GROUP WORK PROJECT - 1

GROUP NUMBER: 9908
MScFE 600: FINANCIAL DATA
World Quant University

July 7, 2025

Group Declaration and Contact Information

FULL LEGAL NAME	LOCATION COUNTRY	EMAIL	MARK "X" FOR NON-CONTRIBUTING MEMBER
Dibakar Sigdel	USA	sigdeldkr@gmail.com	
Huayi TANG			
Pawendtaoré Samuel YAMEOGO			

Statement of Integrity: By typing the names of all group members in the text boxes below, you confirm that the assignment submitted is original work produced by the group (excluding any non-contributing members identified with an "X" above).

Team Member 1	Dibakar Sigdel
Team Member 2	Huayi TANG
Team Member 3	Pawendtaoré Samuel YAMEOG

Use the box below to explain any attempts to reach out to a non-contributing member. Type (N/A) if all members contributed.

(N/A)

1 Task 1: Data Quality

1.1 Structured Data Deficiencies in Financial Datasets

Structured financial datasets, such as those involving securities transactions, are expected to uphold strict standards of accuracy, consistency, and interpretability. However, real-world data often diverge from these expectations. Consider a sample trading dataset involving equity transactions, where multiple quality issues undermine the reliability of downstream analyses.

First, inconsistencies in timestamp formatting are observed. Some entries record time in the ISO 8601 format (e.g., `2023-08-15T09:30:00Z`), while others adopt localized representations (e.g., `Aug 15, 2023 9:30 AM EST`). This inconsistency introduces difficulties in aligning events temporally and can result in erroneous inferences during time-series modeling.

Second, the asset identifiers used in the dataset vary in schema and resolution. Securities are variously represented using ticker symbols (e.g., `AAPL`), international identifiers such as ISINs (e.g., `US0378331005`), and internal proprietary codes (e.g., `EQ123`). The absence of a consistent and unique identifier scheme prevents effective merging across datasets and compromises traceability [11].

Third, anomalous numerical entries such as negative transaction volumes are present. Such entries violate domain-specific validity rules, as securities transactions cannot legitimately involve negative quantities.

1.2 Violation of Foundational Data Quality Dimensions

These issues reflect critical breaches in fundamental dimensions of structured data quality, as identified in data governance literature [9]:

- **Validity**: The presence of negative share volumes breaches the business logic constraints inherent to financial systems.
- **Consistency**: Timestamp heterogeneity impedes chronological alignment and affects any model relying on temporal coherence.
- **Uniqueness**: The use of divergent identification schemes (e.g., tickers vs. ISINs) introduces ambiguity, hindering instrument-level aggregation and auditability.

These violations can significantly distort empirical analyses, from market microstructure studies to portfolio optimization, by introducing structural noise and alignment errors.

1.3 Challenges in Unstructured Financial Data

Beyond structured datasets, unstructured data—such as text from financial social media platforms—presents its own set of quality challenges. For example, a corpus of tweets discussing market developments may exhibit several common defects.

One such issue is the truncation of text, which may arise due to API-imposed character limits or sampling constraints. Truncated messages often lack complete semantic context, impairing sentiment analysis and topic modeling.

Another issue involves the presence of duplicated or automated posts. Bots can repost identical content multiple times per minute, distorting frequency-based metrics and artificially inflating term salience [2].

Additionally, irrelevant or off-topic content—such as memes or humor not related to financial discourse—may be mixed into the dataset, introducing noise into machine learning pipelines intended for financial signal extraction.

1.4 Quality Degradation in Unstructured Domains

These deficiencies in unstructured data reflect further violations of data quality principles:

- **Incompleteness**: Truncated posts omit essential linguistic features necessary for accurate interpretation, reducing the effectiveness of natural language processing (NLP) models.
- **Non-Uniqueness**: Bot-generated duplication compromises the representational fidelity of the dataset, biasing models toward non-informative patterns.
- **Contextual Irrelevance**: Off-topic material violates the *fitness-for-use* principle by introducing thematic incoherence.
- **Unverifiable Provenance**: Absence of reliable user metadata hinders assessments of data trustworthiness, which is critical in compliance-sensitive domains.

