Asset Pricing with Panel Tree under Global Split Criteria

Lin William Cong[1] Guanhao Feng[2] Jingyu He[2] Xin He[3]

[1]Cornell University SC Johnson College of Business and NBER

[2]City University of Hong Kong

[3]Hunan University

Financial Big Data

Distinguishing features and ML solutions

1. High dimensionality (e.g., Cochrane, 2011).
   - Dimension reduction (e.g., Han et al., 2019; Kozak, Nagel, & Santoch, 2019; Feng, Giglio, & Xiu, 2020).
2. Nonlinearity (e.g., Harvey, Liu, & Zhu, 2015; Gu, Kelly, & Xiu, 2019).
   - e.g., splines, Freyberger, Neuhierl, & Weber (2019).
   - Deep NNs (e.g., Feng, Polson, & Xu, 2019; Fan et al., 2021).
3. Interaction versus sparsity.
   - e.g., Trees, Rossi (2018).
4. Low signal-to-noise (e.g., Martin & Nagel, 2019).
5. Non-stationarity/heteroskedasticity.
   - Memory and attention, e.g., Cong et al. (2021); Chen, Pelger, & Zhu. (2021).
6. Multi-sequence panel data.
   - e.g., CAAN, Cong et al. (2019).

Financial Big Data

Distinguishing features and ML solutions

1. High dimensionality (e.g., Cochrane, 2011).
   - Dimension reduction (e.g., Han et al., 2019; Kozak, Nagel, & Santoch, 2019; Feng, Giglio, & Xiu, 2020).
2. Nonlinearity (e.g., Harvey, Liu, & Zhu, 2015; Gu, Kelly, & Xiu, 2019).
   - e.g., splines, Freyberger, Neuhierl, & Weber (2019).
   - Deep NNs (e.g., Feng, Polson, & Xu, 2019; Fan et al., 2021).
3. Interaction versus sparsity.
   - e.g., Trees, Rossi (2018).
4. Low signal-to-noise (e.g., Martin & Nagel, 2019).
5. Non-stationarity/heteroskedasticity.
   - Memory and attention, e.g., Cong et al. (2021); Chen, Pelger, & Zhu. (2021).
6. Multi-sequence panel data.
   - e.g., CAAN, Cong et al. (2019).

Financial Big Data

Distinguishing features and ML solutions

1. High dimensionality (e.g., Cochrane, 2011).
   - Dimension reduction (e.g., Han et al., 2019; Kozak, Nagel, & Santoch, 2019; Feng, Giglio, & Xiu, 2020).
2. Nonlinearity (e.g., Harvey, Liu, & Zhu, 2015; Gu, Kelly, & Xiu, 2019).
   - e.g., splines, Freyberger, Neuhierl, & Weber (2019).
   - Deep NNs (e.g., Feng, Polson, & Xu, 2019; Fan et al., 2021).
3. Interaction versus sparsity.
   - e.g., Trees, Rossi (2018).
4. Low signal-to-noise (e.g., Martin & Nagel, 2019).
5. Non-stationarity/heteroskedasticity.
   - Memory and attention, e.g., Cong et al. (2021); Chen, Pelger, & Zhu. (2021).
6. Multi-sequence panel data.
   - e.g., CAAN, Cong et al. (2019).

Financial Big Data

Distinguishing features and ML solutions

1. High dimensionality (e.g., Cochrane, 2011).
   - Dimension reduction (e.g., Han et al., 2019; Kozak, Nagel, & Santoch, 2019; Feng, Giglio, & Xiu, 2020).
2. Nonlinearity (e.g., Harvey, Liu, & Zhu, 2015; Gu, Kelly, & Xiu, 2019).
   - e.g., splines, Freyberger, Neuhierl, & Weber (2019).
   - Deep NNs (e.g., Feng, Polson, & Xu, 2019; Fan et al., 2021).
3. Interaction versus sparsity.
   - e.g., Trees, Rossi (2018).
4. Low signal-to-noise (e.g., Martin & Nagel, 2019).
5. Non-stationarity/heteroskedasticity.
   - Memory and attention, e.g., Cong et al. (2021); Chen, Pelger, & Zhu. (2021).
6. Multi-sequence panel data.
   - e.g., CAAN, Cong et al. (2019).

Financial Big Data

Distinguishing features and ML solutions

1. High dimensionality (e.g., Cochrane, 2011).
   - Dimension reduction (e.g., Han et al., 2019; Kozak, Nagel, & Santoch, 2019; Feng, Giglio, & Xiu, 2020).
2. Nonlinearity (e.g., Harvey, Liu, & Zhu, 2015; Gu, Kelly, & Xiu, 2019).
   - e.g., splines, Freyberger, Neuhierl, & Weber (2019).
   - Deep NNs (e.g., Feng, Polson, & Xu, 2019; Fan et al., 2021).
3. Interaction versus sparsity.
   - e.g., Trees, Rossi (2018).
4. Low signal-to-noise (e.g., Martin & Nagel, 2019).
5. Non-stationarity/heteroskedasticity.
   - Memory and attention, e.g., Cong et al. (2021); Chen, Pelger, & Zhu. (2021).
6. Multi-sequence panel data.
   - e.g., CAAN, Cong et al. (2019).

Financial Big Data

Distinguishing features and ML solutions

1. High dimensionality (e.g., Cochrane, 2011).
   - Dimension reduction (e.g., Han et al., 2019; Kozak, Nagel, & Santoch, 2019; Feng, Giglio, & Xiu, 2020).
2. Nonlinearity (e.g., Harvey, Liu, & Zhu, 2015; Gu, Kelly, & Xiu, 2019).
   - e.g., splines, Freyberger, Neuhierl, & Weber (2019).
   - Deep NNs (e.g., Feng, Polson, & Xu, 2019; Fan et al., 2021).
3. Interaction versus sparsity.
   - e.g., Trees, Rossi (2018).
4. Low signal-to-noise (e.g., Martin & Nagel, 2019).
5. Non-stationarity/heteroskedasticity.
   - Memory and attention, e.g., Cong et al. (2021); Chen, Pelger, & Zhu. (2021).
6. Multi-sequence panel data.
   - e.g., CAAN, Cong et al. (2019).

Financial Big Data

Distinguishing features and ML solutions

1. High dimensionality (e.g., Cochrane, 2011).
   - Dimension reduction (e.g., Han et al., 2019; Kozak, Nagel, & Santoch, 2019; Feng, Giglio, & Xiu, 2020).
2. Nonlinearity (e.g., Harvey, Liu, & Zhu, 2015; Gu, Kelly, & Xiu, 2019).
   - e.g., splines, Freyberger, Neuhierl, & Weber (2019).
   - Deep NNs (e.g., Feng, Polson, & Xu, 2019; Fan et al., 2021).
3. Interaction versus sparsity.
   - e.g., Trees, Rossi (2018).
4. Low signal-to-noise (e.g., Martin & Nagel, 2019).
5. Non-stationarity/heteroskedasticity.
   - Memory and attention, e.g., Cong et al. (2021); Chen, Pelger, & Zhu. (2021).
6. Multi-sequence panel data.
   - e.g., CAAN, Cong et al. (2019).

Financial Big Data

Distinguishing features and ML solutions

1. High dimensionality (e.g., Cochrane, 2011).
   - Dimension reduction (e.g., Han et al., 2019; Kozak, Nagel, & Santoch, 2019; Feng, Giglio, & Xiu, 2020).
2. Nonlinearity (e.g., Harvey, Liu, & Zhu, 2015; Gu, Kelly, & Xiu, 2019).
   - e.g., splines, Freyberger, Neuhierl, & Weber (2019).
   - Deep NNs (e.g., Feng, Polson, & Xu, 2019; Fan et al., 2021).
3. Interaction versus sparsity.
   - e.g., Trees, Rossi (2018).
4. Low signal-to-noise (e.g., Martin & Nagel, 2019).
5. Non-stationarity/heteroskedasticity.
   - Memory and attention, e.g., Cong et al. (2021); Chen, Pelger, & Zhu. (2021).
6. Multi-sequence panel data.
   - e.g., CAAN, Cong et al. (2019).

