GROUP WORK PROJECT # __1_ MScFE 600: FINANCIAL DATA

GROUP NUMBER: __8902_____

FULL LEGAL NAME	LOCATION (COUNTRY)	EMAIL ADDRESS	MARK X FOR ANY NON-CONTRIBUTING MEMBER
Haowen Shen	China	0xconnor3@gmail.com	
Justine Otieno Kitoto	Kenya	kitotojustin@gmail.com	
Siddharth Dixit			X

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	**Haowen Shen**
Team member 2	**Justine Otieno Kitoto**
Team member 3	

Use the box below to explain any attempts to reach out to a non-contributing member. Type (N/A) if all members contributed.
Note: You may be required to provide proof of your outreach to non-contributing members upon request.

Haowen Shen and Justine Kitoto tried reaching out to Siddharth Dixit through the group discussion forum to no avail.

Report

1. TASK1: Data Quality

a. Example of Poor Quality Structured Financial Data

Company_ID, Revenue, Q1_Profit, Q2-Profit, Q3 Profit, Q4_Profit, Year

ABC123, $2.5M, 450K, $512,000, 498000, missing, 2023

XYZ456, $3.7m, 780K, "in progress", $820,000, $792k, 22-23

LMN789, "$4,200,000", Unknown, $1.1M, -$250k, 1200000, FY2023

PQR321, 3250000, 620000, 680,000, , 710.5k, 2023

b. Analysis of Poor Quality Structured Data

The structured financial data example fails to meet good quality standards in several fundamental ways. First, there is a complete lack of consistency in formatting, with monetary values represented as different units (millions vs. thousands) and using inconsistent notation ($, k, m, missing commas). Second, the data exhibits completeness issues with missing values indicated in different ways ("missing", empty cells) and ambiguous entries like "in progress" or "Unknown" that cannot be used for calculations. Third, the accuracy is compromised by inconsistent date formats (2023, 22-23, FY2023) making it impossible to perform reliable year-over-year comparisons or temporal analysis.

c. Example of Poor Quality Unstructured Financial Data

Financial Analyst Report - Acme Corp

Q3 performance was good, better than last time. Revenue increased but not as much as competitors. Some concerns about cash flow but overall the outlook seems positive-ish. Expenses went up due to that thing that happened with the suppliers. Management mentioned something about new markets during the call. I think the stock might go up. The P/E ratio is reasonable compared to industry standard. There were some regulatory issues pending; I'll check on those later.

Recommendation: Consider buying, but maybe wait for Q4 results.

Note to self: Double-check the numbers from that PDF when I get a chance.

d. Analysis of Poor Quality Unstructured Data

This unstructured financial analysis fails to meet good quality standards through several critical deficiencies. The report lacks specificity and precision, using vague qualitative descriptors ("good," "better than last time," "positive-ish") without providing any concrete metrics or quantitative data to support these assessments. There is a serious problem with completeness, as evidenced by unfinished analysis, undefined references ("that thing"), and explicitly mentioned missing information ("double-check the numbers"). The reliability of the data is compromised by subjective, hedging language ("I think," "might go up," "maybe wait") and inconsistent professional standards, resulting in a document that would be unsuitable for making informed investment decisions or conducting meaningful financial analysis.

TASK2: Yield Curve Modeling: Nelson-Siegel vs. Cubic Spline

1. Introduction

This report presents a comparative analysis of two yield curve modeling techniques—Nelson-Siegel and Cubic Spline—applied to Chinese government securities. Yield curves are fundamental tools in finance that depict the relationship between interest rates and time to maturity. Accurate modeling of these curves is essential for pricing fixed-income securities, managing risk, and informing monetary policy decisions.

2. Data Analysis

For this analysis, we used Chinese government bond yield data from April 11, 2025, across eight key maturities:

Maturity	Yield (%)
3M	1.3800
6M	1.3816
1Y	1.3978
3Y	1.4315
5Y	1.4895
7Y	1.6024
10Y	1.6568
30Y	1.8627

The data reveals a consistently upward-sloping yield curve, which is typical in normal economic environments where investors demand higher yields for longer maturities to compensate for increased risk. The curve shows a gradual increase in yields from short to medium-term maturities, with a more pronounced rise in the long-term segment. This shape suggests market expectations of stable near-term economic conditions with potential for higher growth or inflation in the distant future.

