 引入高阶矩 (Higher Moments): 基于共偏度（Coskewness）定价理论：
   - 目标是构建一个能解释“偏度溢价”的 SDF；
   - 在剪枝时，寻找的组合不仅要有高收益，还要能对冲市场的负偏度风险。

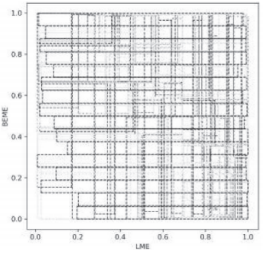
剪枝：论文使用 LASSO/Ridge 进行均值-方差收缩。

- 可以拓展为”Downside-Risk Regularization”。矩条件是 Lower Partial Moments。

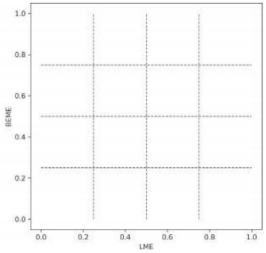
投资者并不是只关心方差，那么“为了最大化下行保护而分组”应该能产生更优的测试资产。

AP-trees vs Traditional Sorting

Dimension	AP Trees	Traditional Sorting
Sorting	**Conditional (Recursive)**	**Unconditional (Indep.)**
Structure	Sorts on B depend on prior sort A.	Intersects indep sorts (e.g., 5×5).
Correlation	Re-sorting within bins.	Causing *empty portfolios*.
Inference	**Interactive**. $(A \times B)$.	**Additive**. $(A + B)$.

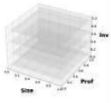


Panel B. Conditional trees

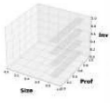


Panel C. Double-sorting

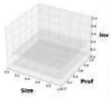
H1: AP Trees maintain economic interpretability



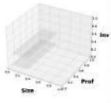
Panel A. The Market (1111.1)



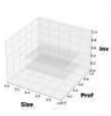
Panel B. Medium Size (50–75%) (1111.121)



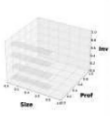
Panel C. Bottom 50% on Operating Profitability (1111.11)



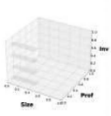
Panel D. Bottom 50% on Investment, Bottom 12.5% on Size (3111.11111)



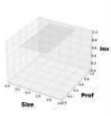
Panel E. Bottom 50% on Investment, Top 50% on Operating Profitability, Bottom 50% on Size (3211.1121)



Panel F. Bottom 50% on Size, Bottom 25% on Operating Profitability (1221.1111)



Panel G. Bottom 25% on Operating Profitability, Bottom 25% on Size (2221.11111)



Panel H. Top 12.5% on Investment, Bottom 50% on Size (3331.12221)

SDF 恢复等价于稳健切点组合（MV 角度）

SDF 恢复角度与稳健切点组合（均值方差角度）完全等价，对应关系如下：

DEFINITION 2（组合优化）	DEFINITION 1（SDF 恢复）
目标收益率约束μ_0	均值收缩参数λ_0
Ridge 惩罚$\lambda_2\|w\|_2^2$	方差收缩$\lambda_2 I_N$
LASSO 惩罚$\lambda_1\|w\|_1$	相同的 LASSO 惩罚
最优权重$\hat{w}_{\text{robust}}$	SDF 权重$\hat{\omega}_{\text{robust}} \propto \hat{w}_{\text{robust}}$

- 二者解决同一问题：DEF 1 从资产定价理论出发，DEF 2 从投资组合实践出发
- 经济意义相同：都是寻找最能代表市场投资机会的稳健、稀疏的基资产集合
- 数学等价：无论从哪个角度出发，最终选择的资产集合和 SDF 是相同的

A. 测试资产与随机贴现因子 (SDF)

为全部个股定价不可行 ⇒ 为特征组合（测试资产）定价。所选管理投资组合（测试资产）所**张成**的投资机会集，与**个股（给定特征下）**的投资机会集相同。

- **个股条件 SDF 投影:**

$$M_t^C = 1 - \sum_{i=1}^{N} b_{t-1,i}(R_{t,i} - \mathbb{E}_{t-1}[R_{t,i}])$$

$$b_{t-1,i} = f(C_{t-1,i})$$

其中 $f(\cdot)$ 是复杂（可能非线性）函数。

- **转化为无条件问题:**

$$M_t^C \approx 1 - \sum_{j=1}^{J} w_j \left(R_{t,j}^{\text{man}} - \mathbb{E}[R_{t,j}^{\text{man}}]\right)$$

$$R_{t,j}^{\text{man}} = \sum_{i=1}^{N} f_j(C_{t-1,i}) R_{t,i}$$

- **管理组合 vs. 基函数:** 组合权重 ↔ 特征空间中的基函数 $f_j(\cdot)$。
- 寻找一组管理组合，使其能最好地近似不可行的个股 SDF。
- **SDF 张成:** 最优测试资产组合应实现**最高夏普比率**，等价于最小化与真实 SDF 的距离。

如果测试资产未能张成 SDF，即使模型完美定价这些资产，也可能严重偏离为个股定价的真实 SDF。

B. 因遗漏测试资产导致的模型设定错误

如果使用的测试资产不完整，会发生什么？

- 真实 SDF 需要两组基函数张成：f^{select}（研究者使用）和 f^{omit}（被遗漏）。
- 对应两组管理组合：R^{select} 和 R^{omit}。
- 研究者仅基于 R^{select} 构建因子模型 F，并声称其完美定价 R^{select}（$\alpha^{\text{select}} = 0$）。

命题 1：遗漏测试资产导致的错误设定被遗漏资产 R^{omit} 的定价误差满足：（其中 $SR(\cdot)$ 为最高可实现夏普比率。）

$$SR^2(R^{\text{select}}, R^{\text{omit}}) - SR^2(F) \leq \alpha^{\text{omit}\top}(\Sigma^{\text{omit}})^{-1}\alpha^{\text{omit}} \leq SR^2(R^{\text{select}}, R^{\text{omit}}) - SR^2(R^{\text{select}})$$

含义 1

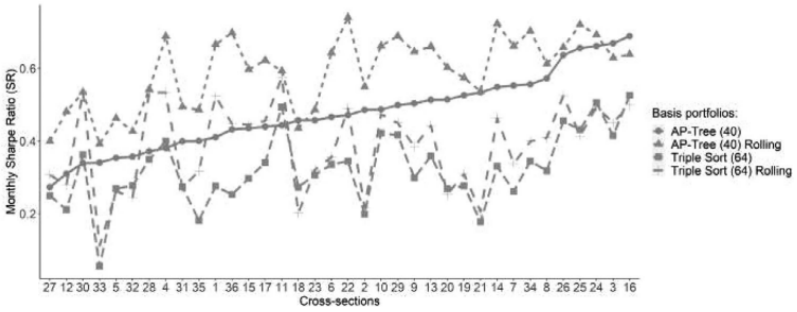
被遗漏资产的定价误差，与**完整资产集**与**部分资产集**的夏普比率平方差直接相关。

含义 2

若 R^{select} 本身未张成 SDF（$SR(R^{\text{select}})$ 较低），即使模型 F 完美定价 R^{select}，它对 R^{omit} 的定价误差也可能很大。

AP Trees 构建的资产集（包含传统组合 + 捕捉交互的组合）具有更高的 SR。

滚动窗口



作者在第 IV.F 节通过滚动窗口估计允许权重随时间变化，并发现这进一步提高了模型表现。