Financial Big Data

Distinguishing features and ML solutions

1. High dimensionality (e.g., Cochrane, 2011).
   - Dimension reduction (e.g., Han et al., 2019; Kozak, Nagel, & Santoch, 2019; Feng, Giglio, & Xiu, 2020).
2. Nonlinearity (e.g., Harvey, Liu, & Zhu, 2015; Gu, Kelly, & Xiu, 2019).
   - e.g., splines, Freyberger, Neuhierl, & Weber (2019).
   - Deep NNs (e.g., Feng, Polson, & Xu, 2019; Fan et al., 2021).
3. Interaction versus sparsity.
   - e.g., Trees, Rossi (2018).
4. Low signal-to-noise (e.g., Martin & Nagel, 2019).
5. Non-stationarity/heteroskedasticity.
   - Memory and attention, e.g., Cong et al. (2021); Chen, Pelger, & Zhu. (2021).
6. Multi-sequence panel data.
   - e.g., CAAN, Cong et al. (2019).

Cornell University

Financial Big Data

Distinguishing features and ML solutions

1. High dimensionality (e.g., Cochrane, 2011).
   - Dimension reduction (e.g., Han et al., 2019; Kozak, Nagel, & Santoch, 2019; Feng, Giglio, & Xiu, 2020).
2. Nonlinearity (e.g., Harvey, Liu, & Zhu, 2015; Gu, Kelly, & Xiu, 2019).
   - e.g., splines, Freyberger, Neuhierl, & Weber (2019).
   - Deep NNs (e.g., Feng, Polson, & Xu, 2019; Fan et al., 2021).
3. Interaction versus sparsity.
   - e.g., Trees, Rossi (2018).
4. Low signal-to-noise (e.g., Martin & Nagel, 2019).
5. Non-stationarity/heteroskedasticity.
   - Memory and attention, e.g., Cong et al. (2021); Chen, Pelger, & Zhu. (2021).
6. Multi-sequence panel data.
   - e.g., CAAN, Cong et al. (2019).

Financial Big Data

Distinguishing features and ML solutions

1. High dimensionality (e.g., Cochrane, 2011).
   - Dimension reduction (e.g., Han et al., 2019; Kozak, Nagel, & Santoch, 2019; Feng, Giglio, & Xiu, 2020).
2. Nonlinearity (e.g., Harvey, Liu, & Zhu, 2015; Gu, Kelly, & Xiu, 2019).
   - e.g., splines, Freyberger, Neuhierl, & Weber (2019).
   - Deep NNs (e.g., Feng, Polson, & Xu, 2019; Fan et al., 2021).
3. Interaction versus sparsity.
   - e.g., Trees, Rossi (2018).
4. Low signal-to-noise (e.g., Martin & Nagel, 2019).
5. Non-stationarity/heteroskedasticity.
   - Memory and attention, e.g., Cong et al. (2021); Chen, Pelger, & Zhu. (2021).
6. Multi-sequence panel data.
   - e.g., CAAN, Cong et al. (2019).

Financial Big Data

Distinguishing features and ML solutions

1. High dimensionality (e.g., Cochrane, 2011).
   - Dimension reduction (e.g., Han et al., 2019; Kozak, Nagel, & Santoch, 2019; Feng, Giglio, & Xiu, 2020).
2. Nonlinearity (e.g., Harvey, Liu, & Zhu, 2015; Gu, Kelly, & Xiu, 2019).
   - e.g., splines, Freyberger, Neuhierl, & Weber (2019).
   - Deep NNs (e.g., Feng, Polson, & Xu, 2019; Fan et al., 2021).
3. Interaction versus sparsity.
   - e.g., Trees, Rossi (2018).
4. Low signal-to-noise (e.g., Martin & Nagel, 2019).
5. Non-stationarity/heteroskedasticity.
   - Memory and attention, e.g., Cong et al. (2021); Chen, Pelger, & Zhu. (2021).
6. Multi-sequence panel data.
   - e.g., CAAN, Cong et al. (2019).

Building ML and AI Models for Finance

1. Economic motivation for ML/AI models.
2. Interpretability and transparency.
    - For new theories and models.
    - Applicability and guarding against overfitting.
    - For policymakers, regulators, and practitioners.
    - Causality: e.g., Athey & Wager (2019); [illegible]
    - Explainable AI, Distillation, etc.

- Asset pricing and investments
    - Prediction exercises with no economic guidance or inference.
    - Economically motivated supervised learning.
- Corporate Finance applications
    - Textual analysis: Hoberg and Phillips (2016); Li et al (2020); ....
    - ML in Corporate finance: Erel et al. (2021); Lyonnet and Stern (2022).

Building ML and AI Models for Finance

1. Economic motivation for ML/AI models.
2. Interpretability and transparency.
   - For new theories and models.
   - Applicability and guarding against overfitting.
   - For policymakers, regulators, and practitioners.
   - Causality: e.g., Athey & Wager (2019); causal BERT;...
   - Explainable AI, Distillation, etc.

- Asset pricing and investments
  - Prediction exercises with no economic guidance or inference.
  - Economically motivated supervised learning.
- Corporate Finance applications
  - Textual analysis: Hoberg and Phillips (2016); Li et al (2020); ....
  - ML in Corporate finance: Erel et al. (2021); Lyonnet and Stern (2022).

Building ML and AI Models for Finance

1. Economic motivation for ML/AI models.
2. Interpretability and transparency.
   - For new theories and models.
   - Applicability and guarding against overfitting.
   - For policymakers, regulators, and practitioners.
   - Causality: e.g., Athey & Wager (2019); causal BERT;...
   - Explainable AI, Distillation, etc.

- Asset pricing and investments
  - Prediction exercises with no economic guidance or inference.
  - Economically motivated supervised learning.
- Corporate Finance applications
  - Textual analysis: Hoberg and Phillips (2016); Li et al (2020); ....
  - ML in Corporate finance: Erel et al. (2021); Lyonnet and Stern (2022).

Building ML and AI Models for Finance

1. Economic motivation for ML/AI models.
2. Interpretability and transparency.
   - For new theories and models.
   - Applicability and guarding against overfitting.
   - For policymakers, regulators, and practitioners.
   - Causality: e.g., Athey & Wager (2019); causal BERT;...
   - Explainable AI, Distillation, etc.

- Asset pricing and investments
  - Prediction exercises with no economic guidance or inference.
  - Economically motivated supervised learning.
- Corporate Finance applications
  - Textual analysis: Hoberg and Phillips (2016); Li et al (2020); ....
  - ML in Corporate finance: Erel et al. (2021); Lyonnet and Stern (2022).

Building ML and AI Models for Finance

1. Economic motivation for ML/AI models.
2. Interpretability and transparency.
   - For new theories and models.
   - Applicability and guarding against overfitting.
   - For policymakers, regulators, and practitioners.
   - Causality: e.g., Athey & Wager (2019); causal BERT;...
   - Explainable AI, Distillation, etc.

- Asset pricing and investments
  - Prediction exercises with no economic guidance or inference.
  - Economically motivated supervised learning.
- Corporate Finance applications
  - Textual analysis: Hoberg and Phillips (2016); Li et al (2020); ....
  - ML in Corporate finance: Erel et al. (2021); Lyonnet and Stern (2022).

Building ML and AI Models for Finance

1. Economic motivation for ML/AI models.
2. Interpretability and transparency.
   - For new theories and models.
   - Applicability and guarding against overfitting.
   - For policymakers, regulators, and practitioners.
   - Causality: e.g., Athey & Wager (2019); causal BERT;...
   - Explainable AI, Distillation, etc.

- Asset pricing and investments
  - Prediction exercises with no economic guidance or inference.
  - Economically motivated supervised learning.
- Corporate Finance applications
  - Textual analysis: Hoberg and Phillips (2016); Li et al (2020); ....
  - ML in Corporate finance: Erel et al. (2021); Lyonnet and Stern (2022).

Building ML and AI Models for Finance

1. Economic motivation for ML/AI models.
2. Interpretability and transparency.
   - For new theories and models.
   - Applicability and guarding against overfitting.
   - For policymakers, regulators, and practitioners.
   - Causality: e.g., Athey & Wager (2019); causal BERT;...
   - Explainable AI, Distillation, etc.

- Asset pricing and investments
  - Prediction exercises with no economic guidance or inference.
  - Economically motivated supervised learning.
- Corporate Finance applications
  - Textual analysis: Hoberg and Phillips (2016); Li et al (2020); ....
  - ML in Corporate finance: Erel et al. (2021); Lyonnet and Stern (2022).

Building ML and AI Models for Finance

1. Economic motivation for ML/AI models.
2. Interpretability and transparency.
   - For new theories and models.
   - Applicability and guarding against overfitting.
   - For policymakers, regulators, and practitioners.
   - Causality: e.g., Athey & Wager (2019); causal BERT;...
   - Explainable AI, Distillation, etc.

- Asset pricing and investments
  - Prediction exercises with no economic guidance or inference.
  - Economically motivated supervised learning.
- Corporate Finance applications
  - Textual analysis: Hoberg and Phillips (2016); Li et al (2020); ....
  - ML in Corporate finance: Erel et al. (2021); Lyonnet and Stern (2022).