In sum, both structured and unstructured data domains require rigorous quality assessments and preprocessing protocols to ensure their suitability for downstream financial modeling. Failure to address these issues can propagate systemic bias and degrade model validity, particularly in high-stakes applications such as trading algorithms and credit risk assessment.

2 Task 2: Yield Curve Modeling

2.1 Data Collection and Description

We retrieved U.S. Treasury yield data from the Federal Reserve Economic Data (FRED) platform using the `fredapi`. The dataset includes yield curve snapshots from January 2024 through June 2025, an 18-month period capturing a broad range of market environments, including inflationary pressures and monetary policy adjustments.

This period was chosen to:

- Capture dynamic economic conditions and central bank rate decisions,
- Provide enough observations for robust model calibration,
- Include maturities from 1 month to 30 years, covering short- to long-term rates.

The yield maturities included are: 1M, 3M, 6M, 1Y, 2Y, 3Y, 5Y, 7Y, 10Y, 20Y, and 30Y. Missing observations (e.g., on holidays) were retained for integrity and omitted only during the curve-fitting steps.

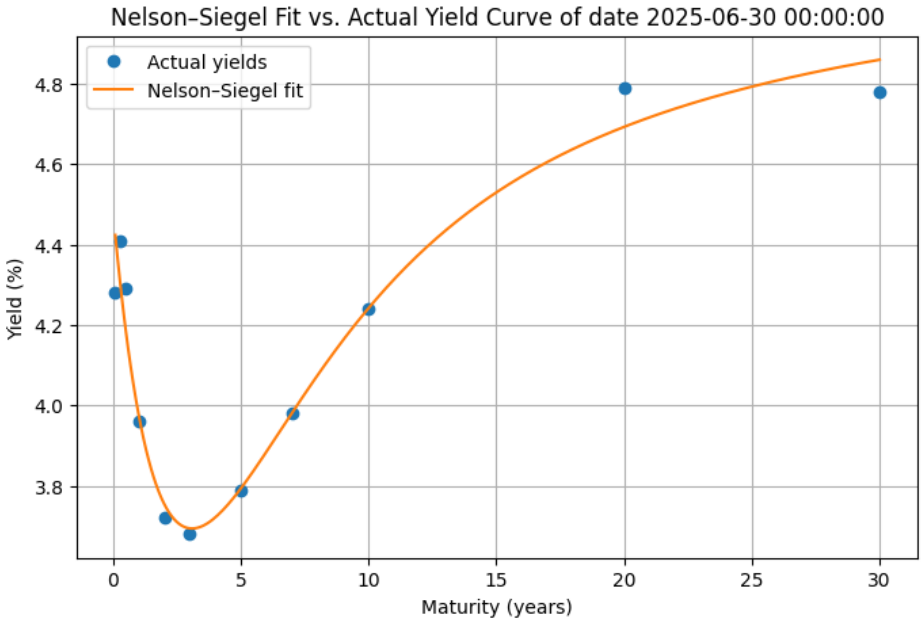

Figure 1: Actual U.S. Treasury Yield Curve (2025-06-30)

2.2 Modeling Approaches

We modeled the yield curve using two methods:

1. Cubic spline interpolation
2. Nelson–Siegel parametric model

2.2.1 Cubic Spline Interpolation

A cubic spline interpolates the observed yields with smooth piecewise polynomials, ensuring the curve passes exactly through all knot points. This method fits the yield curve exactly at the provided maturities, minimizing interpolation error.