3. Nelson-Siegel Model

Model Specification

The Nelson-Siegel model is a parametric approach that describes the yield curve using an exponential function with four parameters:

$$y(t) = \beta_0 + \beta_1[(1-e^{(-t/\tau)})/(t/\tau)] + \beta_2[(1-e^{(-t/\tau)})/(t/\tau) - e^{(-t/\tau)}]$$

Where:

- $y(t)$ is the yield at maturity t
- β_0 represents the long-term interest rate (level)
- β_1 represents the short-term component (slope)
- β_2 captures the medium-term component (curvature)
- τ is a time decay factor controlling the rate at which components decay

Parameter Estimates

Based on our fitting process, we obtained the following parameter values:

Parameter	Value	Interpretation
β_0	1.9729	Long-term interest rate (level)
β_1	-0.5831	Short-term component (slope)
β_2	-0.6955	Medium-term component (curvature)
τ	2.5815	Decay factor

Parameter Interpretation

The parameter values provide meaningful economic insights into the current Chinese bond market:

- **β_0 (1.9729)**: This represents the long-term level of interest rates, suggesting that as maturity approaches infinity, yields would converge to approximately 1.97%. This is slightly higher than the 30-year yield (1.86%), indicating the model anticipates rates to eventually stabilize above current long-term levels.

- **β_1 (-0.5831)**: The negative value confirms an upward-sloping yield curve. This parameter represents the difference between short and long-term rates. The negative value indicates that short-term rates are lower than long-term rates, which is consistent with normal economic expectations.
- **β_2 (-0.6955)**: This negative value influences the medium-term component of the yield curve, creating a slight hump in the curve. The negative value suggests that medium-term yields rise less steeply than what would be predicted by a simple short-to-long term transition.
- **τ (2.5815)**: This decay factor determines how quickly the short and medium-term components decay across the maturity spectrum. The relatively low value indicates that the transition from short to long-term rates occurs primarily in the shorter segment of the curve.

Model Performance

- R-squared: 0.9943
- RMSE: 0.0121

The Nelson-Siegel model provides an excellent fit to the observed data, explaining over 99% of the variation in yields across maturities. The small RMSE value indicates that, on average, the model's predictions deviate from actual yields by only about 1.2 basis points.

4. Cubic Spline Model

Model Specification

The Cubic Spline model is a non-parametric approach that fits a piecewise polynomial of degree 3 between each pair of adjacent knots (maturity points). The resulting function is continuous up to the second derivative.

For each interval [t_i, t_{i+1}], the cubic polynomial is defined as:

$S_i(t) = a_i + b_i(t-t_i) + c_i(t-t_i)^2 + d_i(t-t_i)^3$

Our implementation utilized 8 knots (one for each observed maturity point) and generated 7 sets of polynomial coefficients to connect these points.

Model Performance

- R-squared: 1.0000 (perfect fit)
- RMSE: 0.0000 (no error at observation points)

As expected, the Cubic Spline model provides a perfect fit to the observed data points. This is a mathematical certainty since cubic splines are designed to pass exactly through each knot point.

5. Visual Analysis of Model Comparison

The graph reveals important differences in how the two models fit the yield curve:

1. **General Shape**: Both models capture the overall upward-sloping trend of the yield curve.
2. **Fit at Observation Points**:
    - The Nelson-Siegel model (red line) provides a smooth curve that closely approximates but doesn't exactly match all observed points.
    - The Cubic Spline model (blue line) passes perfectly through every observed yield point.
3. **Behavior Between Observation Points**:
    - The Nelson-Siegel curve maintains a consistent upward trajectory throughout all maturities.
    - The Cubic Spline exhib
    - its unexpected behavior, particularly in the region between 10 and 30 years, where it dips significantly before rising sharply to meet the 30-year yield point.
4. **Long-term Extrapolation**:
    - The Nelson-Siegel model suggests a continued gradual increase in yields beyond 30 years.
    - The Cubic Spline shows an extremely steep increase after 30 years, which is likely unrealistic.