AI Beyond Basic ML: Goal-Oriented Search

1. Automation of repeated physical solutions/processes:
   - Industrial revolution (1750-1850) and Machine Age (1870-1940).
2. Automation of repeated mental/computational solutions/processes:
   - Digital revolution (1950-now) and Information Age.
3. Let machines find solutions themselves.
   - Artificial Intelligence.

- Instead of training through examples (supervised learning), we want to specify a problem and/or goal.
- Requires learning autonomously how to make decisions to achieve goals: essentially a search problem.

AI Beyond Basic ML: Goal-Oriented Search

1. Automation of repeated physical solutions/processes:
   - Industrial revolution (1750-1850) and Machine Age (1870-1940).
2. Automation of repeated mental/computational solutions/processes:
   - Digital revolution (1950-now) and Information Age.
3. Let machines find solutions themselves.
   - Artificial Intelligence.

- Instead of training through examples (supervised learning), we want to specify a problem and/or goal.
- Requires learning autonomously how to make decisions to achieve goals: essentially a search problem.
- Heuristic search (Deep RL and PSA for portfolio management).
- **Greedy search** (panel trees for latent factor asset pricing and uncommon factors for Bayesian asset clusters..

(Deep) Reinforcement Learning as Heuristic Search

The Reward Hypothesis: *Any goal can be formalized as the outcome of maximizing a cumulative reward*.

People learn by **interacting with the environment** in an active and sequential way, to optimize some **rewards**.

1. Fly a helicopter
   - Reward: air time, inverse distance, ...
2. Make a robot walk
   - Reward: distance, speed, ...
3. Play games
   - Reward: win, maximize scores, ...
4. Manage portfolio
   - Reward: returns, Sharpe ratio, ...

- Reward, Value, Policy (Actions).
- Agents: Value-based, Policy-based, Actor Critic, etc.

A Deep RL and XAI Example

"**AlphaPortfolio: Direct Construction Through Deep Reinforcement Learning and Interpretable AI**" Cong, Tang, Wang, & Zhang (2019).

- Why deep reinforcement learning (RL)?
  - Alternative, data-driven, flexible approach for direct optimization.
    - RL: trial-and-error search and delayed rewards (Sutton & Barto, 2017); works well for unlabeled data.
    - Possible interaction with state variables and environments.
    - Offline RL is the most active in AI/CS over the past 5-10 years.
  - AI tailored to portfolio management with superb performance and robustness to economic restrictions.

A Deep RL and XAI Example

"**AlphaPortfolio: Direct Construction Through Deep Reinforcement Learning and Interpretable AI**" Cong, Tang, Wang, & Zhang (2019).

- Why deep reinforcement learning (RL)?
  - Alternative, data-driven, flexible approach for direct optimization.
    - RL: trial-and-error search and delayed rewards (Sutton & Barto, 2017); works well for unlabeled data.
    - Possible interaction with state variables and environments.
    - Offline RL is the most active in AI/CS over the past 5-10 years.
  - AI tailored to portfolio management with superb performance and robustness to economic restrictions.

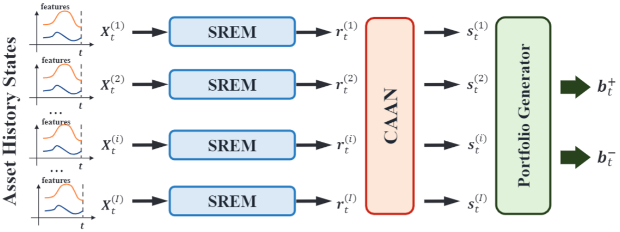
- Economic distillation for interpretable AI:
  - Big data and black-box models: feature selection or performance diagnostics.
  - Explainable AI (XAI): feature importance extraction vs surrogates; instance-based, compression/distillation, etc.
  - Polynomial sensitivity and textual factor analyses: Drivers for portfolio performance and construction choices.
  - Interpretable and extendable tools: projections onto linear modeling and textual spaces

Architecture of AlphaPortfolio

Sequence Representation Extraction Modules:

- Sequence learning in AP (Cong et al., 2020): RNN $\rightarrow$ LSTM $\rightarrow$ Bi-LSTM $\rightarrow$ RNN with Attention $\rightarrow$ **Transformer** (TE) or **Bi-LSTM-HA**.
- History states in look-back window: $\boldsymbol{X}^{(i)} = \left\{ \boldsymbol{x}_1^{(i)}, \ldots, \boldsymbol{x}_K^{(i)} \right\}$.

- Cross-Asset Attention Network (CAAN)
  - Built on self-attention mechanism (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, & Polosukhin, 2017).



Cornell University

AlphaPortfolio Performance on Test Sample

	AP Performance				AP Excess Alpha					
	(1)	(2)	(3)		(4)	(5)	(6)	(7)	(8)	(9)
Firms	All	$> q_{10}$	$> q_{20}$	Factor Models	All		$> q_{10}$		$> q_{20}$	
					$\alpha(\%)$	R^2	$\alpha(\%)$	R^2	$\alpha(\%)$	R^2
Return (%)	17.00	17.10	18.10	CAPM	13.9***	0.005	12.2***	0.088	14.0***	0.102
Std.Dev. (%)	8.50	7.70	8.20	FFC	14.2***	0.052	13.4***	0.381	14.7***	0.465
Sharpe	2.00	2.31	2.21	FFC+PS	13.7***	0.054	12.3***	0.392	13.3***	0.480
Skewness	1.42	1.74	1.91	FF5	15.3***	0.12	13.8***	0.426	14.7***	0.435
Kurtosis	6.33	5.70	5.97	FF6	15.6***	0.128	14.5***	0.459	15.8***	0.516
Turnover	0.26	0.24	0.26	SY	17.4***	0.037	15.8***	0.332	17.0***	0.394
MDD	0.08	0.02	0.02	Q4	16.0***	0.121	15.0***	0.495	16.2***	0.521

Robust to adding economic restrictions and using alternative objectives.
Projection onto linear modeling and natural language spaces.

Panel Tree as Goal-Oriented Greedy Search

Asset Pricing with P-Tree Under Global Split Criteria

- Common factors are used to describe returns and average returns.
- Market Factor, Fama-French-Type Factors, time-varying loadings.
- Machine Learning Methods:
  - Penalized regressions, PCAs, or Deep Learning to generate the stochastic discount factor using multiple firm characteristics.
- Panel Trees with an Application for Asset Pricing:
  - Interpretable (e.g., single decision tree) ML method that suits financial big data.
  - Generate test portfolios that better span the efficient frontier.
  - Guided by economic principles and designed for panel settings (e.g., can accommodate regime-shifts) and factor models for individual AP.

Asset Pricing with P-Tree Under Global Split Criteria

- Common factors are used to describe returns and average returns.
- Market Factor, Fama-French-Type Factors, time-varying loadings.
- Machine Learning Methods:
  - Penalized regressions, PCAs, or Deep Learning to generate the stochastic discount factor using multiple firm characteristics.
- Panel Trees with an Application for Asset Pricing:
  - Interpretable (e.g., single decision tree) ML method that suits financial big data.
  - Generate test portfolios that better span the efficient frontier.
  - Guided by economic principles and designed for panel settings (e.g., can accommodate regime-shifts) and factor models for individual AP.

Asset Pricing with P-Tree Under Global Split Criteria

- Common factors are used to describe returns and average returns.
- Market Factor, Fama-French-Type Factors, time-varying loadings.
- Machine Learning Methods:
  - Penalized regressions, PCAs, or Deep Learning to generate the stochastic discount factor using multiple firm characteristics.
- Panel Trees with an Application for Asset Pricing:
  - Interpretable (e.g., single decision tree) ML method that suits financial big data.
  - Generate test portfolios that better span the efficient frontier.
  - Guided by economic principles and designed for panel settings (e.g., can accommodate regime-shifts) and factor models for individual AP.

Asset Pricing with P-Tree Under Global Split Criteria

- Common factors are used to describe returns and average returns.
- Market Factor, Fama-French-Type Factors, time-varying loadings.
- Machine Learning Methods:
  - Penalized regressions, PCAs, or Deep Learning to generate the stochastic discount factor using multiple firm characteristics.
- Panel Trees with an Application for Asset Pricing:
  - Interpretable (e.g., single decision tree) ML method that suits financial big data.
  - Generate test portfolios that better span the efficient frontier.
  - Guided by economic principles and designed for panel settings (e.g., can accommodate regime-shifts) and factor models for individual AP.