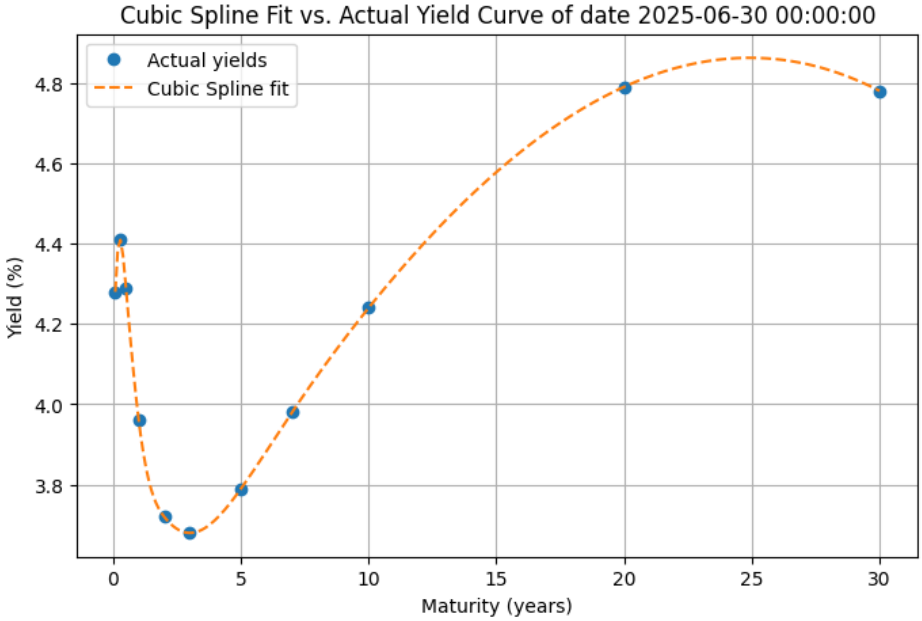

Figure 2: Cubic Spline Fit vs. Actual Yield Curve on 2025-06-30

Key characteristics of the spline fit:

- **Exact interpolation at knot points:** The spline passes through all observed yields by construction.
- **Short-term oscillation:** A spike between 1M and 3M indicates overfitting in low-data regions.
- **Smooth mid/long-term trend:** The spline follows market trends well for maturities from 5Y to 20Y, with a slight hump near 25Y.

2.2.2 Nelson–Siegel Model Specification

The Nelson–Siegel model expresses the yield curve as a function of maturity τ:

$$y(\tau) = \beta_0 + \beta_1 \cdot \frac{1 - e^{-\tau/\tau_d}}{\tau/\tau_d} + \beta_2 \cdot \left(\frac{1 - e^{-\tau/\tau_d}}{\tau/\tau_d} - e^{-\tau/\tau_d} \right) \tag{1}$$

where:

- β_0 captures the *level* or long-term yield,
- β_1 captures the *slope* (short vs. long-term),
- β_2 captures the *curvature* or mid-term hump,

- τ_d is the decay parameter determining the shape transition.

We calibrated the model using non-linear least squares optimization on the observed yield data.

2.3 Model Comparison

2.3.1 Goodness-of-Fit

To compare the models, we evaluated the Mean Squared Error (MSE) at observed maturities.

- **Cubic Spline:** Achieves perfect interpolation (MSE ≈ 0) due to exact fitting.
- **Nelson–Siegel:** Provides a close approximation, though small residuals remain.

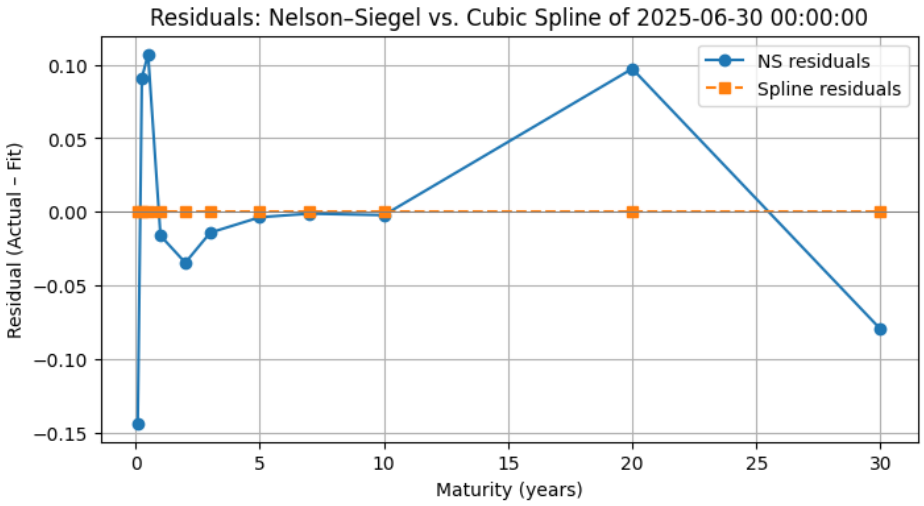

Figure 3: Residuals from Nelson–Siegel vs. Observed Yields

2.3.2 Interpretation

Nelson–Siegel Parameters (2025-06-30):

$$\begin{aligned}
\beta_0 &= 5.1948 \quad \text{(Level — long-run average yield)} \\
\beta_1 &= -0.7140 \quad \text{(Slope — short-term steepness)} \\
\beta_2 &= -3.8303 \quad \text{(Curvature — mid-term hump)} \\
\tau_d &= 2.2120 \quad \text{(Decay rate — time to curvature peak)}
\end{aligned}$$

Cubic Spline Interpretation:

- No global parameters; each maturity interval is fit locally.
- High flexibility, but no intuitive economic meaning for coefficients.
- More prone to local overfitting, especially when data are sparse.