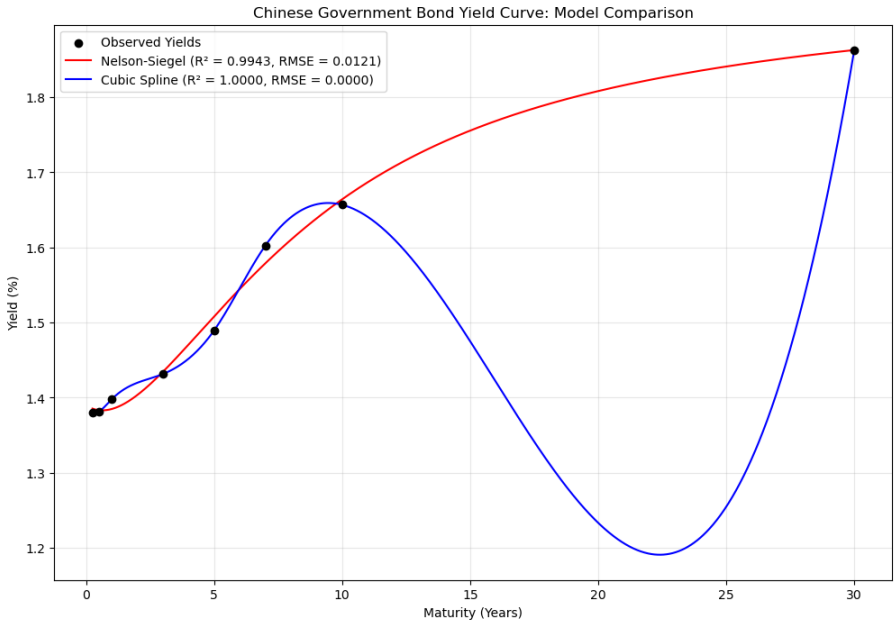


Fig1: Chinese Government Bond Yield Curve: Model Comparison

6. Detailed Model Comparison

Fit Quality

1. **Nelson-Siegel**:
    - Very high R-squared value (0.9943) indicates excellent overall fit
    - Small RMSE (0.0121) shows minimal average deviation from observed yields
    - Provides a smooth, economically plausible curve through the term structure
    - Makes small trade-offs in exact fit to maintain economic consistency
2. **Cubic Spline**:
    - Perfect R-squared (1.0000) and zero RMSE at observation points
    - Between observation points, especially in the 10-30 year range, the model exhibits counterintuitive behavior that is inconsistent with typical yield curve dynamics
    - The dramatic dip between 10-25 years followed by a sharp rise to the 30-year point suggests potential overfitting

Interpretation

1. **Nelson-Siegel**:
    - Provides clear economic interpretation through its parameters
    - The positive β_0 (1.9729) accurately represents the long-term interest rate level
    - The negative β_1 (-0.5831) correctly captures the upward slope of the yield curve
    - The negative β_2 (-0.6955) reflects medium-term dynamics and the slightly accelerated increase in the middle segment of the curve
    - The decay factor τ (2.5815) indicates the transition point where short-term effects begin to diminish
2. **Cubic Spline**:
    - Lacks economic interpretation despite perfect mathematical fit
    - The coefficients are purely mathematical constructs with no financial meaning
    - The behavior between data points, particularly the significant dip between 10-25 years, contradicts economic intuition about interest rate term structures
    - Would be problematic for pricing securities with maturities that fall between observation points

Practical Applications

1. **Nelson-Siegel**:

- Better suited for economic analysis, monetary policy decisions, and forecasting
- More reliable for extrapolating yields beyond the observed range
- Provides consistent yield estimates for securities with non-standard maturities
- Parameters can be tracked over time to monitor changes in market expectations

2. **Cubic Spline**:
   - Excellent for precise interpolation at or very near observed maturities
   - Problematic for risk management due to unrealistic behavior between observations
   - Could lead to mispricing of securities with maturities between observed points
   - The dip in the 10-30 year segment could create arbitrage opportunities that don't exist in reality

7. Ethical Considerations

The question of whether Nelson-Siegel smoothing is unethical requires careful consideration in light of our results. Based on the analysis of Chinese government bond data:

1. **Representational Accuracy**: The Nelson-Siegel model introduces a very small amount of error (RMSE of 0.0121) while maintaining economic consistency. This level of smoothing preserves the essential features of the yield curve while removing potential noise or anomalies. In this context, the smoothing appears justified and ethically sound.
2. **Prevention of Overfitting**: As demonstrated by the Cubic Spline model, perfect mathematical fit can lead to economically unrealistic results. The dramatic dip between 10-30 years in the spline model is likely an artifact of the mathematical technique rather than a genuine economic signal. By avoiding such overfitting, the Nelson-Siegel model may actually provide a more truthful representation of the underlying economic reality.
3. **Transparency**: The ethical use of Nelson-Siegel depends on transparency. As long as the model parameters, methodology, and fit statistics are clearly communicated (as we have done in this report), users can make informed judgments about the reliability of the smoothed curve.
4. **Purpose-Appropriate Modeling**: The Nelson-Siegel model is particularly well-suited for macroeconomic analysis and monetary policy applications where the general shape and economic interpretation of the curve are more important than exact fits at specific points. Using the appropriate model for the intended purpose is ethically sound.