Asset Pricing with P-Tree Under Global Split Criteria

- Common factors are used to describe returns and average returns.
- Market Factor, Fama-French-Type Factors, time-varying loadings.
- Machine Learning Methods:
  - Penalized regressions, PCAs, or Deep Learning to generate the stochastic discount factor using multiple firm characteristics.
- Panel Trees with an Application for Asset Pricing:
  - Interpretable (e.g., single decision tree) ML method that suits financial big data.
  - Generate test portfolios that better span the efficient frontier.
  - Guided by economic principles and designed for panel settings (e.g., can accommodate regime-shifts) and factor models for individual AP.

Asset Pricing with P-Tree Under Global Split Criteria

- Common factors are used to describe returns and average returns.
- Market Factor, Fama-French-Type Factors, time-varying loadings.
- Machine Learning Methods:
  - Penalized regressions, PCAs, or Deep Learning to generate the stochastic discount factor using multiple firm characteristics.
- Panel Trees with an Application for Asset Pricing:
  - Interpretable (e.g., single decision tree) ML method that suits financial big data.
  - Generate test portfolios that better span the efficient frontier.
  - Guided by economic principles and designed for panel settings (e.g., can accommodate regime-shifts) and factor models for individual AP.

Asset Pricing with P-Tree Under Global Split Criteria

- Common factors are used to describe returns and average returns.
- Market Factor, Fama-French-Type Factors, time-varying loadings.
- Machine Learning Methods:
  - Penalized regressions, PCAs, or Deep Learning to generate the stochastic discount factor using multiple firm characteristics.
- Panel Trees with an Application for Asset Pricing:
  - Interpretable (e.g., single decision tree) ML method that suits financial big data.
  - Generate test portfolios that better span the efficient frontier.
  - Guided by economic principles and designed for panel settings (e.g., can accommodate regime-shifts) and factor models for individual AP.

Asset Pricing with P-Tree Under Global Split Criteria

- Common factors are used to describe returns and average returns.
- Market Factor, Fama-French-Type Factors, time-varying loadings.
- Machine Learning Methods:
  - Penalized regressions, PCAs, or Deep Learning to generate the stochastic discount factor using multiple firm characteristics.
- Panel Trees with an Application for Asset Pricing:
  - Interpretable (e.g., single decision tree) ML method that suits financial big data.
  - Generate test portfolios that better span the efficient frontier.
  - Guided by economic principles and designed for panel settings (e.g., can accommodate regime-shifts) and factor models for individual AP.

Motivation: Conditional Stochastic Discount Factor Model

- Explain cross-sectional difference for individual stock returns

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

- A tradable SDF:

$$m_{t+1}=1-w_t^{\top} r_{t+1}=\Sigma_i f(z_{i,t}) R_{i,t+1}, \quad w_t=E_t\left[r_{t+1} r_{t+1}^{\top}\right]^{-1} E_t\left[r_{t+1}\right]$$

Hard to estimate for high dimensional **individual stocks**.

- Researchers use **basis portfolio** (FF 25, industry, etc) instead

$$m_{t+1}=1-W_t R_{t+1}, \quad W_t=E_t\left[R_{t+1} R_{t+1}^{\top}\right]^{-1} E_t\left[R_{t+1}\right], \quad R_{t+1,j}=\Sigma_i f_j(z_{i,t}) R_{i,t+1}.$$

Cornell University

Motivation: Conditional Stochastic Discount Factor Model

- Explain cross-sectional difference for individual stock returns

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

- A tradable SDF:

$$m_{t+1} = 1 - w_t^{\intercal} r_{t+1} = \Sigma_i f(z_{i,t}) R_{i,t+1}, \quad w_t = E_t\left[r_{t+1} r_{t+1}^{\intercal}\right]^{-1} E_t\left[r_{t+1}\right]$$

Hard to estimate for high dimensional **individual stocks**.

- Researchers use **basis portfolio** (FF 25, industry, etc) instead

$$m_{t+1} = 1 - W_t R_{t+1}, \quad W_t = E_t\left[R_{t+1} R_{t+1}^{\intercal}\right]^{-1} E_t\left[R_{t+1}\right], \quad R_{t+1,j} = \Sigma_i f_j(z_{i,t}) R_{i,t+1}.$$

Motivation: Conditional Stochastic Discount Factor Model

- Explain cross-sectional difference for individual stock returns

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

- A tradable SDF:

$$m_{t+1} = 1 - w_t^\intercal r_{t+1} = \Sigma_i f(z_{i,t}) R_{i,t+1}, \quad w_t = E_t\left[r_{t+1} r_{t+1}^\intercal\right]^{-1} E_t\left[r_{t+1}\right]$$

Hard to estimate for high dimensional **individual stocks**.

- Researchers use **basis portfolio** (FF 25, industry, etc) instead

$$m_{t+1} = 1 - W_t R_{t+1}, \quad W_t = E_t\left[R_{t+1} R_{t+1}^\intercal\right]^{-1} E_t\left[R_{t+1}\right], \quad R_{t+1,j} = \Sigma_i f_j(z_{i,t}) R_{i,t+1}.$$

Conditional SDF and Factor Construction from Basis Portfolios

- Time-varying factor loadings and reduced-form estimation using asset characteristics:

$$\beta_{i,t} = \frac{\mathrm{Cov}_t\left(W_t R_{t+1}, r_{i,t+1}\right)}{\mathrm{Var}_t\left(W_t R_{t+1}\right)} = b_0 + b_1^{\intercal} z_{i,t},$$

- FF construct factors by dividing stock universe into six non-overlapped groups.

- SMB and HML are (long-short) portfolios on these six portfolios.

$$SMB = \frac{1}{3}(SV + SM + SG) - \frac{1}{3}(BV + BN + BG)$$
$$HML = \frac{1}{2}(SV + BV) - \frac{1}{2}(SG + BG)$$

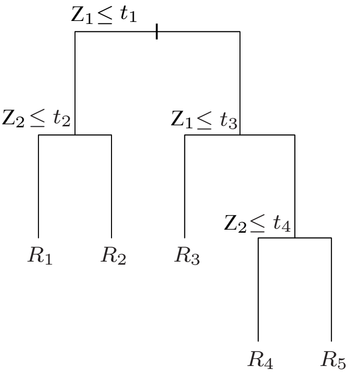
- Assets in the same group **behave similarly given similar risk exposures**.

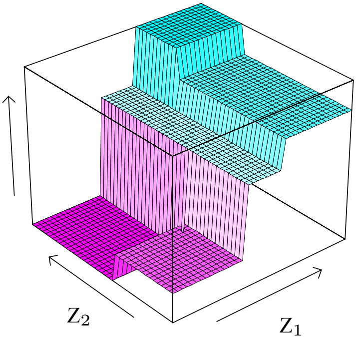
Why decision tree?

- Advantage 1: Generalized conditional sorts **greedy** search instead of costly enumeration of all possible basis portfolios.
- Advantage 2: **Interpretable** ML learns nonlinear interactions and higher order effects of high-dimensional variables.
- Advantage 3: Adaptive to the **low signal-to-noise** environment through data value averaging, ensembles, and error minimization as criterion.
- Advantage 4: Asymptotic normality, unbiasedness, and consistency (Scornet, Biau, & Vert, 2015; Wager, 2016; Athey and Wager, 2018).
- Disadvantages of CART (Breiman et al., 1984) and variants:
  - Constant pricing kernel, assume returns are *i.i.d.*; no time-series splits.
  - Recursion, each leaf splits locally, without any economic consideration.
  - Ensembles not so interpretable; single tree overfits.
- **P-Tree**: More interpretable and flexible class of tree models tailored for AP applications, generating both leaf test portfolios and SDF in a top-down approah.

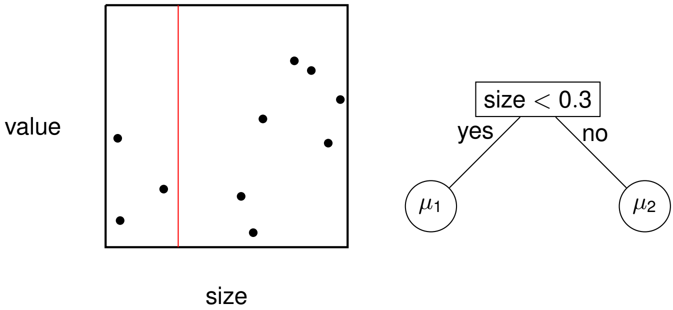
Traditional Regression Trees: Intuition

Hierarchical: use less and less data $\rightarrow$ overfit.



