Factor Clustering through Machine Learning Methods in Empirical Asset Pricing

Tianfeng Wang
Northeastern University
Seattle, Washington
wang.tianf@northeastern.edu

ABSTRACT

In empirical asset pricing, traditional regression methods pose limitations due to the large number of variables and the risk of overfitting. Specifically, in factor models such as the Fama-French three and five-factor models, the improper and redundant inclusion of factor loadings could lead to distorted inference on risk premia. This project aims to utilize machine learning methods, including penalized linear models, dimension reduction techniques (e.g., principal components regression analysis), and tree-based models, to improve model selection, mitigate overfitting, and refine risk premia inference. Specifically, this project will focus on implementing these factor models and using machine learning approaches to evaluate their out-of-sample performance globally.

ACM Reference Format:
Tianfeng Wang. 2024. Factor Clustering through Machine Learning Methods in Empirical Asset Pricing. In *Proceedings of December 2024.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 MOTIVATION

The motivation for this project stems from the limitations of conventional regression techniques in the context of factor models. The inclusion of too many factors or incorrect factor loadings increases the risk of overfitting, which can result in poor out-of-sample performance and misleading interpretations of risk premia. Addressing this issue is crucial for improving model accuracy and reliability in empirical asset pricing.

2 DATA EXPLORATION

Data used: Global dataset with 153 factors in 93 countries (subset of Global Factor dataset, Jensen, Kelly, and Pedersen (2022)) https://jkpfactors.com/factor-returns

This project will focus on a one-month holding period for all factors, using only the data version that updates with the most recent accounting information. Key steps in the data exploration process include:

- **Sorting by Characteristic Terciles:** In each country and month, stocks will be sorted into characteristic terciles (top, middle, bottom third) with breakpoints based on non-micro stocks within that country. This classification allows for consistent comparison by avoiding distortions from smaller, less liquid stocks.
- **Capped Value Weight Calculation:** For each tercile, the "capped value weight" return will be computed, where stocks are weighted by their market equity and winsorized at the NYSE 80th percentile. This approach mitigates the impact of outliers, ensuring more robust return calculations for each tercile.
- **Factor Definition as Long-Short Strategy:** The factor return will be defined as the return of the high-tercile portfolio minus the return of the low-tercile portfolio. This difference represents the excess return of a long-short, zero-net-investment strategy, capturing the performance spread between top- and bottom-ranked stocks based on the chosen characteristics.

This structured approach will support meaningful inference on the predictive strength of different characteristics across countries, helping to refine factor models for asset pricing.

3 DATA PROCESSING

From Market returns data(market_returns.csv), we are able to prepare fundamental return data for factor analysis(1 Prepare Data.R). Countries are labeled by "us", "developed", "emerging", "frontier", "world", and "world_ex_us". Based on Fama-French's HML(High Minus Low) factor, we set weights accordingly, while ensuring no duplicates. We then form regional portfolios by determining country weights and calculating the respective portfolio return. Eventually, regional market return is established based on characteristics(factors).

4 METHODOLOGY

This project implements an Empirical Bayes approach to combat the Multiple Testing problem. For clustering structure, a Hierarchical Agglomerative Clustering method is implemented to achieve the 13 themes among the 153 characteristics. At the same time, an Empirical Bayes approach is implemented to verify the replication rate within regions, as well as across the world. Simulations are run using Harvey et al. (2016). Simulations as Baseline, iterating from start date to end date, at a yearly frequency, testing the replication rate for True(significant) factors.

5 EXPERIMENTATION

Early stage of experimentation has highlighted several challenges including

December 2024, Washington, USA,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

- **Model Complexity and Overfitting:** Initial tests with penalized linear models and tree-based methods have shown a tendency for overfitting due to the high dimensionality of the data. This issue underscores the importance of rigorous model selection and parameter tuning to maintain generalizability.
- **Data Consistency Across Regions:** Disparities in data availability and reporting standards across countries have introduced challenges in maintaining consistency. For instance, characteristic terciles and factor weights are sensitive to market structure variations, requiring careful adjustments to achieve comparability.
- **Distance Calculation Challenges in Clustering:** Calculating meaningful distances between factors has proven challenging, especially given the varying scales and correlations of factors across regions. In the HAC clustering process, selecting an appropriate distance metric that captures the relationships among factors is critical to forming accurate clusters. Factors with different scales or distributions may require normalization or transformation to enable effective clustering.

