Clustering and Machine Learning for Relative Value Trading in Corporate Bond Markets

Issam Fradi

Abstract

This paper presents a machine learning framework for optimizing portfolio strategies in corporate bond markets by clustering issuers based on their market behavior. Corporate bonds are selected due to their distinct risk-return profiles and yield potentials, which compensate for credit and liquidity risks [7]. Utilizing a Minimum Spanning Tree (MST) visualization derived from a correlation-based distance metric, we identify natural groupings of issuers, revealing underlying similarities in credit ratings and industry classifications.

We enhance these clusters using a machine learning approach to ensure they capture meaningful market dynamics. Granger causality tests are applied within clusters to detect predictive relationships among issuers, providing insights that support stock picking and relative value assessment. This methodology provides a robust framework for portfolio managers to enhance bond selection, optimize trading strategies, and manage risk in corporate bond portfolios.

---

1 Introduction and Motivation

Managing corporate bond portfolios presents numerous challenges, especially due to the limitations of traditional credit ratings and the complexities associated with liquidity. Credit ratings from agencies like Moody's, S&P, and MSCI often lag behind real-time market dynamics, leading portfolio managers to seek alternative methods for assessing issuer risk. This reliance on outdated methods can result in suboptimal portfolio decisions and missed opportunities for optimization.

Liquidity is another critical concern, with inconsistencies often complicating trade execution and the identification of alternative issuers with comparable risk-return profiles. These challenges are further amplified in modern corporate bond markets, where traditional assessment methods do not fully address the multifaceted nature of issuer risk and market behavior.

To address these issues, this paper proposes a machine learning-based approach for clustering corporate bond issuers. The methodology begins with constructing a Minimum Spanning Tree (MST) using a correlation-based distance metric, as developed by Mantegna (1999) [1] and refined by López de Prado [5]. This MST helps in identifying the optimal number of clusters by highlighting the structure of market relationships, ensuring that closely linked issuers share similar characteristics, such as credit ratings or industry affiliations.

With this foundational structure, we cluster issuers into groups that reflect meaningful market behaviors. Within these clusters, we apply Granger causality tests [2] to uncover predictive relationships among issuers, providing insights into potential causality in spread movements.

By leveraging these clustering results, we aim to optimize portfolio strategies through relative value assessments. This involves identifying undervalued and overvalued bonds within each cluster, facilitating stock picking and trading strategies that are grounded in both quantitative metrics and fundamental analysis. Ultimately, our framework seeks to enhance portfolio performance and liquidity management in the corporate bond market, bridging the gap between theoretical finance and practical portfolio management.

2 Dataset and Features

The dataset utilized in this study is sourced from the Markit iBoxx platform, covering the period from January 2023. Our analysis focuses specifically on the automobile sector. The dataset includes comprehensive bond-level information such as the Option-Adjusted Spread (OAS), duration, market value, issuer, ISIN, and the Second Best Rating (SBR). While the SBR is not used as a direct input for the clustering and analysis processes, it serves as an important parameter for further interpretation and validation of the results, providing insights into the credit quality of the issuers. For this paper, we worked only with senior bonds to maintain consistency in assessing credit risk profiles across issuers.

Several preprocessing and feature engineering steps were implemented to prepare the data for analysis. First, the logarithm of the OAS (`log_OAS`) was computed to stabilize the variance and address skewness in the distribution of spreads. Next, the percentage change in the log OAS (`log_OAS_var`) was calculated to capture the spread evolution over time for each bond. To aggregate these volatilities at the issuer level, we weighted the `log_OAS_var` by the bond's duration and market value, producing a `weighted_log_OAS_var`. The sum of these weighted values was then divided by the sum of the weights to obtain a weighted average volatility for each issuer on each date (`weighted_avg_log_OAS_var`).

Outliers were identified and linked to specific issuer characteristics, such as companies in default or those with extreme yields associated with high credit risk or imminent bond maturity. These outliers were carefully examined to ensure they did not distort the overall analysis, aligning them with their respective credit ratings and risk profiles.

Different aggregation periods, such as weekly or monthly intervals, were tested to assess their impact on the stability of the volatility estimates. After extensive testing, daily variations were chosen for their ability to capture short-term market dynamics more effectively.

The processed data was then structured into a matrix where each entry represents the weighted average volatility for a given issuer on a specific date. This matrix forms the basis for computing the correlation matrix between issuers, which is subsequently used in the clustering analysis described in the Methods section.

Table 1 provides a summary of the key features derived from the data.

Feature	Description	Calculation
`log_OAS_var`	Percentage change in the log OAS	$\frac{\log(\text{OAS}_t)-\log(\text{OAS}_{t-1})}{\log(\text{OAS}_{t-1})}$
`weighted_log_OAS_var`	Weighted Spread by duration and market value	log_OAS_var × Duration × Market Value

Table 1: Summary of Key Features

By incorporating these steps, the dataset effectively captures the daily aggregated volatility measures for each issuer, providing a solid foundation for clustering issuers with similar market behavior and enhancing the robustness of the subsequent analysis.

3 Methods

3.1 Correlation-Based Metric

In financial analysis, correlations are frequently employed to measure the linear relationships between different assets. However, the standard Pearson correlation coefficient has significant limitations: it does not satisfy the mathematical properties required of a metric, specifically non-negativity, the triangle inequality, and the identity of indiscernibles. These properties are crucial for interpreting "closeness" or "distance" between points in a meaningful way [5].

To address these limitations, we use a correlation-based distance metric, which transforms the Pearson correlation coefficient into a true metric. Consider two random vectors X and Y of size T, with a correlation estimate $\rho(X, Y)$ defined as:

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sigma(X)\sigma(Y)},$$

where $\sigma(\cdot)$ denotes the standard deviation. The correlation-based distance metric $d_p(X, Y)$ is then defined as:

$$d_p(X, Y) = \sqrt{\frac{1}{2}(1 - \rho(X, Y))}.$$

This transformation ensures that the distance metric satisfies all the properties of a true metric, making it particularly effective for clustering analysis [1].