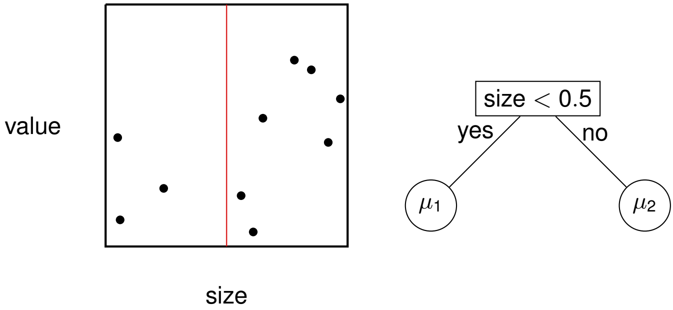


CART: search for optimal split points



- Consider a tentative split point for capturing the cross-sectional variation; similar to sorting.
- Similar to sorting!
- Loop over all possible split points (all variables, all values)

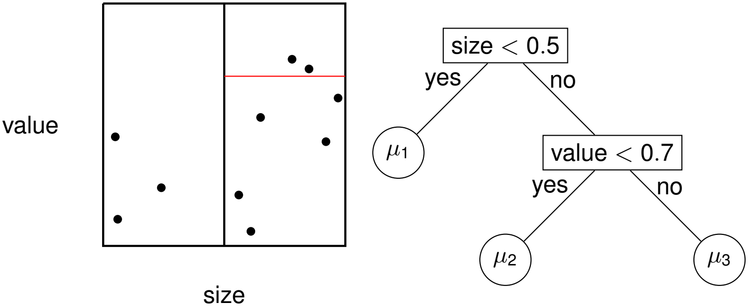
CART: search for optimal split points



- Pick one to optimize the split criterion.
- CART split criterion minimizes L^2 loss or **pricing errors** using a constant pricing kernel:

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

CART grows recursively



- CART assumes observations are *i.i.d.*, which is generally not true for asset return panel data.
- CART grows a tree recursively using local split criterion.
- Easy coding, fast computing, but not crucial or desirable for asset pricing.

Panel Tree (P-Tree) for Asset Pricing

- We use P-Tree to generate factor and use factor to grow P-Tree.
- The squared sum of **pricing errors** is the split criterion.

$$\sum_t \sum_i (r_{i,t} - \mu_{i,t}^{(k)})^2$$

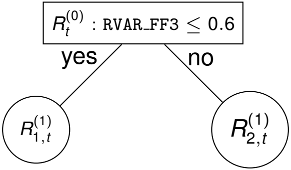
$$\mu_{i,t}^{(k)} = \beta_i f_t^{(k)}$$

- $f_t^{(k)}$ is the factor generated after the k-th split. It is defined using all leaf portfolios.
- The tree has to have a **vectorized outcome** indicating returns of different time periods. But the tree structure models all time periods.
- The split criterion is global, thus the tree has to grow **iteratively**; nevertheless, the greedy search avoids NP hard problems.

Panel Tree Factor Model

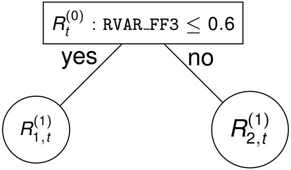
Panel Tree Factor Model: Step I

Consider a split point candidate



- Before splitting, $R_t^{(0)}$ denote the vector of market returns (value weighted portfolio) at the root node.
- $R_{j,t}^{(k)}$ is the leaf-basis portfolio of the j-th terminal node after the k-th split.
- The time series for leaf-basis portfolios can be value / equally weighted – the panel data structure for returns.

Panel Tree Factor Model: Step II



- **Estimate** the SDF $f_t^{(1)}$ based on leaf basis portfolios, a mean-variance efficient portfolio for $R_t^{(1)} = [R_{1,t}^{(1)}, R_{2,t}^{(1)}]$.

$$f_t^{(1)} = \widehat{\Sigma}_1^{-1} \widehat{\mu}_1 R_t^{(1)} = w_{11} R_{1,t}^{(1)} + w_{12} R_{2,t}^{(1)}.$$

- Each **split point candidate** partitions the cross section of individual stocks, providing different leaf basis portfolios and the resulting SDF.

Panel Tree Factor Model: Step III

- The split criterion is the “pricing errors” from a conditional factor model. It also follows the no-arbitrage condition for the asset pricing goal.

$$Ł = \sum_{t=1}^{T} \sum_{i=1}^{N_t} \left(r_{i,t} - \beta(z_{i,t-1}) f_t \right)^2,$$

- $\beta(z_{i,t-1}) = b_0 + b^\intercal z_{i,t-1}$ are conditional factor loadings.

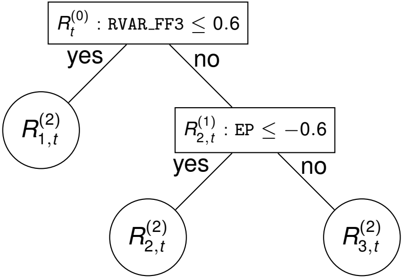
- The above yield the following regression:

$$r_{i,t} = b_0 f_t + b^\intercal Z_{i,t-1} f_t + \epsilon_{i,t}$$

- Quadratic loss for the entire cross section is the split criterion.

- Loop over all characteristics and breakpoints for the optimal model.

Panel Tree Factor Model: Step IV



- The second split gives us three leaf basis portfolios and a updated SDF:

$$f_t^{(2)} = \widehat{\Sigma}_2^{-1} \widehat{\mu}_2 R_t^{(2)} = w_{21} R_{1,t}^{(2)} + w_{22} R_{2,t}^{(2)} + w_{23} R_{3,t}^{(2)},$$

- For the second split, the algorithm searches over all leaf nodes, characteristics, and breakpoints.

- The split criterion is calculated based on the entire cross section, thus P-Tree and its SDF are global.

Boosted P-Trees for Multi-factor Models

- Generate multiple factors using a boosting design (sum of trees).
- The first factor $f_{1,t}$ is generated by the standard tree factor model on excess returns $\{r_{i,t}\}$. We save the $\hat{\beta}_1(z_{i,t-1}), \hat{f}_{1,t}$ from the previous tree.

$$r_{i,t} = \beta_1(z_{i,t-1})f_{1,t} + \epsilon_{i,t}$$

- To generate the second factor $f_{2,t}$, we train the tree factor model on $\{r_{i,t}\}$ **controlling the first factor and first beta.**

$$Ł = \sum_{t=1}^{T}\sum_{i=1}^{N_t}\left(r_{i,t} - \hat{\beta}_1(z_{i,t-1})\hat{f}_{1,t} - \beta_2(z_{i,t-1})f_{2,t}\right)^2$$

- Also allows for a benchmark adjusted model (market adjusted).

Boosted Tree: Market Adjusted Model

- Use the market factor as the first factor $f_{1,t}$
- Fit the stock returns with $f_{1,t}$ and find the beta on the first factor
- Fit the stock returns with $f_{1,t}$, $f_{2,t}$ with the beta on the first factor fixed
- Fit the stock returns with $f_{1,t}$, $f_{2,t}$, $f_{3,t}$ with the beta on the first and second factors fixed
- The process continues...

Duality between MVE and SDF

- Minimum variance of the SDF equals the maximal square Sharpe ratio of the MVE portfolio (Hansen and Jagannathan, 1991).

- P-Tree can incorporate wither asset pricing objective.

- Asset pricing criterion: SDF to explain the cross-sectional variation of stock returns.

$$\mathcal{L}_{\mathrm{A}} = \sum_{t=1}^{T} \sum_{i=1}^{N_t} \left(r_{i,t} - \beta_{i,t-1}^{\mathsf{T}} \mathbf{f}_t \right)^2,$$

- Investmestment-guidede criterion: maximize the Sharpe ratio of SDF.