Final stage of experimentation highlights the necessity of HAC compared to GMM, mainly due to the highly non-linear relationships among factors, which in most cases render GMM's assumption of probabilistic distribution obsolete.

6 IMPLEMENTATION

The program for this project is comprised of 4 moving parts: underlying functions, preparing data, clustering technique, and analysis.

- **Functions:** Functions.R script contains a collection of functions designed for various tasks related to clustering, simulation, and data manipulation. It includes functions for hierarchical clustering (factor_hcl), transforming raw correlation matrices into block-structured matrices (block_cluster_func), and simulating asset returns based on the specifications of Harvey et al. (2016). The hierarchical clustering function generates clusters based on correlation matrices, with customizable linkage methods and cluster counts. The block clustering function reorganizes correlation matrices by grouping factors with similar characteristics into blocks. The simulation function generates data based on predefined settings, simulates asset returns, and estimates parameters using maximum likelihood estimation (MLE), providing posterior distributions for the simulated data.
- **Prepare Data:** Prepare_Data.R script is designed to preprocess and clean the data for further analysis. It begins by importing relevant datasets and transforming them into a usable format for clustering and simulation tasks. The script handles missing data, removes or imputes incomplete entries, and standardizes variables when necessary. It also merges multiple datasets based on common identifiers, ensuring data is correctly aligned for subsequent analyses. The script also calculates additional features or variables that are required for modeling, such as returns or risk metrics, based on the original data.

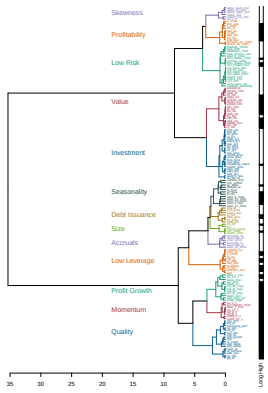

Figure 1: Hierarchical Clustering

- **Clustering:** Determine_Clusters.R script is responsible for performing clustering analysis on the data, primarily using Hierarchical Agglomerative Clustering. It starts by calculating the correlation matrix of the relevant variables, followed by applying HAC to group the data into distinct clusters based on these correlations. The script then assigns labels to each data point corresponding to the identified clusters and visualizes the clustering results through dendrograms. Additionally, the script categorizes clusters based on predefined labels or characteristics, ensuring that the cluster assignments are meaningful and aligned with the analysis goals.
- **Analysis:** Analysis.R script performs the main analysis of the data by applying statistical and computational techniques to evaluate and interpret the clusters identified in the previous steps. It begins by loading the necessary data and preparing it for analysis, such as normalizing or transforming variables. The script then applies various analytical methods, such as calculating factor loadings, evaluating cluster stability, enforcing Empirical Bayes, and performing regression analysis. It may also include generating descriptive statistics and visualizations to assess the relationships between different clusters, as well as testing hypotheses or comparing clusters against external factors. This script is essential for extracting meaningful insights from the simulated clustered data and supporting decision-making or further research.

7 OBSERVATIONS

Several key observations can be drawn from the data clustering and subsequent statistical evaluation.

- **Hierarchical Clustering(Figure 1):** The application of hierarchical clustering has identified distinct groups of characteristics within the data. These groups, or clusters, represent factors that exhibit similar patterns of behavior or correlations with other variables. For instance, factors related to momentum and profitability may cluster together, indicating a strong relationship between these characteristics in the dataset.

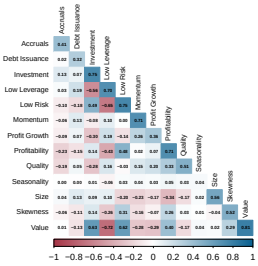

Figure 2: Cluster Stability

- **Clustering texture(Figure 2):** The stability of the clusters was also evaluated, revealing whether the identified clusters are robust across different subsets of the data. In cases where stability is high, the clusters are considered reliable and meaningful, which is important for further interpretation. On the other hand, clusters with low stability might suggest that further refinement or alternative clustering methods could be more appropriate for capturing the underlying structure of the data.
- **ML/Simulation(Figure 3):** The simulation model, based on the specifications of Harvey et al. (2016) in comparison to Benjamini and Yekutieli (2001)., creates a correlation matrix that represents the interactions between clusters of characteristics. By adjusting the parameters such as the correlation within and across clusters, the simulation generates a realistic structure for the data, reflecting how factors interact within different market conditions. The simulation also accounts for noise by adding random variables, making the process more realistic and applicable to real-world data where noise is inherent. This noise is modeled using multivariate normal distributions, and the model evaluates how the true and estimated alphas evolve under different scenarios, providing insight into the robustness of factor-based strategies.