However, the effectiveness of this metric relies on the assumption that the random vectors X and Y follow a bivariate normal distribution [6]. Without this assumption, the linear relationship quantified by the correlation coefficient may be misleading, potentially compromising the reliability of the distance metric.

The bivariate normal distribution is defined for two random variables X and Y with means μ_X and μ_Y, standard deviations σ_X and σ_Y, and correlation coefficient ρ. The joint probability density function (PDF) is given by:

$$f(X, Y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(X-\mu_X)^2}{\sigma_X^2} + \frac{(Y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(X-\mu_X)(Y-\mu_Y)}{\sigma_X\sigma_Y}\right]\right).$$

This joint PDF captures the dependence structure between X and Y and is crucial for validating the use of correlation as a measure of linear dependence. If X and Y do not follow a bivariate normal distribution, the correlation may not fully capture their relationship, making the correlation-based metric less meaningful.

Other measures, such as marginal and joint entropy, conditional entropy, and mutual information, were also considered for capturing the dependencies between variables (Appendices A.2–A.5). While these methods can provide a more comprehensive view of dependencies, they do not satisfy the properties of a true metric, such as the triangle inequality, which limits their applicability in metric-based clustering approaches [5].

Therefore, the correlation-based metric remains a suitable choice due to its compatibility with metric-based clustering methods and its alignment with the assumptions of bivariate normality common in fixed income markets. For further mathematical derivations, please see Appendices A.1 and A.2 [5, 6].

3.2 Clustering Methodology for Issuer Analysis

Clustering techniques in corporate bond markets are employed to group issuers exhibiting similar market behaviors, offering a dynamic alternative to traditional credit ratings for analyzing issuer relationships. By using a correlation-based distance metric, we capture structural relationships between issuers based on their market activity. This approach enables the formation of clusters that reflect real-time dependencies, making them highly relevant for portfolio management decisions.

The clustering process begins with constructing a distance matrix D derived from the correlation matrix, where correlations are transformed into distances to satisfy metric properties such as non-negativity and the triangle inequality. Ensuring that the clustering method adheres to these metric properties is crucial for the robustness of the clustering outcomes.

To refine the clustering, we employ the Minimum Spanning Tree (MST) approach, calculated using Kruskal's algorithm (Algorithm 1). The MST provides a visual representation of the closest relationships between issuers, where each node represents an issuer, and the edges denote the minimal connections based on the distance metric. This method effectively highlights clusters that not only capture statistical similarities but also align with fundamental characteristics, such as credit ratings and industry sectors (Figure 3).

Algorithm 1 Kruskal's Algorithm for Minimum Spanning Tree

Data: Graph $G = (V, E)$ with vertices V and edges E
Result: Minimum Spanning Tree T
begin
  Initialize an empty tree T and a forest where each vertex is its own tree Sort all edges E in non-decreasing order of their weights **foreach** *edge* (u, v) *in sorted edge list* **do**
    Find the root of the trees containing u and v **if** *the roots are different* **then**
      Add edge (u, v) to T Merge the trees containing u and v
    end
  end
  return T
end

Kruskal's algorithm, with a time complexity of $O(E \log E)$, is computationally efficient for analyzing relationships among a large number of issuers, as the sorting of edges is the most intensive step. This efficiency makes it well-suited for applications requiring quick and accurate analysis of issuer connectivity.

By examining the MST structure, we observe that issuers with similar credit ratings or those within the same industry tend to cluster together, forming groups that are both statistically robust and economically meaningful. The optimal number of clusters is determined using the Elbow method, which involves plotting the within-cluster sum of squared distances (WSS) against different values of k. The point where the rate of decrease slows significantly—known as the elbow—suggests the optimal number of clusters:

$$\text{WSS}(k) = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2,$$

where C_i represents the i-th cluster and μ_i is its centroid. This method, combined with MST analysis, ensures that the clustering results are grounded in both quantitative metrics and fundamental analysis.

These clusters are particularly useful for liquidity management. When facing liquidity constraints with a specific issuer, portfolio managers can identify alternative issuers within the same cluster that exhibit similar market behaviors, allowing for portfolio adjustments without significantly altering the risk-return profile.

Overall, this data-driven clustering methodology provides a powerful framework for issuer analysis, enabling portfolio managers to uncover hidden relationships, manage liquidity risks, and optimize their trading strategies. By integrating quantitative metrics with fundamental insights through the MST, the resulting clusters are both practical and actionable for portfolio management.

3.3 Granger Causality Test

The Granger Causality Test (GCT) is a statistical hypothesis test used to determine whether one time series can predict another. In the context of our fixed income analysis, the GCT is applied within clusters to explore potential causal relationships between issuers, aiding in identifying which issuer's movements may influence another. This insight can be leveraged to develop more informed trading strategies [2].

Consider two stationary time series X_t and Y_t. We say that X_t "Granger-causes" Y_t if past values of X_t provide statistically significant information about future values of Y_t that is not already contained in past values of Y_t.

Hypotheses

- **Null Hypothesis (H_0)**: X_t does not Granger-cause Y_t, meaning that past values of X_t do not provide any additional information about the future values of Y_t beyond what is already provided by past values of Y_t.
- **Alternative Hypothesis (H_1)**: X_t Granger-causes Y_t, meaning that past values of X_t do provide additional information about the future values of Y_t.

The test involves fitting two autoregressive models:

1. **Restricted model (without X_t)**:

$$Y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \epsilon_t$$

2. **Unrestricted model (with X_t)**:

$$Y_t = \beta_0 + \sum_{i=1}^{p} \beta_i Y_{t-i} + \sum_{i=1}^{p} \gamma_i X_{t-i} + \epsilon_t$$

where ϵ_t is the error term, p is the number of lags, and α_i, β_i, and γ_i are coefficients to be estimated. The test statistic is based on the difference between the residual sums of squares (RSS) of these two models.