2.4 Ethical Considerations of Smoothing

As discussed in Module 2, smoothing techniques can raise ethical concerns if used to obscure risks or distort data interpretation.

- The Nelson–Siegel model preserves essential yield curve structure (level, slope, curvature) using transparent and interpretable parameters.
- It is widely adopted in academia and industry for forecasting, valuation, and regulatory stress testing.
- Unlike spline methods, it does not fabricate new data points or exaggerate curve shape between observations.
- Ethical misuse occurs only when model assumptions are hidden or manipulated to support a specific narrative.

Conclusion: The Nelson–Siegel model is not inherently unethical. It is acceptable—provided its assumptions and limitations are clearly communicated—to use it for yield curve analysis and forecasting.

Summary

This task applied two powerful modeling techniques to the U.S. Treasury yield curve:

- Cubic spline interpolation offers exact fits but lacks interpretability and can overfit.
- The Nelson–Siegel model captures macroeconomic structure with interpretable parameters and smoother trends.

Both models serve valuable roles, but the Nelson–Siegel approach provides a more robust framework for economic interpretation and risk analysis in fixed-income applications.

3 Task 3: Exploiting Correlation

Modern financial data analysis extends beyond descriptive statistics and prediction to uncovering the underlying structure of market behaviors. A critical step in this process is identifying co-movements among financial variables—such as yields across maturities—using dimensionality reduction techniques. This section investigates the role of correlation and latent factor modeling through Principal Component Analysis (PCA), applied to both simulated and real-world yield data.

3.1 Simulated Data Analysis

To establish a theoretical baseline, we begin with synthetic data designed to contain no intrinsic correlations. Specifically, five independent Gaussian time series were generated, each representing hypothetical yield changes with a mean of zero and low variance. This setup mirrors daily movements in yield changes that are purely stochastic and uncorrelated.

PCA was then applied to this dataset. As expected in such an unstructured system, the explained variance was distributed relatively evenly among the five principal components. The absence of dominant eigenvalues confirmed the lack of latent common factors—an ideal characteristic of white noise-like data.

For this particular simulation:

- The first component explained 26.19% of the total variance,
- The second, 22.14%,
- And the third, 19.49%.

These results are consistent with theoretical expectations from random matrix theory, where variance is uniformly dispersed under null correlation conditions [7].

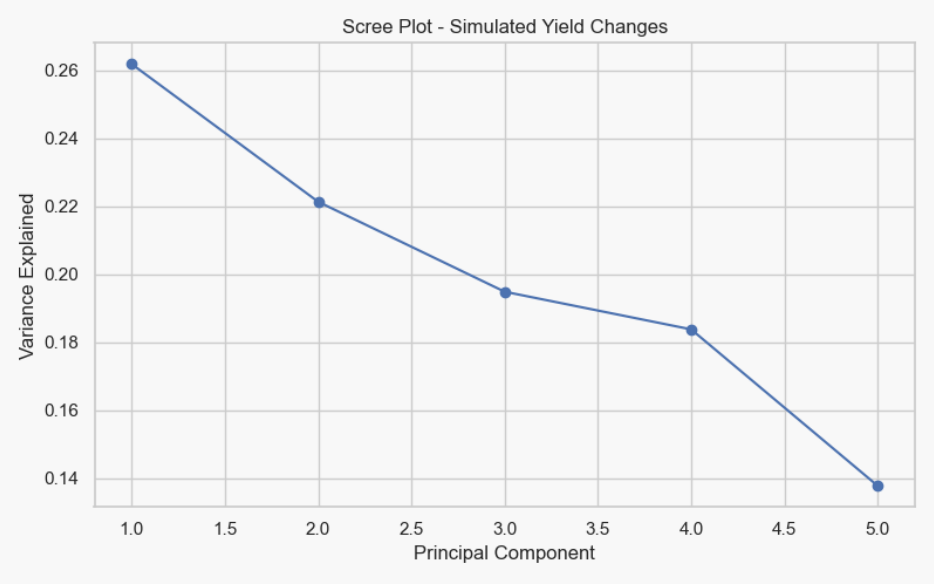

Figure 4: Scree plot of simulated yield changes

As shown in Figure 4, the gradual and nearly linear drop in explained variance across components is emblematic of an unstructured dataset, where no single latent dimension dominates.