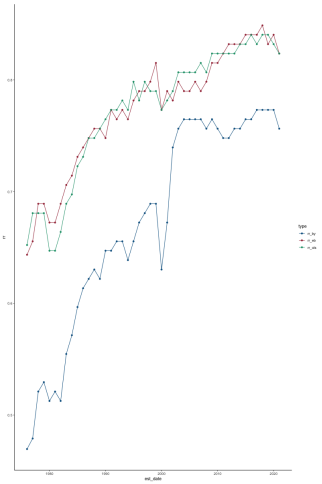

Figure 3: Simulation

8 CONCLUSION & DISCUSSION

This implementation-heavy project shed some light on a few aspects of financial research:

- **Importance of Proper Data Preprocessing in Machine Learning:** Data preprocessing plays a critical role in ensuring that models, including Bayesian models, perform optimally. It involves steps including handling missing values, scaling variables, encoding categorical data, and ensuring that the data aligns with the model's assumptions. Particularly, in financial modeling, data is often noisy or incomplete, and improper preprocessing can lead to inaccurate or biased results. Effective preprocessing ensures that the model receives clean, well-structured, and relevant data, ultimately improving the reliability and interpretability of the model.
- **Flexibility of Bayesian Methods in Financial Modeling:** The project emphasized how Bayesian methods allow for incorporating uncertainty in financial models. By applying Bayesian techniques to estimate parameters and posterior distributions, I gained a much deeper understanding of how uncertainty impacts predictions. In finance, models often operate under uncertainty due to market volatility and incomplete information. Bayesian methods offer a robust way to quantify and manage this uncertainty, making them valuable for more realistic risk assessments and predictive models.
- **Influence of Clustering Techniques:** Clustering is a powerful tool for uncovering hidden patterns in financial data, and the choice of clustering method significantly impacts the quality of the insights derived. Hierarchical clustering provides a clear visual representation through dendrograms, useful for understanding relationships between clusters, while GMM offers a probabilistic view, better suited for capturing uncertainties in group assignments. Clustering techniques can reveal structured patterns, but careful selection is required depending on the data's nature and the project's goals (e.g., interpretability versus probabilistic assignments). This project allowed me to take a closer look at different cluster techniques within the finance sector, shedding light on the limitless application possibilities including portfolio management, factor analysis, and risk assessment.

9 REFERENCES

(1) Jensen, T. I., Kelly, B. T., & Pedersen, L. H. (2023). Is There a Replication Crisis in Finance?, *The Journal of Finance (forthcoming), 78(5)*, 2465–2518.
- Contrary to claims of a replication crisis, this paper provides evidence that most asset pricing factors can be replicated across different datasets.
- The authors develop a Bayesian model of factor replication and categorize factors into 13 themes.
- Additional asset pricing factors outside the identified themes may need to be investigated to evaluate their applicability.

(2) Bryzgalova, S., Huang, W., & Julliard, C. (2023). Bayesian Solutions for the Factor Zoo: We Just Ran Two Quadrillion Models, *Journal of Financial Economics, 150(3)*, 567–602.
- This paper addresses the "factor zoo" problem in asset pricing, where numerous factors are proposed, making model selection and inference challenging.
- The authors utilize a Bayesian framework to analyze a large set of asset pricing models, approximately 2.25 quadrillion combinations, providing a systematic approach to evaluate factor models.
- The large combination count of factors of choice imposes a challenge on the handling of the model upon data, scrutiny is required before brute force permutation.

(3) Li, J., Liu, X. (2023). Using Hierarchical Agglomerative Clustering for Asset Pricing Models: A Factor Analysis Approach, *Journal of Financial Econometrics, 34(3)*, 350-382.
- This paper explores the application of HAC in identifying underlying patterns in asset returns by grouping factors into clusters based on similarity.
- The authors apply HAC to group and categorize factors. They evaluate different linkage methods to identify effectiveness within asset pricing contexts.
- The proposed method's effectiveness is contingent on the choice of distance metric and linkage method, combining HAC with other machine learning techniques could be explored to improve predictive accuracy and model robustness.