F-Test Statistic

$$F = \frac{(RSS_r - RSS_{ur})/p}{RSS_{ur}/(n-2p-1)}$$

where RSS_r and RSS_{ur} are the residual sums of squares for the restricted and unrestricted models respectively, n is the number of observations, and p is the lag length.

The significance level α is set to 0.05. If the F-statistic is significant at this level, we reject the null hypothesis, concluding that X_t Granger-causes Y_t. Otherwise, we do not reject the null hypothesis.

The GCT provides valuable insights into issuer relationships within clusters. For example, if an issuer A Granger-causes issuer B, then movements in the spreads of issuer A might predict movements in issuer B. This information could be critical in constructing a relative value trading strategy, allowing portfolio managers to anticipate market movements and enhance trading decisions.

4 Results and Discussion

This section presents the findings from our clustering analysis, combined with Granger causality tests, to understand the relationships among bond issuers in the automobile sector. The clustering was conducted on all issuers in the sector, while the Granger causality test focused specifically on the issuers closest to Ford Motor Credit Co LLC, identified from the clustering results.

Ford Motor Credit Co LLC was specifically chosen for the Granger causality analysis due to its unique position with fundamentally connected peers that span both investment grade (IG) and high yield (HY) categories. Ford's credit ratings of BBB- (S&P), Ba1 (Moody's), and BBB- (Fitch), along with its Second Best Rating (SBR) of BBB-, place it at the borderline of IG and HY. This makes Ford an ideal candidate for examining cross-credit relationships, as its connections include a mix of other companies and suppliers with varied credit profiles. This approach provides a comprehensive view of issuer relationships based on market behavior and potential predictive linkages.

4.1 Clustering Analysis

The clustering algorithm, leveraging a correlation-based distance metric [5], categorized the issuers into three distinct groups (see Figure 7). Each cluster captures different market behaviors and credit risk profiles:

- **Cluster 1: High Yield Issuers** (Figure 4) primarily consists of issuers with lower credit ratings, such as BB+ to B. These bonds exhibit higher risk and volatility, which is characteristic of high-yield bonds that appeal to risk-seeking investors looking for higher returns.

- **Cluster 2: Investment Grade Issuers** (Figure 5) includes bonds with stronger credit ratings, such as A and BBB. These issuers have lower default risk and more stable returns, aligning with the behavior expected from investment-grade bonds. This cluster fits well with portfolio strategies aiming to balance risk and return while maintaining credit quality.

- **Cluster 3: Mixed Category Issuers** (Figure 6) is a heterogeneous cluster containing both investment-grade and high-yield bonds. As Table 2 shows, this cluster spans a wide range of credit ratings from A to BB. This mixed composition suggests that Cluster 3 captures issuers whose market behavior does not strictly follow traditional credit classifications. The presence of both high-yield and investment-grade bonds indicates complexities that neither fundamental credit risk nor conventional metrics could easily disentangle, which is why the cluster appears as a mix. From a portfolio management perspective, Cluster 3 could be further refined by reallocating issuers based on qualitative assessments or keeping them grouped to leverage the potential diversification benefits.

4.2 Integration with Granger Causality Tests

To further analyze the relationships among issuers, we conducted Granger causality tests on the closest issuers to Ford Motor Credit Co LLC, as identified by the clustering process. Figure 9 displays the results, revealing several statistically significant predictive relationships. For instance, issuers such as BMW International Investment BV, ZF Finance GmbH, and Gestamp Automocion SA exhibit causality with Ford Motor Credit Co LLC, indicating that their spread movements may influence Ford's spread dynamics. Figure 10 further visualizes these co-integration relationships, highlighting the interconnectedness of credit risk within the sector. The acyclic graph shows directed edges that indicate causality directions, which can serve as leading indicators for market behavior. This analysis is valuable for portfolio managers as it provides insights into the propagation of spread movements, enabling anticipatory adjustments in bond positions. It is important to note that Granger causality tests are typically used to assess systematic contagion risk rather than systematic trading strategies, underscoring their broader implications in understanding sector-wide credit conditions. To complement this analysis, we employed the NETS (Network Estimation for Time Series) framework introduced by Barigozzi and Brownlees (2019) [8], which models a large panel of time series as a sparse Vector Autoregression (VAR) system. The NETS approach captures both Granger causality and contemporaneous partial correlations using LASSO (Least Absolute Shrinkage and Selection Operator) estimation. The main LASSO estimation formula in NETS is given by:

$$\hat{\theta}_T = \arg\min_{\theta \in \mathbb{R}^m} \left[\frac{1}{T} \sum_{t=1}^{T} \ell(\theta; y_t, \hat{c}_T) + \frac{\lambda_G}{T} \sum_{k=1}^{p} \sum_{i,j=1}^{n} |\alpha_{ijk}| \hat{w}_{T,ijk} + \frac{\lambda_C}{T} \sum_{l,h=1}^{n} |\rho_{lh}| \hat{w}_{lh} \right]$$

where λ_G and λ_C are tuning parameters, and $\hat{w}_{T,ijk}$ and $\hat{w}_{lh}$ are adaptive LASSO weights. This estimation method simultaneously captures the network structure of the time series data, effectively identifying significant interconnections within a high-dimensional dataset. The NETS framework, applied to our distance matrix of bond issuers, proves robust for financial econometrics and risk analysis, offering a nuanced view of how issuer relationships and risk dynamics unfold within the market.

