Multifactors Risk Research of China Stock Listed Banks

Chuxuan Ma

Lehigh University P.C. ROSSIN COLLEGE

April 2025

Outline

Problem Description

Weighted Average Cost of Capital (WACC)

Critical financial indicator for assessing a company's capital cost:

$$\mathbf{r}_{wacc} = \frac{E}{E+D} \times \mathbf{r}_e + \frac{D}{E+D} \times \mathbf{r}_d \times (1-t)$$

Where:

- E: Market value of equity
- D: Market value of debt
- $\mathbf{r}_e$: Cost rate of equity
- $\mathbf{r}_d$: Cost rate of debt
- t: Corporate tax rate

Risk Premium Component

$$\mathbf{r}^*_{wacc} = \mathbf{r}_{wacc} - \mathbf{r}_f$$

Where $\mathbf{r}_f$ is the risk-free interest rate (SHIBOR)

Research Goal

Primary Objective

Explore causal relationships between financial factors and risk premiums in China's banking sector

Risk Premium Decomposition

- Common (market-driven)
- Idiosyncratic (bank-specific)

Risk Types Analysis

- Default Risk
- Liquidity Risk
- Market Risk

Expected Contribution

Provide insights into fundamental risk structures of China's banking industry

China's Banking Context

Unique Characteristics

- Mixed ownership structures
- Ongoing financial reforms
- Varying government influence
- Evolving regulations

Research Significance

- Different from Western banks
- Requires specialized analysis
- Critical for financial stability
- Informs regulatory policy

Dataset Overview

Data Scope

- 42 listed Chinese banks
- 978 trading days (2019-2022)
- JoinQuant API data source
- Captures pre/post-pandemic conditions

Data Components

Financial Indicators

421 factors including:

- Technical factors
- Fundamentals
- Macroeconomics
- Money flows
- Securities margins
- Industry metrics
- Index data

Bank Information

- Listing details
- Enterprise values
- Market capitalization
- Return rates

Model Parameters

- WACC components
- Risk-free rates
- Market returns
- FCF model inputs

Data Characteristics

Data Preparation

- **Standardized**: All factors normalized
- **Tail-shrinked**: Outliers processed using [median -5 IQR, median $+5$ IQR]
- **Industry-neutralized**: Residualized from industry effects
- **Market cap-neutralized**: Adjusted for size effects

Missing Values Processing

- Distinguished between pre-listing/post-delisting missing values and disclosure-related missing values
- Used IterativeImputer with XGBoost for sophisticated imputation

Enterprise Value Weighting

Sectoral Factors Methodology

- Created weighted sectoral factors by enterprise value
- Formula:

$$\text{BankIndustryIndicator}_{j,t} := \frac{\mathbf{v}_{j,t}^{T} \cdot \textbf{Indicator}_{j,t}}{\sum_{i=1}^{42} v_{i,j,t}}$$

- Where:
  - i: Bank index ($i \in \{1, 2, ..., 42\}$)
  - j: Indicator index ($j \in \{1, 2, ..., 421\}$)
  - t: Trading day
  - v: Enterprise value

Methodological Framework

Key Components

- Double Machine Learning framework for causal inference
- Controls for confounding effects across 421 financial indicators
- Combines multiple advanced techniques:
  - K-Shape clustering for risk classification
  - AdaBoost regression with recursive feature elimination
  - Causal Forest models for treatment effect heterogeneity

Risk Classification System

Domain Knowledge Classification

Initial categorization based on banking finance theory:

- Default Risk
- Liquidity Risk
- Market Risk

K-Shape Clustering

- Time-series clustering based on shape similarity
- Updates knowledge-based classifications
- Adapts to observed patterns in factors
- Reveals hidden relationships

K-Shape Risk Reclassification

K-Shape Algorithm Mathematical Formulation

For time series factors **x** and **y** of length m:

- Cross-correlation at shift q:

$$CC_q(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{m-q} x_i \cdot y_{i+q}$$

- Shape-based distance (SBD):

$$SBD(\mathbf{x}, \mathbf{y}) = 1 - \max_q \frac{CC_q(\mathbf{x}, \mathbf{y})}{\sqrt{CC_0(\mathbf{x}, \mathbf{x}) \cdot CC_0(\mathbf{y}, \mathbf{y})}}$$

- Clustering $\mathcal{C}^*$ objective: Minimize within-cluster SBD sum

$$\min_{\mathcal{C}} \sum_{k=1}^{K} \sum_{\mathbf{x} \in \mathcal{C}_k} SBD(\mathbf{x}, \mu_k)$$

where μ_k is the shape-based centroid of cluster $\mathcal{C}_k$.