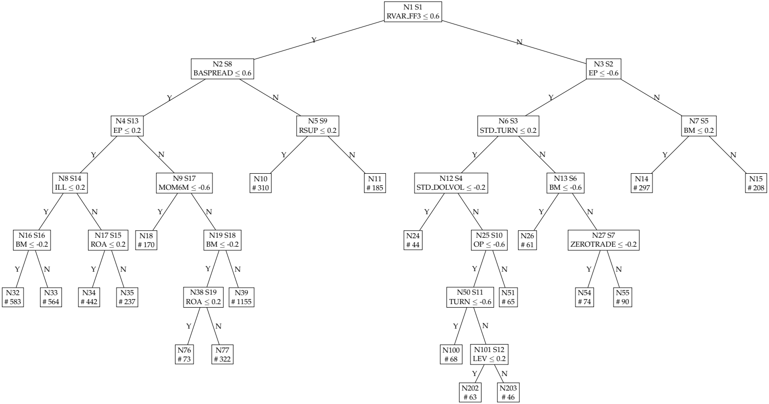
$$\mathcal{L}_{\mathrm{I}} = -\boldsymbol{\mu}_{\mathbf{F}}' \boldsymbol{\Sigma}_{\mathbf{F}}^{-1} \boldsymbol{\mu}_{\mathbf{F}},$$

Empirical Findings

U.S. Equities

- 1981-2020 monthly observation for US equities
- Returns and lag-one-month characteristics
- Standardize the characteristics in the cross-section into Uniform $[-1, 1]$
- 61 characteristics in 6 categories: momentum, value-versus-growth, investment, profitability, intangibles, and frictions
- Periods 1981-2000 and 2001-2020 as training and test samples.

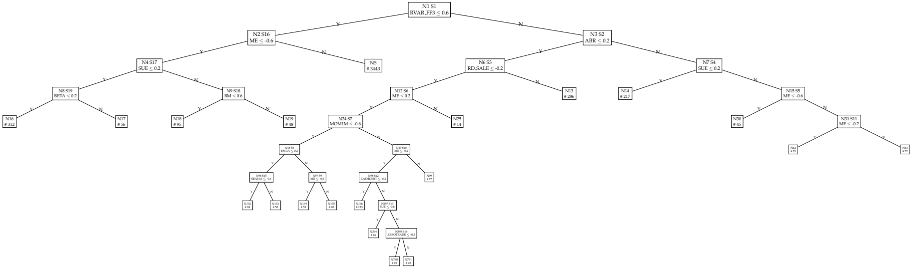
Asset Pricing Tree Structure



- rvar_ff3 (idiosyncratic volatility)
- ep (earnings-to-price)

Investment-Guided P-Tree Structure

Squared Sharpe Ratio as the Objective Function



- rvar_ff3 (idiosyncratic volatility)
- abr (abnormal return around ernings anouncement)
- rd_sales (R&D expense to sales)
- sue (standard unexpected earnings)

Cornell University

Variable Importance via Random P-Forest

- Study importance of variables using bagging (random forest) strategy.
- Fit a tree to bootstrapped return data (randomly draw 20 characteristics out of 61) repeat 1000 times independently.
- Any characteristic is considered about 330 times out of 1,000 subsamples for fitting the P-Forest.
- Two measurements of variable importance

$$\text{Selection Probability(K)} = \frac{\#(\text{Selected at first K splits})}{\#(\text{Randomly drawn})}$$

$$\text{Char. Importance} = E(\text{loss function}|\text{with char}_i) - E(\text{loss function}|\text{without char}_i)$$

Variable Importance: Top Splits

Random Forest

	1	2	3	4	5
Top1	RVAR_FF3 0.40	RVAR_CAPM 0.40	ME 0.39	SVAR 0.32	CFP 0.25
Top2	ME 0.45	RVAR_FF3 0.41	RVAR_CAPM 0.40	CFP 0.35	EP 0.33
Top3	ME 0.45	RVAR_FF3 0.41	RVAR_CAPM 0.41	CFP 0.37	EP 0.36

Measure of Asset Pricing Performance

- Pricing the individual stocks

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

where $\widehat{r}_{i,t} = \beta(z_{i,t-1}) f_t$

$$\text{Stock CS } R^2 = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}(r_{i,t} - \widehat{r}_{i,t})\right)^2}{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T} r_{i,t}\right)^2},$$

- Standard asset pricing test for portfolios (FF25, Ind49)

$$\text{Portfolio CS } R^2 = 1 - \frac{\sum_{i=1}^{N}\left(\bar{r}_i - \widehat{\bar{r}}_i\right)^2}{\sum_{i=1}^{N} \bar{r}_i^2},$$

Asset Pricing Performance

	Individual Stocks				Portfolios			
	In-Sample		Out-of-Sample		Entire Sample			
	Tot	CS	Tot	CS	FF25	Ind49	Leaf20	Leaf40
				Panel A: P-Tree				
PTree2	11.1	25.5	11.1	10.4	77.8	92.9	85.4	66.1
PTree5*	13.0	22.7	13.7	16.5	77.9	63.2	50.8	67.3
				Panel B: Other Benchmark Models				
CAPM	7.0	1.3	8.4	0.6	91.4	88.1	-219.1	-36.6
FF3	10.5	7.5	10.7	5.1	94.9	85.4	-204.7	-30.6
FF5	11.0	13.1	11.3	5.1	96.1	78.5	-72.7	22.7
Q5	10.9	18.1	11.5	6.4	96.1	88.7	32.5	62.6
RP-PCA5	12.1	18.3	13.6	15.0	69.7	48.6	-66.5	23.2
IPCA5	13.8	27.8	14.9	17.7	90.4	57.3	31.4	63.0

- P-Tree factors are strong at explaining stock returns.
- P-Tree gives 20 test portfolios; difficult to price by other models.
- Squared sharpe ratio to select no. of factors (Barillas and Shaken, 2017).

Investment Performance: Tradable, High Sharpe and Alpha

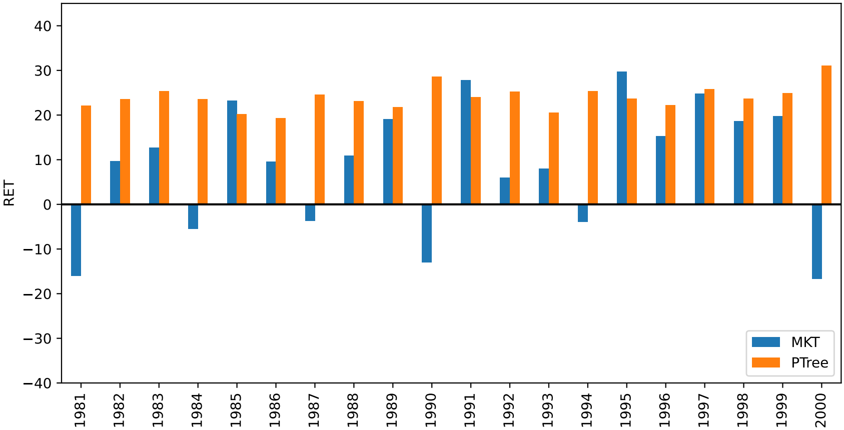
	In-Sample (1981-2000)						Out-of-Sample (2001-2020)					
	MVE			$1/N$			MVE			$1/N$		
	AVG	SR	α	AVG	SR	α	AVG	SR	α	AVG	SR	α
Panel A: Asset Pricing P-Tree												
PTree2	1.75	1.58	1.51***	1.31	1.34	0.86***	0.30	0.31	0.29	0.45	0.56	0.17
PTree5*	1.26	3.47	1.20***	1.06	1.69	0.72***	0.80	1.93	0.76***	0.70	1.14	0.42***
Panel B: Investment P-Tree												
PTree2	1.78	10.41	1.76***	1.27	1.94	0.90***	1.07	2.78	1.10***	0.86	1.36	0.56***
PTree5*	1.36	12.55	1.35***	0.78	1.93	0.56***	0.76	2.96	0.78***	0.48	1.34	0.32***
Panel D: Other Benchmark Models												
FF3	0.53	1.16	0.40***	0.38	0.85	0.20***	0.22	0.30	-0.06	0.28	0.40	0.01
FF5	0.45	1.48	0.38***	0.38	1.34	0.33***	0.27	0.64	0.13*	0.25	0.59	0.12
Q5	0.77	2.78	0.74***	0.63	2.10	0.53***	0.34	1.22	0.34***	0.31	1.10	0.25***
RP-PCA5	0.82	3.48	0.76***	1.07	1.77	0.75***	0.34	1.49	0.32***	0.50	1.00	0.27***
IPCA5	1.50	10.37	1.48***	0.90	3.15	0.80***	0.97	4.60	0.98***	0.73	2.14	0.61***

- P-Tree factors are tradable, with high Sharpe Ratio and Jensen's Alpha.