4.3 Yield Curve Visualization and Interpretation

To gain deeper insights into the risk-return profiles of the clustered issuers, we utilized the annual yield curve as a visualization tool instead of the Option-Adjusted Spread (OAS) initially used for clustering (Figure 8). The interpolated yield curves for each cluster highlight distinct market behaviors, enabling a more nuanced assessment of bond valuations: bonds positioned below the yield curve are identified as expensive or rich, while those above are considered cheaper. This rich-cheap differentiation is pivotal for relative value trading strategies, where portfolio managers can exploit these market inefficiencies by purchasing undervalued bonds (above the curve) and selling overvalued ones (below the curve).

Understanding these rich-cheap dynamics goes beyond quantitative metrics and necessitates a thorough grasp of the underlying issuer fundamentals. This combination of analysis is essential for portfolio managers and analysts to make informed decisions, especially when executing strategies like mean reversion or pairs trading. For example, mean reversion strategies can be employed to capitalize on the tightening or widening spreads of bonds that deviate from their expected positions relative to the interpolated yield curve. Within each cluster, bonds with ISINs significantly above the curve (indicating widening spreads) may be preferred as they present potential opportunities for tightening back to the mean, thereby offering attractive entry points for portfolio adjustments and optimizing returns.

Beyond relative value trading, the insights from this yield curve analysis extend to broader portfolio management applications. For macro and credit strategists, this approach provides a valuable tool for assessing sector-wide dynamics and refining market positioning and risk assessment. By identifying clusters with distinct yield curve slopes and shapes, strategists can develop targeted approaches, such as deploying Credit Default Swap (CDS) strategies or constructing benchmarks, to hedge risks or enhance portfolio performance. The ability to pinpoint specific ISINs that are outliers within their clusters allows portfolio managers to make tactical adjustments that align with both broader market conditions and issuer-specific developments.

Overall, this yield curve visualization approach is not only integral to traditional portfolio optimization but also opens pathways for more sophisticated trading strategies, including mean reversion and pairs trading, which can be tailored to the unique characteristics of each bond cluster. By integrating this analysis into a comprehensive risk management framework, portfolio managers can navigate the complexities of corporate bond markets more effectively, leveraging both quantitative metrics and in-depth issuer knowledge to drive superior investment decisions. This multi-dimensional approach strengthens the portfolio's resilience and responsiveness to market shifts, ultimately enhancing performance in a dynamic market environment.

4.4 Implications for Portfolio Management and Trading

The integration of clustering and Granger causality tests provides a robust framework for optimizing corporate bond portfolios. The clustering results help portfolio managers construct optimal portfolios by identifying groups of issuers with similar market behaviors, which is particularly advantageous in low-liquidity scenarios. By selecting alternative issuers within the same cluster, managers can maintain desired exposure while effectively navigating liquidity constraints.

Granger causality tests further refine this framework by uncovering predictive relationships among issuers, allowing portfolio managers and traders to anticipate spread movements and implement relative value strategies with greater precision. This forward-looking insight enables managers to position ahead of expected market shifts, capturing identified inefficiencies to optimize returns and mitigate risk.

Overall, this integrated approach balances quantitative rigor with market intuition, equipping portfolio managers and traders with a comprehensive toolkit for constructing resilient portfolios that are responsive to evolving market dynamics. By enhancing liquidity management, optimizing trade execution, and deepening the understanding of issuer relationships, this methodology contributes to improved portfolio performance in the corporate bond market.

Table 2: Credit Rating Proportion by Clusters

Cluster SBR_issue	1	2	3
A+	-	12.2	-
A	-	44.8	26.2
A-	-	9.8	-
BBB+	-	13.3	10.3
BBB	-	8.6	14.8
BBB-	17.1	11.3	25.8
BB+	46.9	-	22.9
BB	22.2	-	-
BB-	12.3	-	-
B	1.5	-	-

5 Conclusion

This study presents a novel approach to understanding issuer relationships in the corporate bond market by integrating machine learning techniques with traditional financial analysis. Using a correlation-based distance metric for clustering, we identified groups that reflect meaningful market behaviors and credit risk profiles. Yield curve interpolation of these clusters further revealed bond valuation inefficiencies exploitable for trading strategies.

Applying Granger causality tests on issuers close to Ford Motor Credit Co LLC demonstrated that some issuers' spread movements predict others, enhancing the clustering framework with a predictive dimension. This combined approach supports more informed decisions on liquidity management and relative value trading, improving portfolio performance.

While quantitative methods significantly enhance the assessment of issuer risk and clustering compared to traditional credit ratings, they are sensitive to market data and underlying assumptions. The diverse nature of Cluster 3 highlights challenges in clustering mixed-characteristic issuers, suggesting the need for nuanced interpretation and potential refinement of clustering techniques.

Overall, this research provides a robust framework for understanding issuer relationships, managing risk, and optimizing returns in portfolio management. By bridging theoretical finance with practical application, this approach offers potential for broader use across different sectors, paving the way for further innovations in portfolio optimization and risk management.

6 Directions for Future Research

While the current methodology provides a strong foundation for clustering bond issuers and analyzing their market behavior, several avenues for enhancement could further refine its scope and precision. A key direction is the integration of hybrid clustering methods that combine quantitative data with qualitative insights from credit analysts, offering a more adaptable and context-aware clustering approach, especially for complex cases like Cluster 3.

Exploring alternative distance metrics, such as those based on entropy or mutual information (Appendix A.2–A.5), could better capture non-linear relationships among issuers, leading to a more nuanced understanding of market dynamics. Additionally, incorporating macroeconomic variables, liquidity measures, or issuer-specific events into the clustering process could provide a comprehensive view of the factors influencing bond spreads.

Expanding the Granger causality tests beyond issuers closest to Ford Motor Credit Co LLC to include broader sector or cross-sector relationships may uncover extensive predictive linkages. Furthermore, leveraging advanced machine learning techniques, including deep learning or natural language processing (NLP), could analyze textual data from news, reports, and market commentary, providing additional insights that complement the current quantitative framework.