3.2 Real Government Yield Data Analysis

We contrast this theoretical baseline with PCA conducted on real-world yield data. Specifically, we collected daily closing prices of five U.S. Treasury bond ETFs—TLT, IEF, SHY, VGSH, and EDV—covering a range of maturities and credit durations. The analysis spans six months, from 2025-01-08 to 2025-07-07, with daily log returns computed to induce stationarity and ensure comparability across assets [10].

The resulting PCA yielded a strikingly different profile. The first principal component alone accounted for 97.52% of the total variance, followed by the second (2.29%) and third (0.12%). This overwhelming dominance of the first component suggests that most of the variation in these bond yields can be explained by a single systemic factor—a phenomenon well-documented in empirical fixed-income research [8].

Interpretations of the principal components follow well-established economic logic:

- The first component—often labeled the *level factor*—represents parallel shifts in the entire yield curve, typically driven by macroeconomic variables such as inflation expectations or central bank interest rate policy.
- The second component corresponds to the *slope factor*, characterizing differences in movement between short- and long-term yields, commonly associated with changes in the yield curve steepness.
- The third component reflects *curvature*, capturing localized shifts around the mid-term maturities while holding the ends more stable.

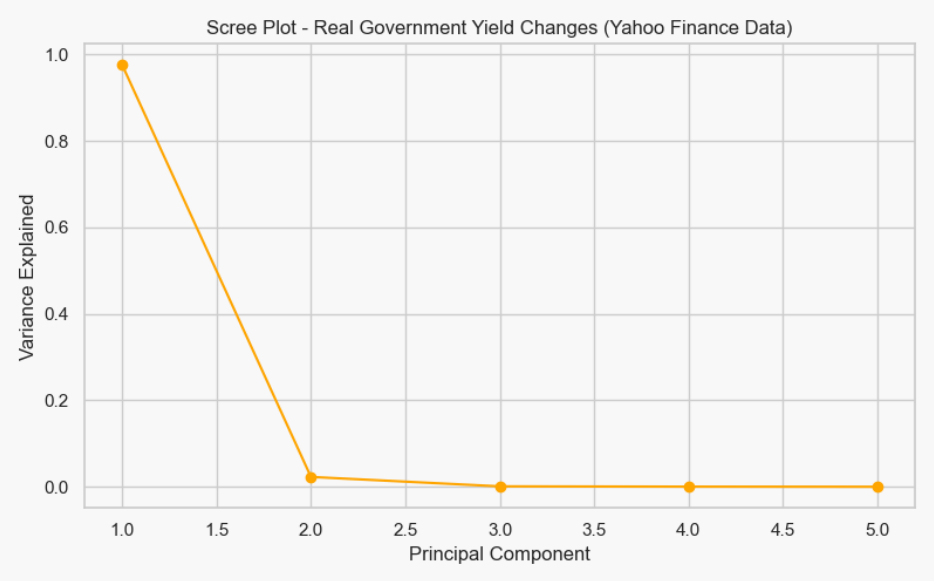

Figure 5: Scree plot of real government yield changes

Figure 5 vividly illustrates the dominance of the first principal component and the rapidly diminishing explanatory power of higher-order components. This reflects the underlying economic intuition that yield movements are often governed by a small number of systemic forces.

3.3 Comparison and Implications

The juxtaposition of simulated and empirical results emphasizes the fundamental role of correlation in financial data. In the simulated scenario, each variable contributes independently, resulting in a flat scree curve. In contrast, the real dataset exhibits strong latent structure, with co-movements driven by macro-level dynamics.