Factor Spanning Alpha Tests

	In-Sample			Out-of-Sample		
	FF5	Q5	IPCA5	FF5	Q5	IPCA5
Panel A: Market-Adjusted P-Tree factors						
RVAR_FF3-EP	130***	101***	107**	12	4	31
BM_IA-III	35**	33	-100***	107***	110***	34
MOM12M-STD_DOLVOL	82***	53***	-24	25	22	-95***
ME-RDM	52***	48***	109***	29***	27***	13
MVE (4 factors + mkt)	58***	45***	-21	36***	34***	-11
1/N (4 factors + mkt)	60***	47***	49*	35***	33***	1
Panel B: Market-Adjusted Investment P-Tree factors						
RVAR_FF3-ABR	354***	341***	227***	215***	201***	69***
BM_IA-LGR	46***	58***	96***	13	16	-20
STD_TURN-LEV	36***	32***	85***	-21**	-19**	-18
CFP-MOM12M	53***	49***	90***	45**	47**	5
MVE (4 factors + mkt)	248***	241***	175***	147***	139***	42**
1/N (4 factors + mkt)	98***	96***	131***	50***	49***	12
Panel C: Other Test Assets						
MVE-FF25	55***	42***	27*	19***	15**	10
MVE-IND49	13*	20	-14	10	8	28*
1/N-FF25	-8***	-8	57***	3	8**	5
1/N-IND49	63*	30	62	-2	10	4

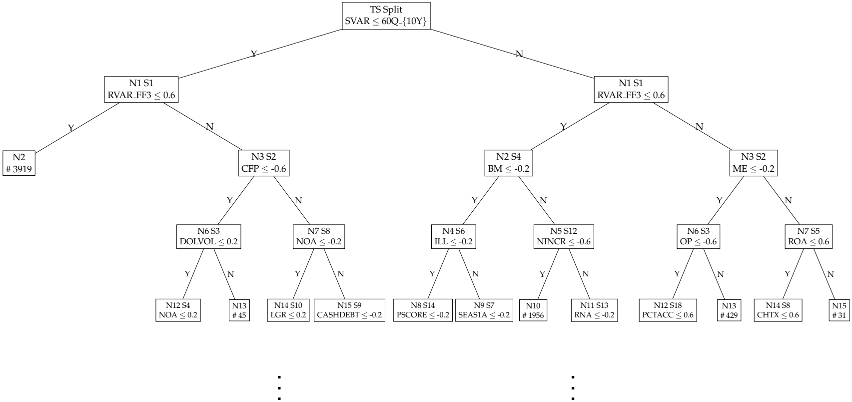
Investing in P-Tree Factors



Time-Series Split

- Asset returns are panel data with two dimensions.
- In addition to cross-section split, we can also include time-series split.
- The asset pricing tree model can be different under different macroeconomic conditions.
- When building the tree, we simply split the time-series before splitting the cross-section.

Asset Pricing Tree under High/Low Stock Variance



- Adapt to different macroeconomic conditions.
- Empirically, our model finds Stock Variance is the key indicator.
- We have all the empirical results for Time-Series P-Tree in the paper.

Cornell University

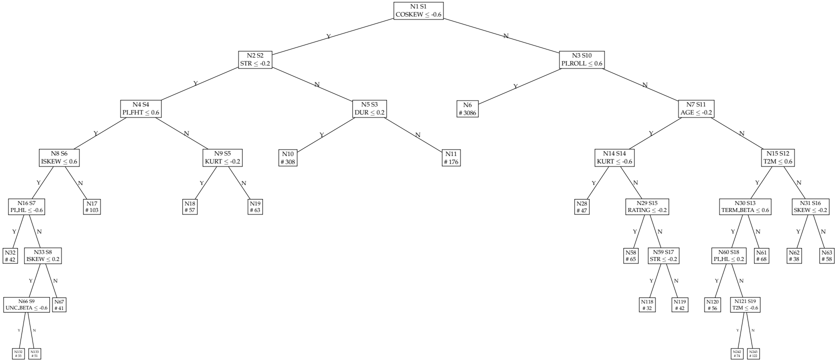
Extensions: Interaction to Strengthen or Resurrect Anomalies

- Fama-French type Factors - Long-short Portfolios sorted on one firm characteristic (or bivarate sorted with market equity).
- Characteristics or factor interaction is rarely explored.
- Possible to enhance factor risk premium by considering (asymmetric) interactions.
- Possible to resurrect insignificant factors by considering (asymmetric) interactions.
    - Maximum daily returns (Bali et al., 2021) has has a significant premium in the training sample but disappears in the test sample. Interacting with abnormal returns around earnings announcement (ABR) on the short portion and industry-adjusted size (ME IA) on the long portion earns 67 basis points for monthly average returns and 111 basis points for alpha.

Corporate Bonds Data

- 2002-2019 monthly observation for US corporate bonds
- Trade Reporting and Compliance Engine (TRACE)
- Transaction-level data
- Returns and lag-one-month characteristics
- Standardize the characteristics in the cross-section into Uniform $[-1, 1]$
- 40 characteristics in 4 categories: interest risk or maturity, beta (risk measures), liquidity, past return

Panel Tree for Corporate Bonds



- Corporate bond is an important and interesting market, with rich cross-sectional characteristics.
- P-Tree works well in corporate bond.

Takeaways

- P-Tree offers an alternative top-down solution to generalized sorting.
- Generated basis portfolios help construct factors for asset pricing, and serving as test assets.
- Using U.S. equity and corporate bond data, P-Tree models outperform standard factor models in pricing and return prediction.
- High-dimensionality, nonlinearity, interactions, low signal-to-noise, time heteroskedasticity, panel data $+$ Interpretable!
- A new class of models that provides a unified framework to
  - (i) analyze potentially non-i.i.d., unbalanced panel data, and
  - (ii) accommodate global split criteria (guided by economics).

  All while preserving trees' interpretability, computational feasibility, and suitability for financial big data.

Other Applications of the P-Tree Framework

"**Uncommon Factors and Bayesian Asset Clusters**"
(Cong, Feng, He, and Li, 2022).

- Do different assets follow different factor models – uncommon factors?
- How to separate assets for different models – observation clustering?
- How to choose factors for different clusters of assets – variable selection?

Motivation: Uncommon Factors

- Factor models —— explain the cross-sectional return dynamics
  - Well-known risk factors: Market, Beta, Size, Value, Momentum $\cdots$
- Long-standing topic to searching for the true or universal (factor) model that is not rejected by asset pricing tests.
  - For example, FF 5 factors explain 5×5 ME-B/M portfolios, but significant alpha for small-growth (Fama and French, 2015).

Motivation: Uncommon Factors

- Factor models —— explain the cross-sectional return dynamics
  - Well-known risk factors: Market, Beta, Size, Value, Momentum $\cdots$
- Long-standing topic to searching for the true or universal (factor) model that is not rejected by asset pricing tests.
  - For example, FF 5 factors explain 5×5 ME-B/M portfolios, but significant alpha for small-growth (Fama and French, 2015).
- There are a few directions of research
  - Missing factors? The literature keeps fishing more.
  - Factor zoo? Factor selection and model comparison.
  - Time variation? Unconditional v.s. Conditional model.
  - Choices of test assets? Unstable factor loadings, or weak factors?
  - Some assets may be just mispriced.
- Take a step back; maybe no one-size-fits-all empirically.

Motivating Uncommon Models and Observation Clustering

- Standard factor modeling for the holy grail of empirical asset pricing:

$$\begin{aligned} r_{1,t} &= \alpha_{1,t} + \beta_{1,1,t} f_{1,t} + \cdots + \beta_{1,k,t} f_{k,t} + \epsilon_{1,t} \\ &\vdots \\ r_{n,t} &= \alpha_{n,t} + \beta_{n,1,t} f_{1,t} + \cdots + \beta_{n,k,t} f_{k,t} + \epsilon_{n,t} \end{aligned}$$

  - LHS observations/assets are heterogeneous; grouped heterogeneity.
  - Burden all on RHS model estimation and selection.
  - Sorting/test asset construction for common models & cross-cluster spread.

- A novel approach for *jointly* considering observation clustering and heterogeneous model selection:
  - Model selection on RHS: homogeneous observations following one common factor model.
  - Observation clustering on LHS: split the cross-section such that each cluster has a model with potentially uncommon factors.
  - Data-driven yet incorporating economic principles/finance theory and preserving interpretability.

Clustering in Finance

- Pre-specified clustering in asset pricing
  - Industry classification (Fama and French, 1997).
  - International finance: sorted portfolios (Karolyi and Stulz, 2003; Hou et al, 2011) and individual assets (Chaieb et al, 2021).
- Characteristics-based Clustering
  - Security sorting on characteristics clusters individual stocks (to form sorted portfolios) for similar risk exposures (Berk 2000).
  - Panel tree for splitting the cross section (Cong et al., 2022)
- The correct cluster is unknown (no observed labels).
  - Supervised clustering based on factor model fitness (Patton and Weller, 2019; Cong et al., 2022).
  - Unsupervised clustering using return correlation (Ahn et al., 2009).