By pursuing these directions, future research could extend the framework, offering even more robust tools for managing corporate bond portfolios in dynamic and complex market environments.

References

[1] R. N. Mantegna, "Hierarchical Structure in Financial Markets," *The European Physical Journal B*, vol. 11, pp. 193–197, 1999.

[2] G. Kirchgässner, J. Wolters, and U. Hassler, *Introduction to Modern Time Series Analysis*, 2nd ed., Springer, Berlin, Heidelberg, 2013, ch. 3, pp. 95–122. [Section used: Granger Causality Test]

[3] J. B. Kruskal, "On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem," *Proceedings of the American Mathematical Society*, vol. 7, pp. 48–50, 1956.

[4] H. Loberman and A. Weinberger, "Formal Procedures for Connecting Terminals with a Minimum Total Wire Length," *Journal of the ACM*, vol. 4, no. 4, pp. 428–433, 1957.

[5] M. López de Prado, *Machine Learning for Asset Managers*, Cambridge University Press, 2020. [Chapter used: Chapter 2, A Correlation-Based Metric]

[6] Y. L. Tong, *The Multivariate Normal Distribution*, Springer-Verlag, New York, 1990, ch. 2, pp. 6–21. [Section used: The Bivariate Normal Distribution]

[7] L. Martellini, P. Priaulet, and S. Priaulet, *Fixed-Income Securities: Valuation, Risk Management and Portfolio Strategies*, John Wiley & Sons, 2003.

[8] M. Barigozzi and C. Brownlees, "NETS: Network estimation for time series," *Journal of Applied Econometrics*, vol. 34, pp. 347–364, 2019. [DOI: 10.1002/jae.2676]

APPENDIX

A.1. CORRELATION-BASED METRIC

Consider two real-valued vectors X, Y of size T, and a correlation variable $\rho[X,Y]$, with the only requirement that $d[X,Y] = \rho[X,Y]\sigma[X]\sigma[Y]$, where $\sigma[X,Y]$ is the covariance between the two vectors, and $\sigma[\cdot]$ is the standard deviation. Note that Pearson's is not the only correlation to satisfy these requirements.

Let's prove that $d[X,Y] = \sqrt{\frac{1}{2}(1-\rho[X,Y])}$ is a true metric. First, the Euclidean distance between the two vectors is $d[X,Y] = \sqrt{\sum_{t=1}^{T}(x_t - y_t)^2}$. Second, we z-standardize those vectors as $x = \frac{X-\bar{X}}{\sigma[X]}$, $y = \frac{Y-\bar{Y}}{\sigma[Y]}$. Consequently, $0 \leq \rho[X,Y] = \rho[P,Y]$. Third, we derive the Euclidean distance $d[x,y]$ as,

$$d[x,y] = \sqrt{\sum_{t=1}^{T}(x_t - y_t)^2} = \sqrt{\sum_{t=1}^{T}x_t^2 + \sum_{t=1}^{T}y_t^2 - 2\sum_{t=1}^{T}x_t y_t} = \sqrt{T + T - 2T\sigma[x,y]}$$

$$= \sqrt{2T\left(1-\rho[x,y]\right)} = \sqrt{4Td[x,Y]}$$

In other words, the distance $d[X,Y]$ is a linear multiple of the Euclidean distance between the vectors $\{X,Y\}$ after z-standardization, hence it inherits the true-metric properties of the Euclidean distance.

Similarly, we can prove that $d[X,Y] = \sqrt{1-|\rho[X,Y]|}$ is also a true metric. In order to do that, we redefine $y = \frac{Y-\bar{Y}}{\sigma[Y]}\text{sgn}[\rho[X,Y]]$, where $\text{sgn}[\cdot]$ is the sign operator, so that $0 \leq \rho[X,Y] = |\rho[X,Y]|$. Then,

$$d[x,y] = \sqrt{2T\left(1-|\rho[x,y]|\right)} = \sqrt{2Td[x,Y]}$$

A.2. Marginal and Joint Entropy

The concept of entropy helps to quantify the uncertainty in a random variable. Let X be a discrete random variable that takes a value x from the set S_X with probability $p[x]$. The entropy $H[X]$ of X is defined as:

$$H[X] = -\sum_{x \in S_X} p[x]\log p[x],$$

where the convention $0\log[0] = 0$ is used, and entropy can be interpreted as the amount of uncertainty associated with X.

For two discrete random variables X and Y, the joint entropy $H[X,Y]$ is defined as:

$$H[X,Y] = -\sum_{x \in S_X}\sum_{y \in S_Y} p[x,y]\log p[x,y].$$

The joint entropy represents the uncertainty in both X and Y.

A.3. Conditional Entropy

The conditional entropy of X given Y, denoted $H[X|Y]$, represents the amount of uncertainty remaining about X after Y is known. It is defined as:

$$H[X|Y] = H[X,Y] - H[Y] = -\sum_{y \in S_Y} p[y]\sum_{x \in S_X} p[x|y]\log p[x|y],$$

where $p[x|y]$ is the conditional probability of X given Y. The conditional entropy is zero if knowing Y completely determines X.

A.4. Mutual Information

Mutual information $I[X,Y]$ measures the reduction in uncertainty of one random variable due to the knowledge of another random variable. It is defined as:

$$I[X,Y] = H[X] + H[Y] - H[X,Y] = \sum_{x \in S_X}\sum_{y \in S_Y} p[x,y]\log\frac{p[x,y]}{p[x]p[y]}.$$

Mutual information is always non-negative and symmetric, i.e., $I[X,Y] = I[Y,X]$. It is zero when X and Y are independent. However, it is important to note that mutual information is not a metric as it does not satisfy the triangle inequality. Despite this, it is a useful measure in clustering contexts.

The concept of mutual information helps in understanding the dependencies between variables and has been used for clustering algorithms that aim to maximize the mutual information between clusters and the original data.