PCA thus emerges as a crucial tool for revealing latent drivers in yield curves. Its ability to distill high-dimensional data into a small set of orthogonal components is particularly valuable in portfolio management, risk factor modeling, and hedging strategies [4]. For instance, a fixed-income manager may use the level, slope, and curvature factors as proxies for systemic risks and construct immunized portfolios accordingly.

Overall, this exercise underscores the theoretical power and practical utility of PCA in understanding correlation structures, particularly in fixed-income markets where co-movements are per

4 Task 4: Empirical Analysis of ETFs: XLRE Sector Exposure

Exchange-Traded Funds (ETFs) serve as a liquid and diversified investment vehicle representing sector-specific or broad-market exposures. This section presents an empirical analysis of XLRE, the Real Estate Select Sector SPDR Fund, using advanced statistical methods to uncover the underlying sources of return variation across its constituents. Specifically, we employ covariance matrix analysis, Principal Component Analysis (PCA), and Singular Value Decomposition (SVD) to detect latent risk factors and co-movement patterns.

4.1 Data Collection and Return Computation

The dataset comprises daily adjusted closing prices for XLRE and its top 25 constituent holdings (including PLD, AMT, CCI, EQIX, PSA, SPG, DLR, and others), retrieved from Yahoo Finance over the period spanning January 8, 2025, to July 7, 2025, encompassing 121 trading days. Daily logarithmic returns were computed from the adjusted prices to ensure stationarity and comparability across securities [10].

Log-returns are particularly suited for financial modeling due to their additive properties over time and better distributional characteristics relative to simple percentage returns. They serve as a normalized representation of price changes and form the basis for numerous applications in finance, including volatility estimation, portfolio optimization, and factor modeling [3].

4.2 Covariance Matrix Analysis

We begin by analyzing the covariance matrix of the computed returns, which captures the joint variability of asset returns. A positive covariance implies that two assets tend to move in the same direction, while a negative covariance indicates opposite movement. In sector-focused ETFs such as XLRE, constituents are expected to exhibit positively correlated behavior due to shared exposure to macroeconomic drivers such as interest rates and real estate cycles [1].

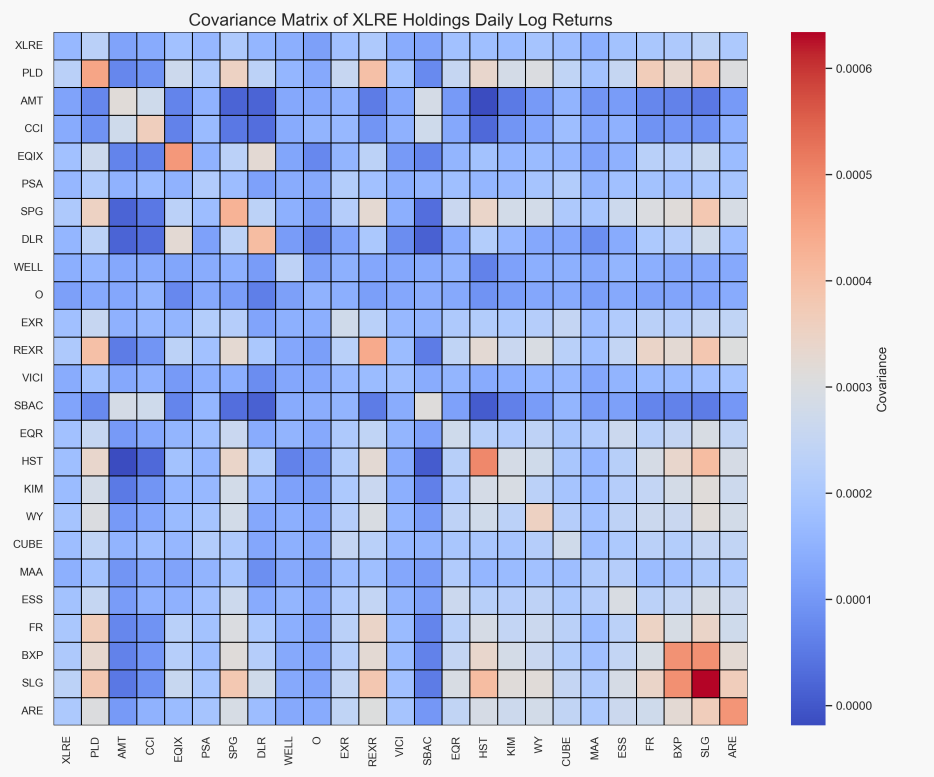

Figure 6: Covariance Matrix of XLRE Holdings Daily Log Returns

As illustrated in Figure 6, the covariance matrix exhibits predominantly positive values across constituent pairs. This reflects the inherent co-movement within the real estate sector and underscores the importance of further decomposition techniques to distinguish systematic from idiosyncratic components of risk.