Machine Learning Architecture

CustomRegressor $\mathtt{rc}(\cdot)$: Multi-stage Ensemble

1. **Base Learner 1**: Decision Tree with Bayesian hyperparameter optimization and all features
2. **Strong Learner 1**: AdaBoost with optimized Base Learner, all features and Bayesian hyperparameter optimization
3. **Feature Selection**: Recursive Feature Elimination with Cross-Validation and Strong Learner
4. **Base Learner 2**: Decision Tree with Bayesian hyperparameter optimization and selected features
5. **Strong Learner 2**: AdaBoost with optimized Base Learner and Bayesian hyperparameter optimization and selected features

Risk Premium Decomposition

Mathematical Decomposition Framework

Across bank sector at all trading days (2019-2022), we decompose the WACC risk premium:

$$\mathbf{r}^*_{wacc} = \mathbf{r}_{wacc} - \mathbf{r}_f \quad \text{(Total risk premium)}$$

Prediction-based decomposition:

$$\hat{r}^*_{common} = \mathrm{rc}(\mathbf{r}_{wacc}) \quad \text{(Prediction from WACC)}$$
$$\mathbf{r}^*_{idio} = \mathbf{r}^*_{wacc} - \hat{r}^*_{common} \quad \text{(Residual component)}$$

Treatment risk decomposition using K-Shape clusters:

$$\hat{r}^*_{\mathrm{XR}} = \mathrm{rc}(\mathbf{r}_{\mathrm{XR}}) \quad \text{(Prediction from WACC)}$$

where $X_{\mathrm{XR}} = \{x_i \mid i \in \mathcal{C}_{\mathrm{XR}}\}$ (Liquidity risk factors) and $\mathrm{XR} \in \{\mathrm{DR}, \mathrm{LR}, \mathrm{MR}\}$

- Common risk represents the predictable component based on WACC
- Idiosyncratic risk represents unexplained variations
- Treatment risks are constructed from factors classified by K-Shape clustering

Causal Inference Framework

Double Machine Learning Mathematical Framework

$$Y = \theta T + g(X) + \varepsilon_Y, \varepsilon_Y \sim \mathcal{N}(0, \sigma^2)$$
$$T = m(X) + \varepsilon_T, \varepsilon_T \sim \mathcal{N}'(0, \sigma'^2)$$

Orthogonalization process:

$$\tilde{Y} = Y - \overbrace{\mathbb{E}[Y|X]}^{\theta m(X)+g(X)} = \theta\epsilon_T + \epsilon_Y \quad \text{(Residualized outcome)}$$
$$\tilde{T} = T - \overbrace{\mathbb{E}[T|X]}^{m(X)} = \epsilon_T \quad \text{(Residualized treatment)}$$
$$\text{ATE} := \theta = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[\tilde{Y}|\tilde{T}] = \mathbb{E}[\text{CATE}(X)] \quad \text{(Average Treatment Effect)}$$

For heterogeneous effects, we estimate:

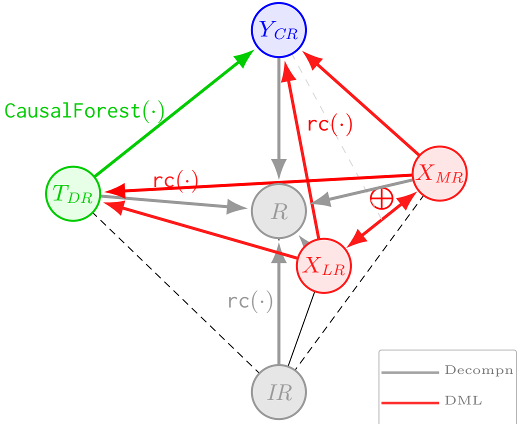
$$\text{CATE}(x) = \tau(x, 1, 0) = \mathbb{E}[Y(1) - Y(0)|X = x] \quad \text{(Conditional Treatment Effect)}$$

where:

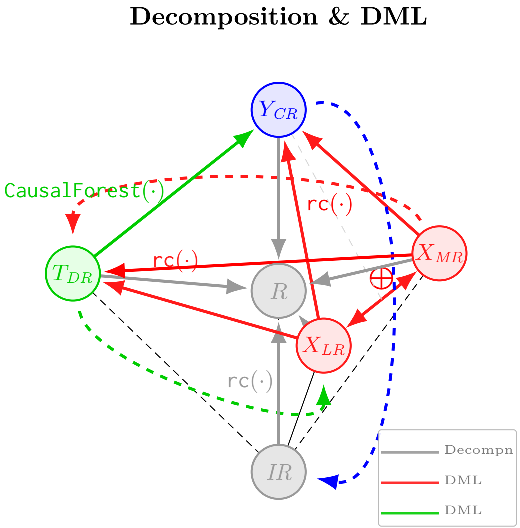
- $\tau(x, t, t') := \mathbb{E}[Y(t) - Y(t')|X = x]$.
- We apply $\hat{\mathbb{E}}[Y|X] = \text{rc}$, $\hat{\mathbb{E}}[T|X] = \text{rc}$ and $\tau = \texttt{CausalForest}$ here.
- Outcome $Y \in \{Y_{\text{CR}}, Y_{\text{IR}}\}$: Common Risk or Idiosyncrasy Risk.
- Treatment $T \in \{T_{\text{DR}}, T_{\text{LR}}, T_{\text{MR}}\}$: Default Risk or Liquidity Risk or Market Risk.
- Confounders $X \in \{X_{\text{LR}} \oplus X_{\text{MR}}, X_{\text{DR}} \oplus X_{\text{MR}}, X_{\text{DR}} \oplus X_{\text{LR}}\}$: Liquidity Risk & Market Risk or Default Risk & Market Risk or Default Risk & Liquidity Risk.