A.5. Bivariate Normal Distribution

Let $Z_1, Z_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, which we will use to build a general bivariate normal distribution.

$$f(z_1, z_2) = \frac{1}{2\pi} \exp\left[-\frac{1}{2}\left(z_1^2 + z_2^2\right)\right]$$

We want to transform these unit normal distributions to have the following arbitrary parameters: $\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho$.

$$X = \sigma_X Z_1 + \mu_X$$
$$Y = \sigma_Y[\rho Z_1 + \sqrt{1-\rho^2} Z_2] + \mu_Y$$

First, let's examine the marginal distributions of X and Y:

$$X = \sigma_X Z_1 + \mu_X = \sigma_X \mathcal{N}(0,1) + \mu_X = \mathcal{N}(\mu_X, \sigma_X^2)$$
$$Y = \sigma_Y[\rho Z_1 + \sqrt{1-\rho^2} Z_2] + \mu_Y = \sigma_Y[\rho \mathcal{N}(0,1) + \sqrt{1-\rho^2}\mathcal{N}(0,1)] + \mu_Y = \mathcal{N}(\mu_Y, \sigma_Y^2)$$

Second, we can find $\text{Cov}(X,Y)$ and $\rho(X,Y)$:

$$\text{Cov}(X,Y) = E[(X - E(X))(Y - E(Y))] = \sigma_X \sigma_Y \rho$$
$$\rho(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = \rho$$

Consequently, if we want to generate a Bivariate Normal random variable with $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ where the correlation of X and Y is ρ, we can generate two independent unit normals Z_1 and Z_2 and use the transformation:

$$X = \sigma_X Z_1 + \mu_X$$
$$Y = \sigma_Y[\rho Z_1 + \sqrt{1-\rho^2} Z_2] + \mu_Y$$

We can also use this result to find the joint density of the Bivariate Normal using a 2d change of variables. The joint density of X and Y is given by:

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right)\right]$$

Obviously, the density for the Bivariate Normal is complex, and it only becomes more difficult when considering higher-dimensional joint densities. We can write the density in a more compact form using matrix notation:

$$f(x) = \frac{1}{2\pi(\det\Sigma)^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]$$

Where:

$$\mu = \begin{pmatrix}\mu_X \\ \mu_Y\end{pmatrix}, \quad \Sigma = \begin{pmatrix}\sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2\end{pmatrix}$$

We can confirm our results by checking the value of $(\det\Sigma)^{-1/2}$ and $(x-\mu)^T\Sigma^{-1}(x-\mu)$ for the bivariate case.

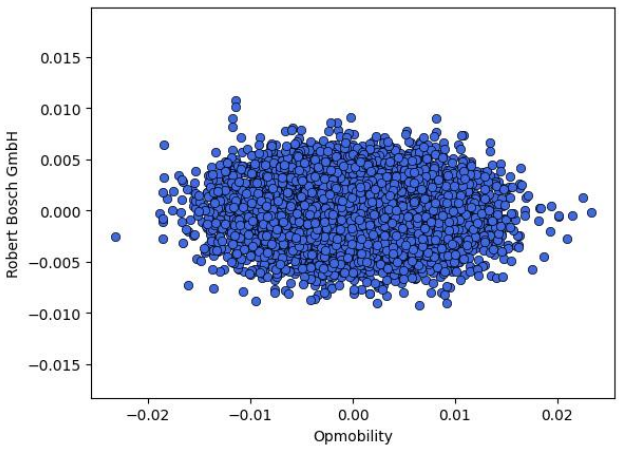

Figure 1: Opmobility and Robert Bosch GmbH

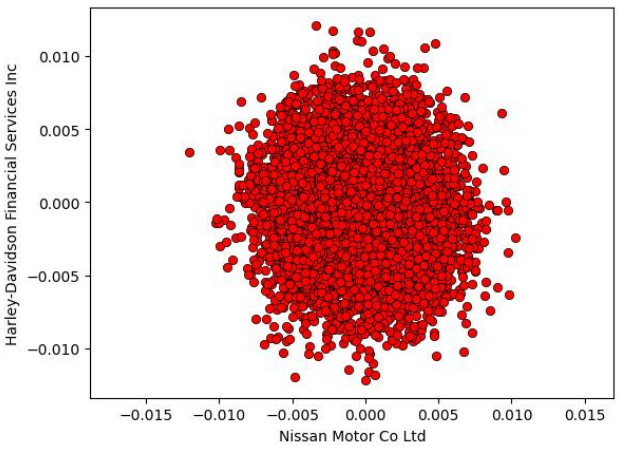

Figure 2: Nissan Motor Co Ltd and Harley-Davidson Financial Services Inc

As illustrated in Figures 1 and 2, the joint distributions of the weighted average log OAS variances for the selected issuers are examined to validate the appropriateness of the cross-correlation metric used in the clustering methodology. Figure 1, showing Opmobility and Robert Bosch GmbH, depicts an elliptical shape typical of a bivariate normal distribution, albeit with some deviations indicating market-specific idiosyncrasies.

Conversely, Figure 2, which represents the relationship between Nissan Motor Co Ltd and Harley-Davidson Financial Services Inc, exhibits a tighter and near-linear pattern, closely fitting a bivariate normal distribution. This alignment supports the use of the cross-correlation metric, affirming that the spread behaviors of these issuers follow the assumptions required for accurate clustering.

These pairs exemplify the broader validation process conducted across all issuer pairs, ensuring that the distributional assumptions of the cross-correlation metric hold, thereby enhancing the robustness and reliability of subsequent clustering outcomes.