4.3 Principal Component Analysis (PCA)

PCA is a widely used statistical method for reducing data dimensionality while preserving the variance structure. It achieves this by transforming correlated variables into orthogonal principal components (PCs), which are ranked by their explanatory power of the total variance [7].

The first few principal components usually capture most of the meaningful variation, especially in financial datasets where common economic factors drive asset returns. In our analysis, PCA applied to the XLRE returns revealed:

- The first component explains 59.50% of the total variance.
- The second explains 13.23%.
- The third explains 5.90%.

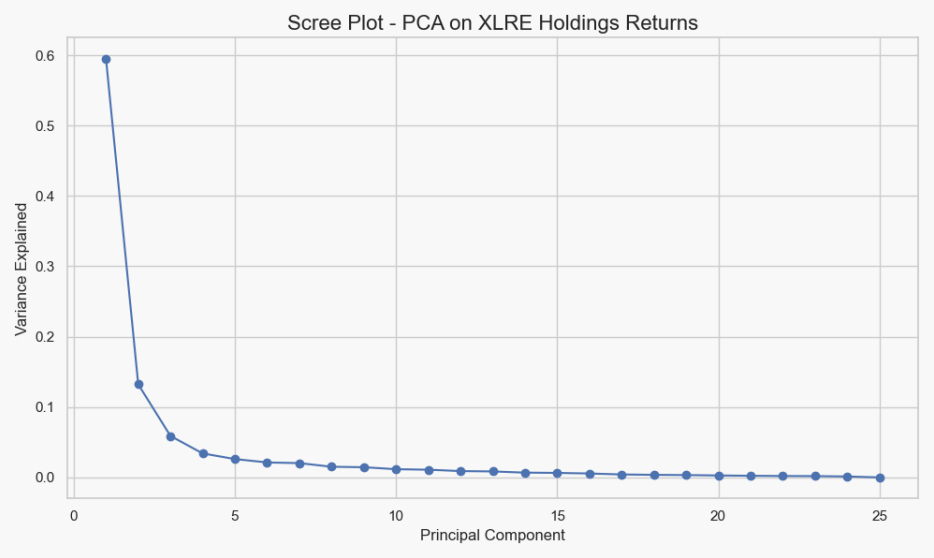

Figure 7: Scree Plot - PCA on XLRE Holdings Returns

As shown in Figure 7, the scree plot highlights a sharp drop in explained variance after the first component, suggesting the presence of a dominant market-wide factor. The first principal component typically represents a general real estate market trend—possibly linked to broad economic factors like interest rate sensitivity or GDP expectations. The second and third components often capture more specific exposures such as sub-sector behavior (e.g., commercial vs. residential REITs) or style tilts (e.g., growth vs. value).

4.4 Singular Value Decomposition (SVD)

SVD provides a numerically robust method of matrix factorization and is deeply connected to PCA when applied to mean-centered data. Specifically, SVD decomposes a matrix $X \in R^{n \times p}$ into:

$$X = USV^T$$

where U and V are orthogonal matrices and S is a diagonal matrix of singular values. These singular values correspond to the square roots of the eigenvalues of the covariance matrix, thus linking directly to the variance explained in PCA [6].