Based on these considerations, the Nelson-Siegel smoothing of the Chinese government bond yield curve appears ethically justified, especially given:

- The high quality of fit (R-squared of 0.9943)
- The small magnitude of smoothing (RMSE of only 0.0121)
- The clear economic interpretability of the resulting parameters

- The avoidance of unrealistic behavior between observation points

8. Conclusion

Both the Nelson-Siegel and Cubic Spline models offer valuable approaches to yield curve modeling, but with significantly different characteristics that become apparent in our analysis of Chinese government bonds:

1. **Nelson-Siegel** provides an excellent balance between fit quality (R-squared: 0.9943) and economic interpretability. Its smooth, consistent curve avoids the potential pitfalls of overfitting while capturing the essential features of the term structure. The economically meaningful parameters offer valuable insights into market expectations about future interest rates.
2. **Cubic Spline** achieves mathematical perfection at observation points but at the cost of economic plausibility between those points. The significant dip in the 10-25 year segment followed by a sharp rise to the 30-year point demonstrates how purely mathematical approaches can generate curves that contradict financial theory and market behavior.

The optimal choice between these models depends on the specific application:

- For pricing securities with standard maturities that match the observation points, both models perform equally well.
- For economic analysis, monetary policy decisions, and pricing securities with non-standard maturities, the Nelson-Siegel model provides more reliable and economically consistent results.
- For applications where exact interpolation at specific points is the primary concern and the behavior between those points is less important, the Cubic Spline may be preferable.

This analysis demonstrates that successful yield curve modeling requires balancing mathematical fit with economic theory. The Nelson-Siegel model, with its strong theoretical foundation and excellent empirical performance, provides a more holistic representation of the Chinese government bond yield curve despite introducing a small amount of smoothing. This smoothing is not only ethically justified but may enhance the economic validity of the resulting curve.

Task 4: Empirical Analysis of ETFs

1. **<u>Understanding Daily Returns, PCA, and SVD in the Context of XLRE Holdings</u>**

In this analysis, we examined the 30 largest holdings of the Real Estate Select Sector SPDR Fund (XLRE) to understand their daily return behavior and uncover underlying patterns in risk and performance using advanced techniques like Principal Component Analysis (PCA) and Singular Value Decomposition (SVD).

2. **Daily Returns and Rolling Metrics**

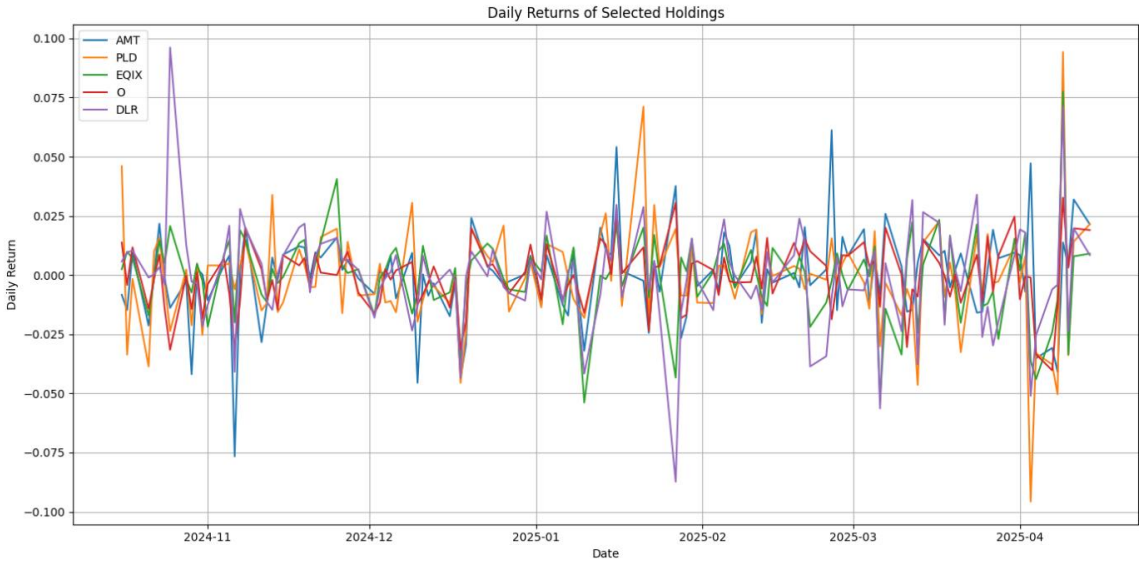
Daily returns, which show the percentage change in a security's value from day to day, are a fundamental part of financial data analysis. They enable analysts to use a normalized scale to gauge the relative performance and volatility of assets. It is easier to compare securities of varying sizes and takes compounding into account when returns are used instead of raw prices. Furthermore, returns are crucial for any risk or portfolio analysis since statistical characteristics like mean, standard deviation, covariance, and correlation have greater importance on return series than on raw prices.