B.1. Minimum Spanning Tree

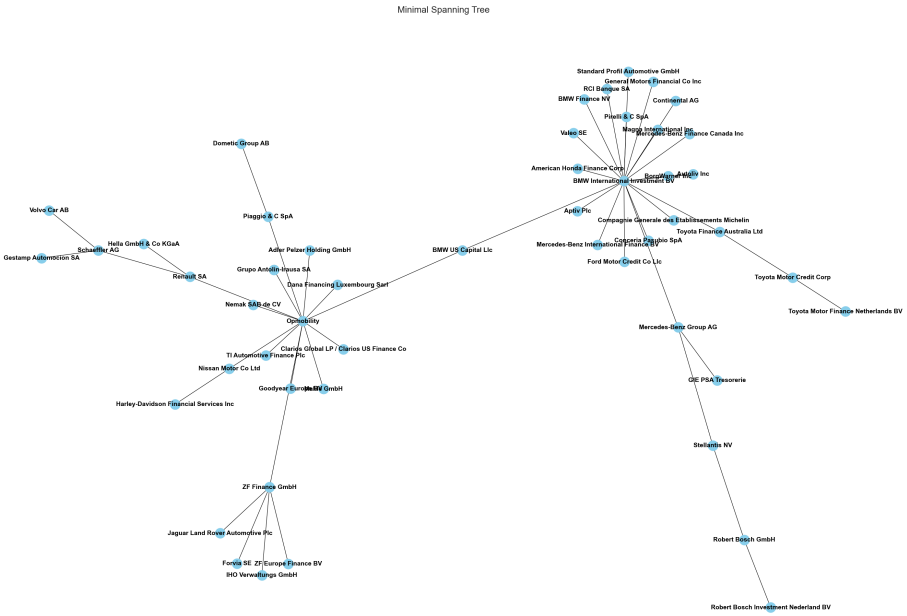

Figure 3: Minimum Spanning Tree generated using Kruskal's algorithm

The Minimum Spanning Tree (MST) highlights clear distinctions between issuer types. Vertices surrounding BMW International Investment BV predominantly include Investment Grade (IG) issuers, reflecting lower risk, while the branches extending from OPMobility are associated with High Yield (HY) issuers, indicating higher market volatility and risk.

B.2. Visualization of Clustering Results

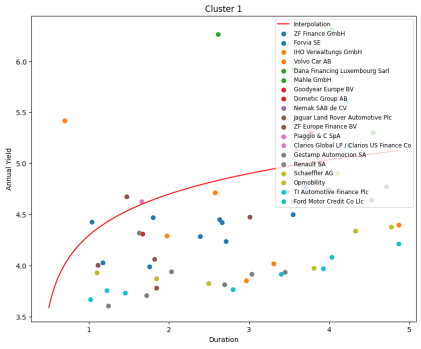

Figure 4: Cluster 1: High Yield Issuers

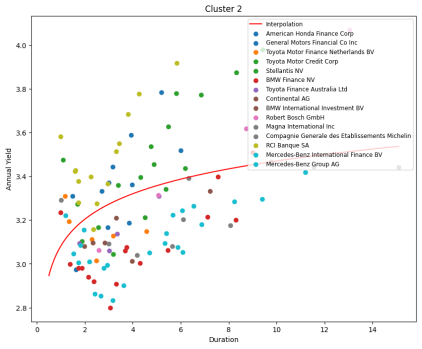

Figure 5: Cluster 2: Investment Grade Issuers

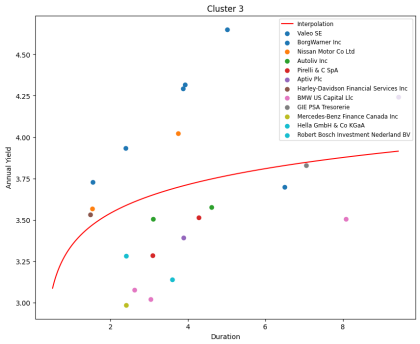

Figure 6: Cluster 3: Mixed Category Issuers

Figure 7: Comparison of Clustering Results: Visualization of Issuer Clusters based on Market Behavior.

B.3. Yield Curve Interpolation Across Clusters

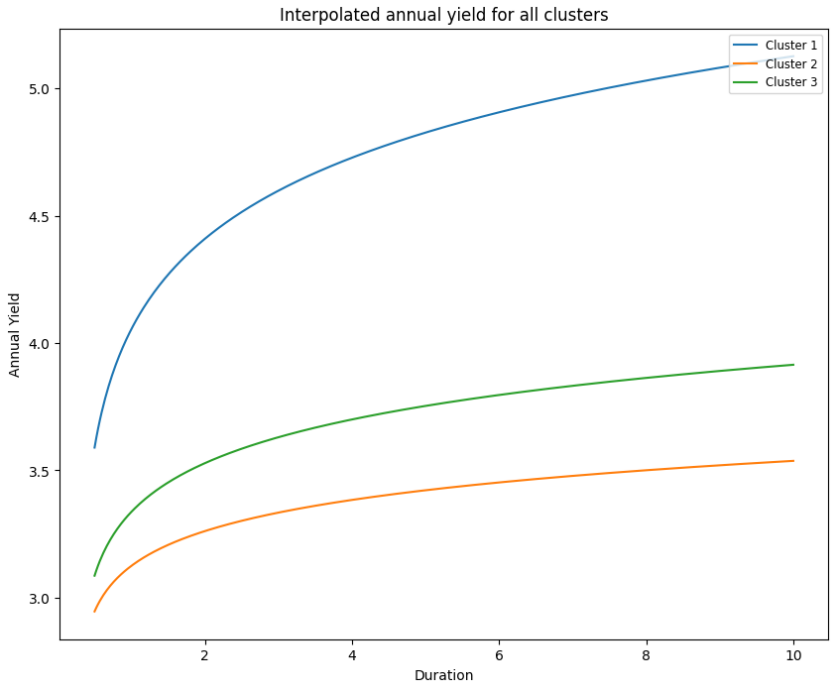

Figure 8: Interpolated Yield Curves for Different Clusters: Comparison of the annual yield curves for Clusters 1, 2, and 3 based on duration. Each curve reflects the distinct risk-return profiles and market behaviors of the clustered issuers.