Visual Pipline of Decomp & DML

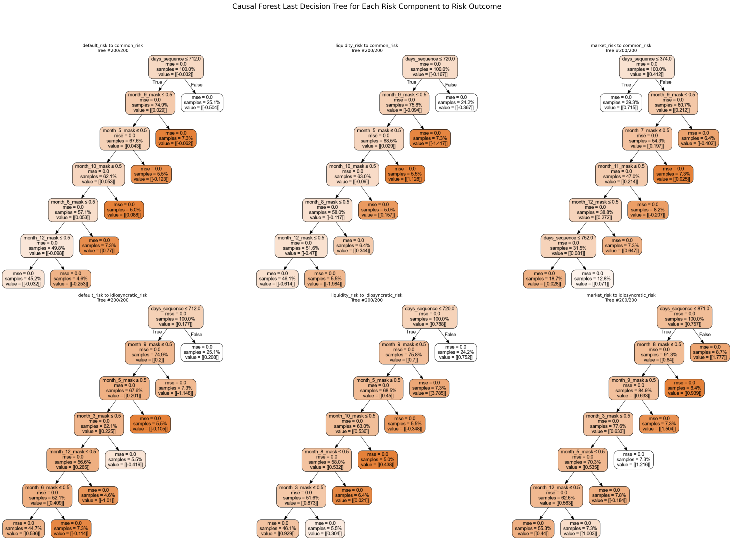
Decomposition & DML



Visual Pipline of Decomp & DML Iteration



CATE Trees in Causal Forest



CATE Patterns Overview

Directional Effects

- Most relationships: positive CATEs
- Market risk → Common risk: strongest positive effects
- Default risk: moderate positive
- Liquidity risk: variable effects

Magnitude & Stability

- Effect range: -0.1 to +0.3
- Mostly clustered: 0.05-0.15
- Market risk: high stability
- Liquidity risk: high variability

Temporal Heterogeneity

Effect Evolution Over Time

- Stronger effects in earlier periods
- Gradual decay over sample timeframe
- Clear structural breakpoints
- Evolving risk transmission mechanisms

Key Temporal Anomalies

- Market risk: maintains consistent effect strength
- Liquidity risk $\rightarrow$ idiosyncratic risk: sign inversions
- Default risk $\rightarrow$ idiosyncratic risk: complex dependencies

Seasonal Heterogeneity

Reporting Cycle Effects

- Quarter-end months show significant effects
- Financial reporting influences risk relationships
- Regulatory disclosure timing matters

Annual Pattern Breakpoints

- January-February: year-start effects
- July-August: mid-year transitions
- Season-specific risk behavior

Risk-Specific Seasonal Patterns

- Market risk: strong summer effects
- Default risk: amplified year-end effects
- Liquidity risk: negative mid-year, positive year-end

Integrated Temporal-Seasonal Effects

- Risk relationships show increasing seasonal dependency over time
- Banking sector risk transmission increasingly synchronized with:
  - Regulatory cycles
  - Reporting periods
  - Fiscal quarters
- Pattern suggests evolving market maturity and institutional adaptation

Key Findings

Risk Relationship Insights

- Complex causal networks between financial factors and risk premiums
- Different risk types have varying impacts on common vs. idiosyncratic components
- Market risk shows strongest and most stable effects across conditions
- Significant temporal and seasonal patterns in risk transmission

Implications

- **For Regulators**: Consider timing of policy implementation
- **For Investors**: Account for seasonal risk patterns in portfolios
- **For Bank Management**: Adjust risk frameworks for heterogeneous effects

Limitations & Future Research

Limitations

- Period-specific findings (2019-2022)
- Sample limited to listed banks
- Potential unobserved confounders
- Modeling assumptions

Future Research

- Extend to non-listed banks
- Compare with international markets
- Incorporate policy event studies
- Develop predictive risk models

Acknowledgements

- JoinQuant platform for comprehensive financial data
- Open-source Python libraries:
  - pandas, numpy (data processing)
  - scikit-learn (machine learning)
  - xgboost (gradient boosting)
  - econml (causal inference)
  - matplotlib, seaborn (visualization)
  - graphviz (decision tree visualization)
  - scikit-optimize (Bayesian optimization)
  - tslearn (time series clustering)
  - tqdm (progress tracking)

Thank You!

Questions?

chm524@lehigh.edu