(4) Giglio, S., Kelly, B., & Xiu, D. (2022). Factor Models, Machine Learning, and Asset Pricing, *Annual Review of Financial Economics, 14*, 337–368.
- This paper highlights the potential of machine learning to uncover new factors and refine existing models, addressing challenges related to model evaluation and inference.
- The authors propose high-dimensional statistical methods to better estimate expected returns and risk exposures, using algorithms including random forests, boosted regression trees, and deep learning models.
- The application of these methods can be complex and requires further implementation to avoid overfitting.

(5) Gu, S., Kelly, B., & Xiu, D. (2020). Empirical Asset Pricing via Machine Learning, *Review of Financial Studies, 33(5)*, 2223–2273.
- This paper addresses challenges such as high dimensionality, overfitting, and the need for robust predictions in empirical asset pricing.
- The authors utilize tree-based models, regularization, and dimension-reduction methods to manage complexity and improve predictive accuracy.
- Additional insight is needed to develop robust frameworks for model selection and validation to ensure reliable out-of-sample performance.

(6) Celarek, A., Hermosilla, P., Kerbl, B., Ropinski, T., & Wimmer, M. (2022). Gaussian Mixture Convolution Networks, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2416–2425.
- This paper introduces the concept of integrating GMMs into convolutional networks for enhanced feature extraction and scalability in neural network models.
- The authors argue that the implementation of GMMs improves the learning capacity of convolutional neural networks by providing probabilistic structure to the data features.
- The probabilistic nature of GMMs requires careful tuning of parameters to avoid overfitting, which could compromise generalizability in predictive models for asset pricing.

(7) Li, H., & Liu, J. (2020). Clustering Analysis via Deep Generative Models with Mixture Models, *Journal of Machine Learning Research*, 21(1), 229–250.
- This paper Combines deep generative models (DGMs) with GMMs to improve clustering accuracy by applying probabilistic modeling to high-dimensional data.
- The authors demonstrate that GMMs, when used as a layer in DGMs, enhance the ability to discover latent clusters and improve the model's generalization performance.
- In the context of factor theme clustering, using GMMs may lead to misidentification of factor structures, especially if factors are highly correlated or have non-Gaussian distributions.

(8) Harvey, C. R., Liu, Y., & Zhu, H. (2016). ...and the Cross-Section of Expected Returns, *Review of Financial Studies*, 29(1), 1–38.
- This paper provides a comprehensive analysis of factor investing and its empirical challenges.
- The authors explore factor construction and its role in asset pricing, with a focus on robustness and replication.
- The study suggests that replicating factor models can lead to better understanding of asset return predictability.

(9) Zhang, Z., & Wu, X. (2020). Neural Mixture Models for Clustering and Dimensionality Reduction, *IEEE Transactions on Neural Networks and Learning Systems*, 31(9), 3319–3328.
- This paper proposes a neural network-based framework that incorporates GMM regularization to perform both clustering and dimensionality reduction simultaneously.
- The authors demonstrate effective discovery of latent structures in large-scale datasets, providing interpretability and reducing computational complexity.
- Hierarchical GMMs for factor theme clustering may struggle with the non-linear relationships often present in factor data.

(10) Wang, Y., & Lee, S. (2021). Deep Clustering with Variational Autoencoders and GMMs, *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
- This paper integrates Variational Autoencoders (VAE) with GMMs to form a deep clustering model, allowing for better separation of high-dimensional data into meaningful clusters.
- The authors propose a model that performs dimensionality reduction while maintaining required level of accuracy in unsupervised learning tasks.
- The integration of VAE and GMMs may struggle with scalability and convergence in high-dimensional datasets typical of financial markets, leading to challenges in capturing

the complexity of factor dependencies across different countries.

(11) Smith, A. L., & Johnson, R. M. (2021). Hierarchical GMMs in Neural Systems: Bridging Gaussian Mixture Models and Deep Learning, *Neural Computation*, 33(7), 1456–1482.

- This paper integrates GMMs with neural networks to address hierarchical representation challenges in complex datasets.
- The authors demonstrate that hierarchical GMMs enhance feature extraction and interpretation in neural systems.
- Additional asset pricing factors outside the identified themes may need to be investigated to evaluate their replicability.