C.1. Granger Causality Test and Co-Integration Results

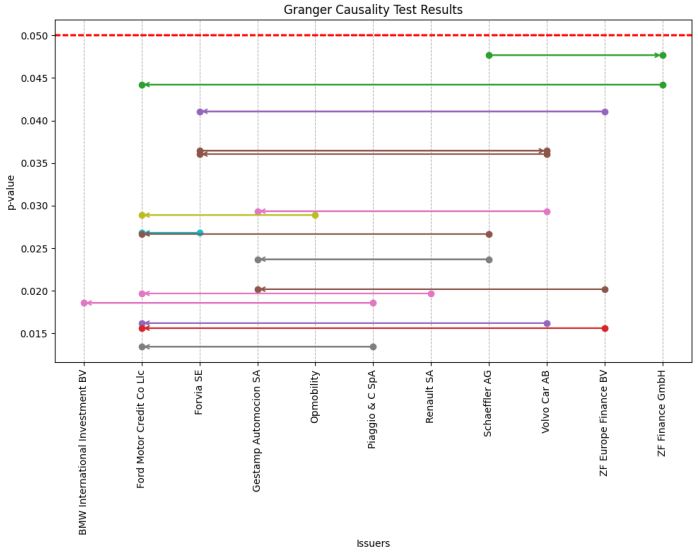

Figure 9: Granger Causality Test Results for Issuers Closest to Ford Motor Credit Co LLC

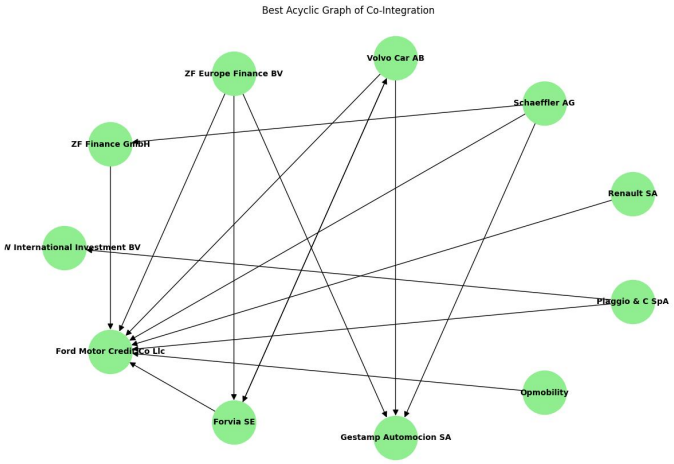

Figure 10: Best Acyclic Graph of Co-Integration Among Issuers

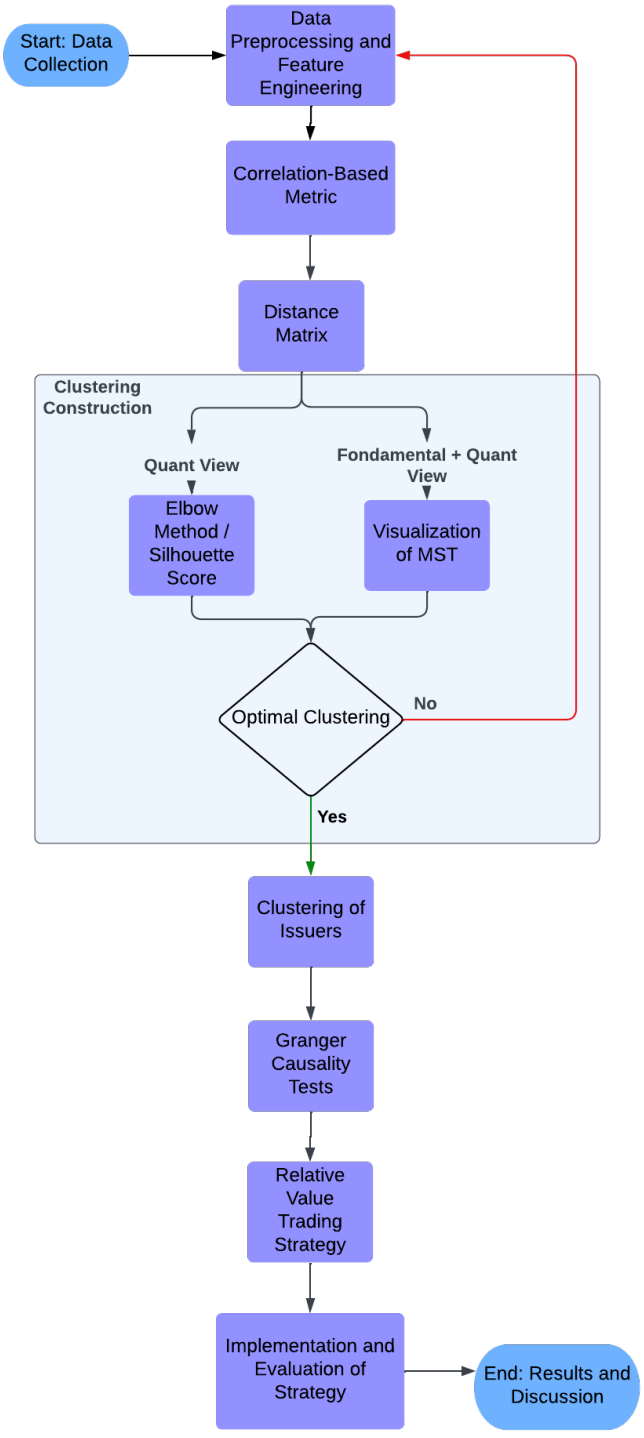

Figure 11: Flowchart: Data-Driven Clustering and Trading Strategy Development