We began by extracting and plotting the **daily returns** of each REIT using their adjusted closing prices over the past 6 months. This allowed us to standardize their behavior and compare them on a relative scale.

In order to document changing patterns and hazards throughout time, we superimposed:

- **Rolling Averages (21-day):** highlighted short- to medium-term momentum and indicated which stock was steadily increasing or decreasing in strength.
- **Rolling Volatility (21-day):** The variability of returns was measured using volatility (21-day). Volatility spikes in several equities indicated periods of increased uncertainty, which could have been related to company-specific changes or macroeconomic news.

The tactical distribution of assets depends on these measurements, which also aid in spotting risk warnings or opportune entrances.

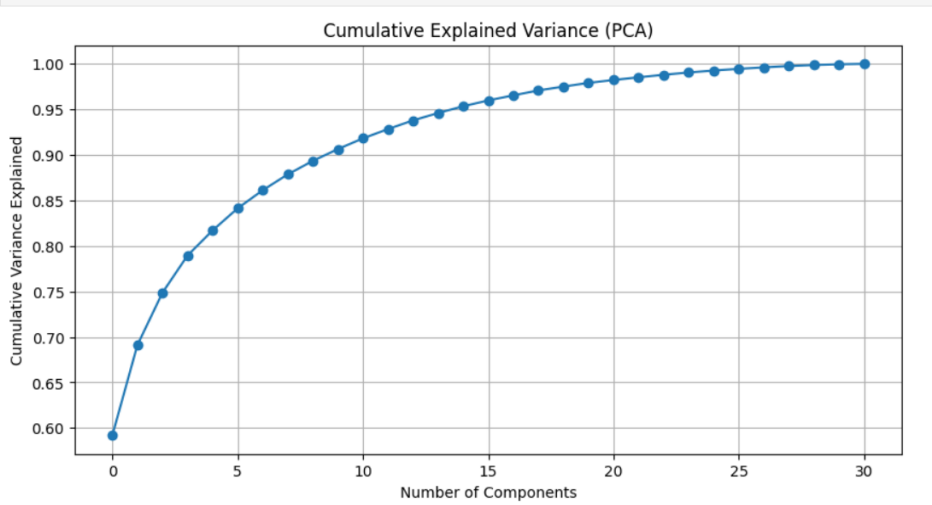


3. **Principal Component Analysis(PCA)**

According to QuantNext, The Principal Component Analysis or PCA is a statistical technique for reducing the dimension of a large dataset. van der Maaten et al. (2009) compare PCA with 12 front-ranked nonlinear dimensionality reduction techniques by applying each on self-created and natural tasks. CA aims to construct a low-dimensional representation of the data while maintaining the maximal variance and covariance structure of the data (Jolliffe, 1986).

PCA was used to uncover the primary causes of portfolio variance and minimize dimensionality in the standardized return matrix.

The first principal component (PC1) indicated the presence of a wide market/sector, which explained the majority of the variance.
More peculiar behaviors or subgroup trends, such as the distinctions between residential and industrial REITs, were represented by later components.
The scree plot revealed that the behaviour of the portfolio was primarily explained by the top three to four components, indicating that the REITs in XLRE are highly correlated and motivated by similar risk considerations.