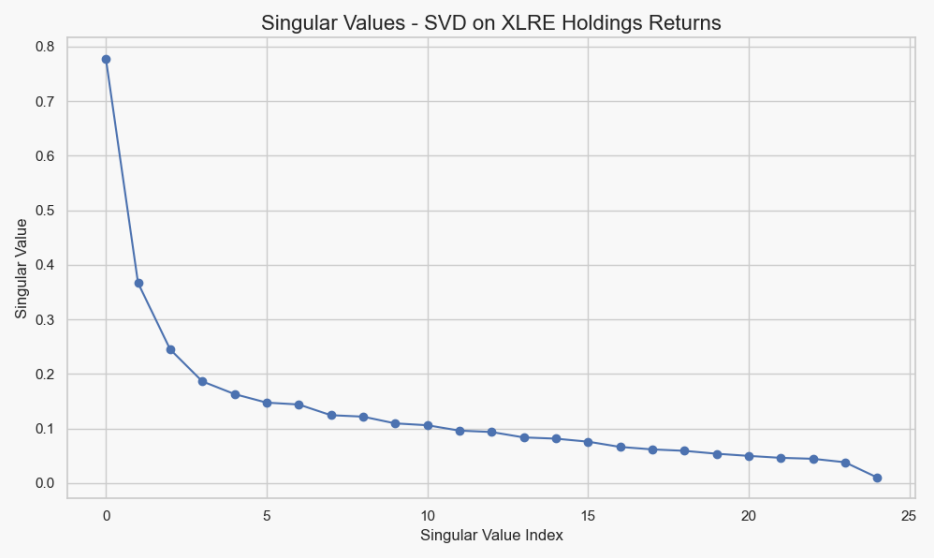

Figure 8: Singular Values - SVD on XLRE Holdings Returns

Figure 8 reveals that the largest singular values mirror the leading components from PCA. This validates the dominance of a few systemic drivers and demonstrates that both PCA and SVD are capable of uncovering the underlying low-rank structure of return variation in the real estate sector.

4.5 Comparison of PCA and SVD

Although PCA and SVD are mathematically interrelated, their computational perspectives differ. PCA is typically defined through the eigendecomposition of the covariance matrix, whereas SVD operates directly on the data matrix. From a numerical stability standpoint, SVD is often preferred, especially in high-dimensional settings where the covariance matrix may be ill-conditioned or singular [5].

For mean-centered data X, the principal components derived from PCA correspond exactly to the right singular vectors V from SVD, while the singular values provide a scale-invariant measure of component strength. Both methods converge on the same conclusion: a small number of orthogonal components explain most of the observed variation in the XLRE holdings.

4.6 Conclusion

The results of this empirical investigation underscore the importance of latent factor modeling in ETF analysis. The high degree of co-movement among XLRE constituents suggests that the portfolio is driven by a handful of systematic factors—primarily interest rate sensitivity and real estate market conditions.

The combination of PCA and SVD offers a rigorous approach to identifying these hidden structures. Such insights are invaluable for:

- **Factor-Based Investing:** Designing portfolios that target specific return drivers while mitigating exposure to undesired factors.
- **Risk Management:** Monitoring and hedging latent risk exposures within a sector.

- **Portfolio Simplification:** Reducing dimensionality while retaining essential economic information.

These tools provide a foundation for more robust and interpretable asset allocation strategies, especially in sector-specific or thematic investing contexts.

References

[1] Andrew Ang. *Asset Management: A Systematic Approach to Factor Investing.* Oxford University Press, 2009.

[2] Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *Proceedings of the Findings of EMNLP*, pages 1644–1650, 2020.

[3] John Y Campbell, Andrew W Lo, and Craig A MacKinlay. *The Econometrics of Financial Markets.* Princeton University Press, 1997.

[4] Frank J. Fabozzi, Lionel Martellini, and Philippe Priaulet. Principal component analysis of yield curves. In *Handbook of Fixed-Income Securities.* McGraw-Hill, 2010.

[5] Gene H Golub and Charles F Van Loan. *Matrix Computations.* Johns Hopkins University Press, 4th edition, 2013.

[6] Per Christian Hansen. Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion. *SIAM Review*, 34(4):561–580, 1988.

[7] Ian T Jolliffe. *Principal Component Analysis.* Springer, 2002.

[8] Robert Litterman and José Scheinkman. Common factors affecting bond returns. *Journal of Fixed Income*, 1(1):54–61, 1991.

[9] Diane M Strong, Yang W Lee, and Richard Y Wang. Data quality in context. *Communications of the ACM*, 40(5):103–110, 1997.

[10] Ruey S. Tsay. *Analysis of Financial Time Series.* Wiley, 3rd edition, 2010.

[11] Anuja Wickramasinghe and Anurag Sharma. Data quality management in financial analytics: A review. In *Proceedings of the 53rd Hawaii International Conference on System Sciences*, pages 5402–5411, 2020.