Risk Factor Selection

- Current factor/characteristic selection studies focus on aggregate signals
  - Factor Selection in Time-Series Regression (betas) (Hwang and Rubesam, 2020; Avramov et al., 2022).
  - Factor Selection in Cross-Sectional Regression or SDF model (risk price) (Kozak et al., 2020; Feng et al., 2020; Bryzgalova et al., 2022).
  - Characteristics selection for Future Return Predictability (Freyberger et al., 2020; Gu et al., 2020).
- Weak factors (Kkan and Zhang, 1999; Giglio, Xiu, and Zhang, 2022).
  - factors to which the test assets have little or no exposure
  - standard estimation and inference incorrect
- Uncommon factors —— an alternative to overcome empirical challenges.

Bayesian Methods in Finance

- Why use Bayesian methods?
  - Parameter uncertainty (Kandel and Stambaugh, 1996; Barberis, 2000).
  - Model uncertainty
    - Model Averaging (Avramov, 2002; Avramov et al., 2022).
    - Shrinkage Prior (Hwang and Rubesam, 2020; Bryzgalova et al., 2022).
  - Economic Prior Beliefs (Pastor, 2000; Paster and Stambaugh, 2000; Avramov and Chordia, 2006; Avramov and Wermers, 2006).
  - Posterior probabilities for factor usefulness, credible interval for model parameters, and predictive distribution for risk assessment.
- How to use Bayesian methods to compare factor models?
  - Bayesian marginal likelihood (Barillas and SHanken, 2018; Chib et al, 2020) considers and integrates parameter uncertainty or/and model uncertainty.
- Therefore, marginal likelihood is a natural and interpretable global split and stopping criterion for clustering ——- splitting the cross section.

Single Leaf Model: A Bayesian Factor Model

For all assets in the same j-th leaf,

- $r_{i,t}$: a panel of individual stock returns
- f_t: traded risk factors (MktRF, SMB, HML, RMW, CMA, MOM, etc.)
- $z_{i,t-1}$: prespecified firm characteristics

$$
\begin{aligned}
r_{i,t} &= A(_{i,t-1}) + B(_{i,t-1})\mathbf{f}_t + \epsilon_{i,t} \\
A(_{i,t-1}) &= \alpha_j \\
B(_{i,t-1}) &= \boldsymbol{\beta}_j(_{i,t-1}) \\
\boldsymbol{\beta}_j(\mathbf{z}_{i,t-1}) &= \mathbf{b}_{j,0} + \mathbf{b}_{j,1}\left(I_K \otimes \mathbf{z}_{i,t-1}\right) \\
\epsilon_{i,t} N(0, \sigma^2_{i,t}), &\quad \sigma^2_{i,t} = \sigma^2_j,
\end{aligned}
$$

Estimate a pooled model for all assets with idiosyncratic betas and alphas driven by $z_{i,t-1}$. Plug dynamic $\alpha_j(\cdot)$ and $\boldsymbol{\beta}_j(\cdot)$:

$$
r_{i,t} = \alpha_j + \mathbf{b}_{j,0}\mathbf{f}_t + \mathbf{b}_{j,1}\left(\mathbf{f}_t \otimes \mathbf{z}_{i,t-1}\right) + \epsilon_{i,t},
$$

Model Estimation and Factor Selection using Spike-and-Slab

- SS as Bayesian variable selection prior for selecting $\mathbf{f}_t$.
- Skeptical investor ($w_i = 0.1$) versus Agnostic investor ($w_i = 0.5$).
- Bayesian variable/factor selection assuming independent SS priors on $\mathbf{b}_{j,0}$:

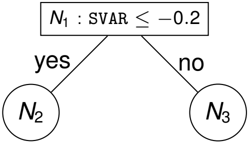
$$
\begin{aligned}
\pi(b_{j,0,k} \mid \sigma_j^2, \boldsymbol{\gamma}_j) &= (1-\gamma_{j,\mathbf{f},k})N(0,\xi_0^2\sigma_j^2) + \gamma_{j,\mathbf{f},k}N(0,\xi_1^2\sigma_j^2);\, k=1,\cdots,K,\\
\pi(b_{j,1,k,i} \mid \boldsymbol{\gamma}_j) &= (1-\gamma_{j,\mathbf{f},k})N(0,\xi_0^2\sigma_j^2) + \gamma_{j,\mathbf{f},k}N(0,\xi_1^2\sigma_j^2);\, k=1,\cdots,K;\, i=1,\cdots,M,\\
\pi(a_{j,0} \mid \sigma_j^2) &= N(0,\xi^2\sigma_j^2),\\
\pi(\sigma_j^2) &= \text{inverse-Gamma}(S_0, v_0),\\
\pi(\boldsymbol{\gamma}_j) = \pi(\boldsymbol{\gamma}_{j,\mathbf{f}}) &= \prod_{k=1}^{K} w_k^{\gamma_{j,\mathbf{f},k}}(1-w_k)^{(1-\gamma_{j,\mathbf{f},k})}.
\end{aligned}
$$

Latent γ denotes the prior on coefficient bing “spike” or “slab.”.

$$
\boldsymbol{\gamma}_j = (\gamma_{j,1}, \gamma_{j,2}, \cdots, \gamma_{j,K+KM}) = (\underbrace{\boldsymbol{\gamma}_{j,\mathbf{f}}}_{K\times 1}, \underbrace{\boldsymbol{\gamma}_{j,\mathbf{f}\bullet\mathbf{z}}}_{KM\times 1}),
$$

From Single Leaf to a Tree: Marginal Likelihood as Global Split Criterion

- Split the cross section according to asset characteristics



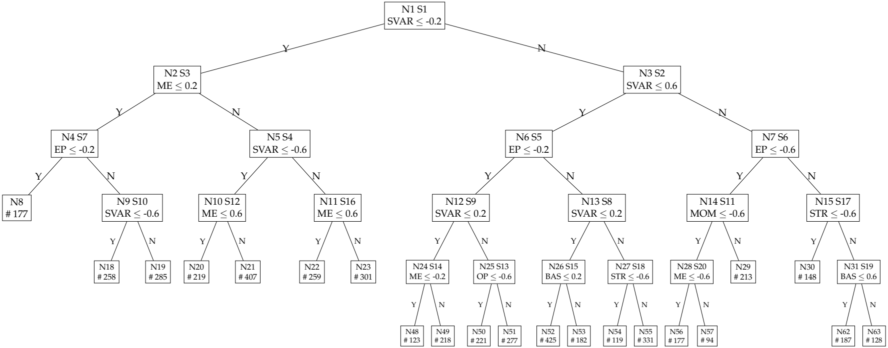
- "Goodness" of a candidate split: joint marginal likelihood of the models on two child nodes.
- Model parameters can be integrated out a priori:

$$p(\mathcal{A}_0) := p(\mathbf{R} \mid \mathbf{Z}, \mathbf{F}) = \int p(\mathbf{R} \mid \mathbf{Z}, \mathbf{F}, \gamma_j, \alpha_j, \mathbf{b}_{j,0}, \mathbf{b}_{j,1}, \sigma_j^2)$$
$$\times\, \pi(\alpha_j \mid \sigma_j^2)\pi(\mathbf{b}_{j,0}, \mathbf{b}_{j,1} \mid \sigma_j^2, \gamma_j)\pi(\sigma_j^2 \mid \gamma_j)\pi(\gamma_j) d\alpha_j d\mathbf{b}_{j,0} d\mathbf{b}_{j,1} d\sigma_j^2 d\gamma_j.$$

- Separation of tree growth and mis-specification/estimation.
- Parameter and model uncertainties captured in closed form.

Splitting the Cross Section into Asset Clusters

- Four major cluster groups driven by SVAR (-0.2), ME (-0.2), SVAR (-0.6).
- Low-vol and size-related anomalies as grouped heterogeneity: low SVAR loads not on IVOL, high SVAR loads not on BAB.
- Robustness in Size-adjusted trees.



Key Findings

- Asset returns exhibit grouped heterogeneity.
- BCM applied to U.S. individual stock returns identifies market, size, and short-term reversal as common factors, and several uncommon factors that lose exposure to some clusters during tree growth.
- Differential factor exposure and potential segmentation manifest primarily through differential stock variance, followed by market equity and earnings-to-price ratio.
- Built on leaf clusters, a tangency portfolio on cluster-selected factor models delivers exceptional in-sample and out-of-sample performance.
- Cluster alphas indicate arbitrage opportunities and can generate an out-of-sample monthly average return of 2.22% using LS hedged alpha portfolios.
- More skeptical prior beliefs lead to less prediction risk and better coverage.