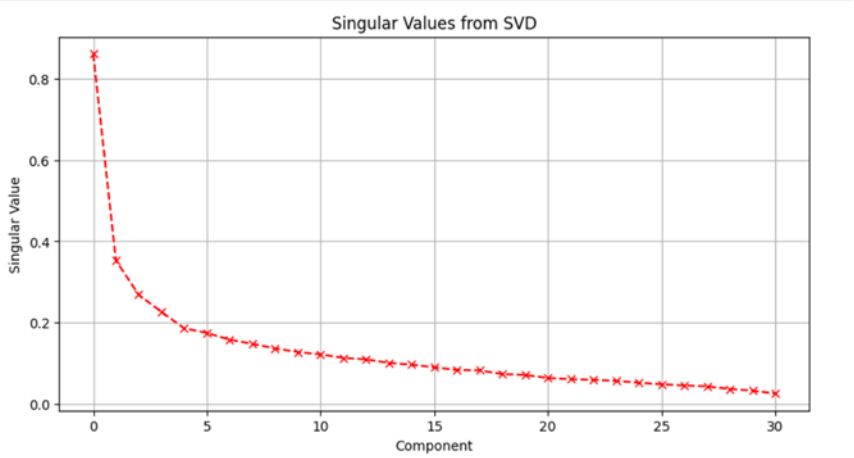


4. **Singular Value Decomposition (SVD)**

SVD further confirmed these results. By decomposing the return matrix, we observed:

- **A sharp drop-off in singular values**, emphasizing that only a few directions (factors) in the return space carry substantial informational weight.
- The **alignment with PCA** reinforced the robustness of our insights.

This dimensionality reduction suggests strong opportunities for **factor-based portfolio construction** or risk forecasting.



5. **Interpretation of Key Mathematical Concepts**

   a. **Eigenvectors (PCA) -** Determine which data directions best account for variance; in other words, the "hidden drivers" of the market.
   b. **Eigenvalues (PCA) -** Show the amount of variance that each component can account for.
   c. **Singular Values (SVD) -** Represent the strength of each factor, similar to eigenvalues in PCA, but from a decomposition standpoint.

Recommended Course of Action

Based on the analysis of daily returns, PCA, and SVD of the 30 largest holdings in XLRE, several strategic recommendations emerge for portfolio management and future analysis:

i. **Focus on Principal Drivers**

A limited number of components explain most of the variance in returns, according to the PCA and SVD results. This implies that a small number of important macroeconomic or industry-specific factors—such as interest rates, inflation forecasts, and demand for commercial real estate—drive a significant portion of the movement in XLRE constituents.
Action: Finding and keeping an eye on these latent elements rather than following every single stock helps simplify portfolio exposure. For hedging or risk management, concentrate on the main components that are dominating.

ii. Selective Diversification

A small number of stocks may contribute to distinctive variation, indicating idiosyncratic behavior, even while PCA demonstrates correlation across several holdings. These may offer advantages for diversification.

Action: To maximize diversification while remaining within the sector, find and think about adding stocks that are underrepresented or have weak correlations.

iii. Risk Timing Using Volatility

A few securities may contribute to distinctive variance, indicating idiosyncratic behavior, even while PCA demonstrates correlation across multiple holdings. These can offer advantages for diversity. Take Action: To maximize diversification while remaining within the sector, find and think about adding stocks that are underrepresented or have weak correlations.

iv. Dynamic Allocation Based on Rolling Averages

Rolling returns can help with tactical allocation by exposing short-term patterns.

Action: Put momentum-based weighting into practice by underweighting holdings with falling trends and overweighting those with consistent positive rolling returns.

v. Factor-Based ETF Alternatives

It may be possible to approximate exposure to XLRE's risk profile with fewer ETFs or factor-based instruments because a small number of components can explain a large portion of the return structure.

Action: Explore REIT-focused ETFs with factor tilts (e.g., quality, momentum) to reduce complexity while maintaining exposure.

vi. Enhance Monitoring Systems

This analysis shows the importance of real-time statistical tools (PCA, rolling volatility, etc.) in navigating sector dynamics. Action: Develop or adopt dashboard tools to track PCA loadings, rolling volatilities, and cross-correlations in real time.

[i]References

1. [i ii] . Jolliffe, I. T., and I. T. Jolliffe. "Generalizations and adaptations of principal component analysis." *Principal component analysis* (1986): 223-234.
2. Van Der Maaten, Laurens, Eric Postma, and Jaap Van den Herik. "Dimensionality reduction: a comparative." *J Mach Learn Res* 10.66-71 (2009).