2025-2 IMEN891M – Financial BigData Analysis

Final Presentation

오승준, 최우혁, Isac Johnsson

2025.10.22.

0 Index

1 Introduction (Revisited)

Background

- High dimensional data analysis is essential, yet poses significant challenges in modern econometrics.
- Classic mean-variance **Markowitz portfolio theory** often fails into higher dimension of data due to unstable covariance estimation, omitted network dependencies and the curse of dimensionality.
- To address these challenges, various methodologies has before been proposed, a few important ones are:

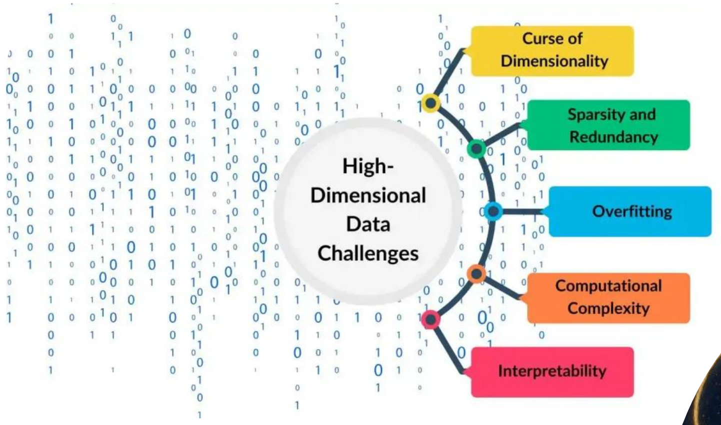


Image Source: Van Ottten, Neri; Spotintelligence.com

POET - High dimensional covariance estimation (Fan, Liao & Mincheva, 2013)

- By applying **factor models** and **sparsity** in the residual covariance matrix, it efficiently estimates **high-dimensional covariance matrices**.

SAR - Spatial Autoregression (Cliff & Ord, 1981; Baltegi et al., 2014)

- Model that **captures dependencies** structures across assets and markets.
- From here, we take some ideas with subgrouping.

LASSO - Regularized regression (Tibshirani, 1996)

- Imposes **sparsity on portfolio weights**, which improves stability in portfolio optimization in high dimensions.

Image Source: OpenAI

Therefore, we integrate these High-Dimensional Methods for Robust Portfolio Construction.

2 Data

POSTECH

DATA SELECTION

- We employ a cross-asset dataset (2015.01.01 - 2024.12.31) spanning totally **71** series. (Data Source : investing.com , Yahoo Finance, etc.)

Equities

- Global indices and stocks (From S&P 500 and it's sectoral indices, MSCI World, Major regional stocks and indexes like Nikkei, KOSPI etc…)

Fixed Income

- U.S. Treasury yields (3M, 5Y, 10Y, 30Y, corporate bond spreads)

Commodities

- Gold, Oil(Brent, WTI), industrial metals, agricultural futures, etc..

Currencies

- USDKRW, JPYKRW, EURKRW, CNYKRW

Others

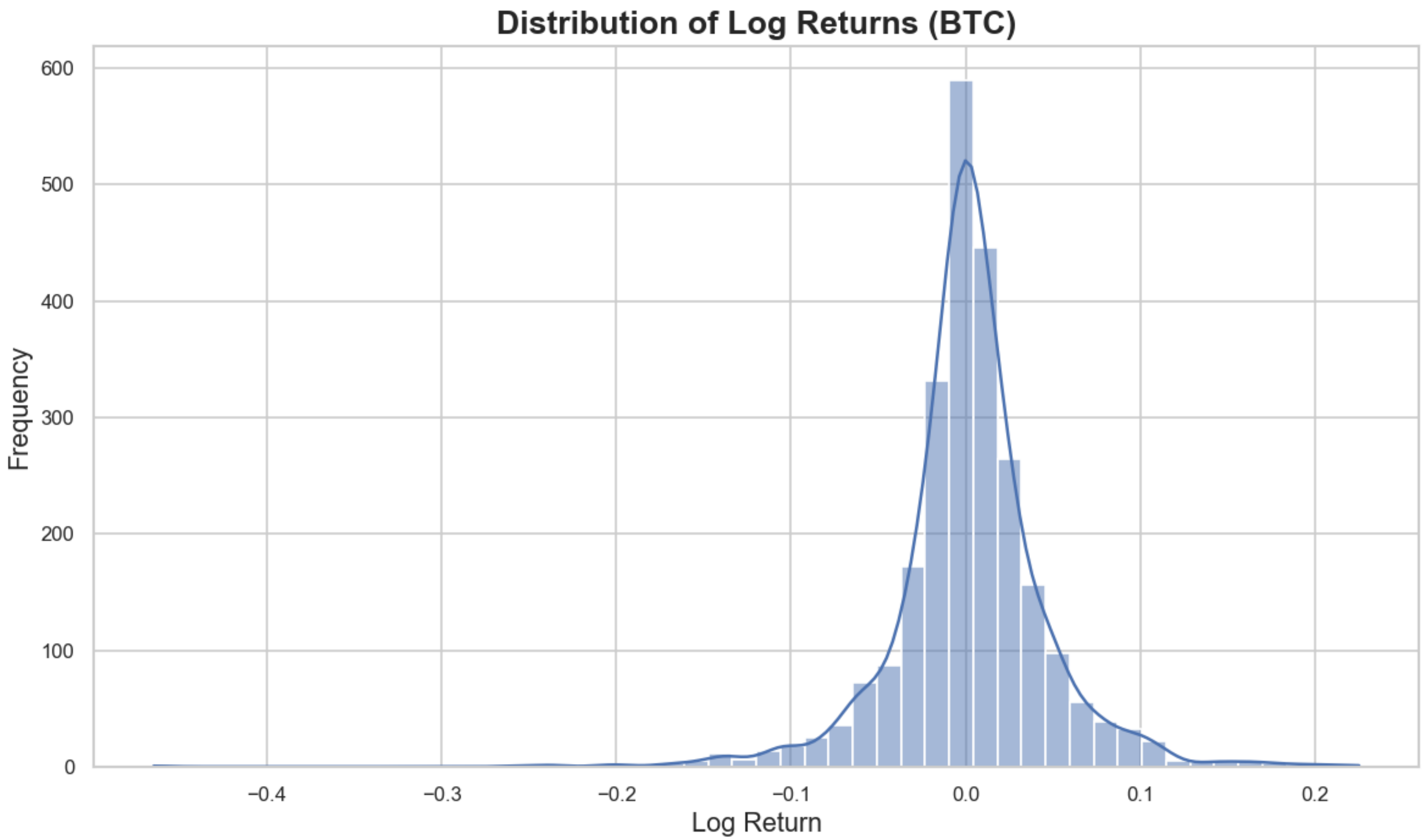
- Bitcoin (Cryptocurrencies)
- VIX(Volatility Measures)
- Macro indicator proxies (ex. CPI, Dollar Index, etc..)

POSTECH

Explanatory Data Analysis

Distribution of Log Returns

- As a representative of all assets, **Bitcoin**, the asset with the highest average return and volatility, has been selected.
- The distribution of BTC log returns is sharply **peaked around zero**, indicating most daily changes are small.
- The **heavy tails** on both sides show the presence of large price swings compared to a normal distribution.
- The curve is **slightly asymmetric**, suggesting mild skewness in return behavior.
- This pattern reflects Bitcoin's **high volatility and non-Gaussian nature**, common in crypto markets.

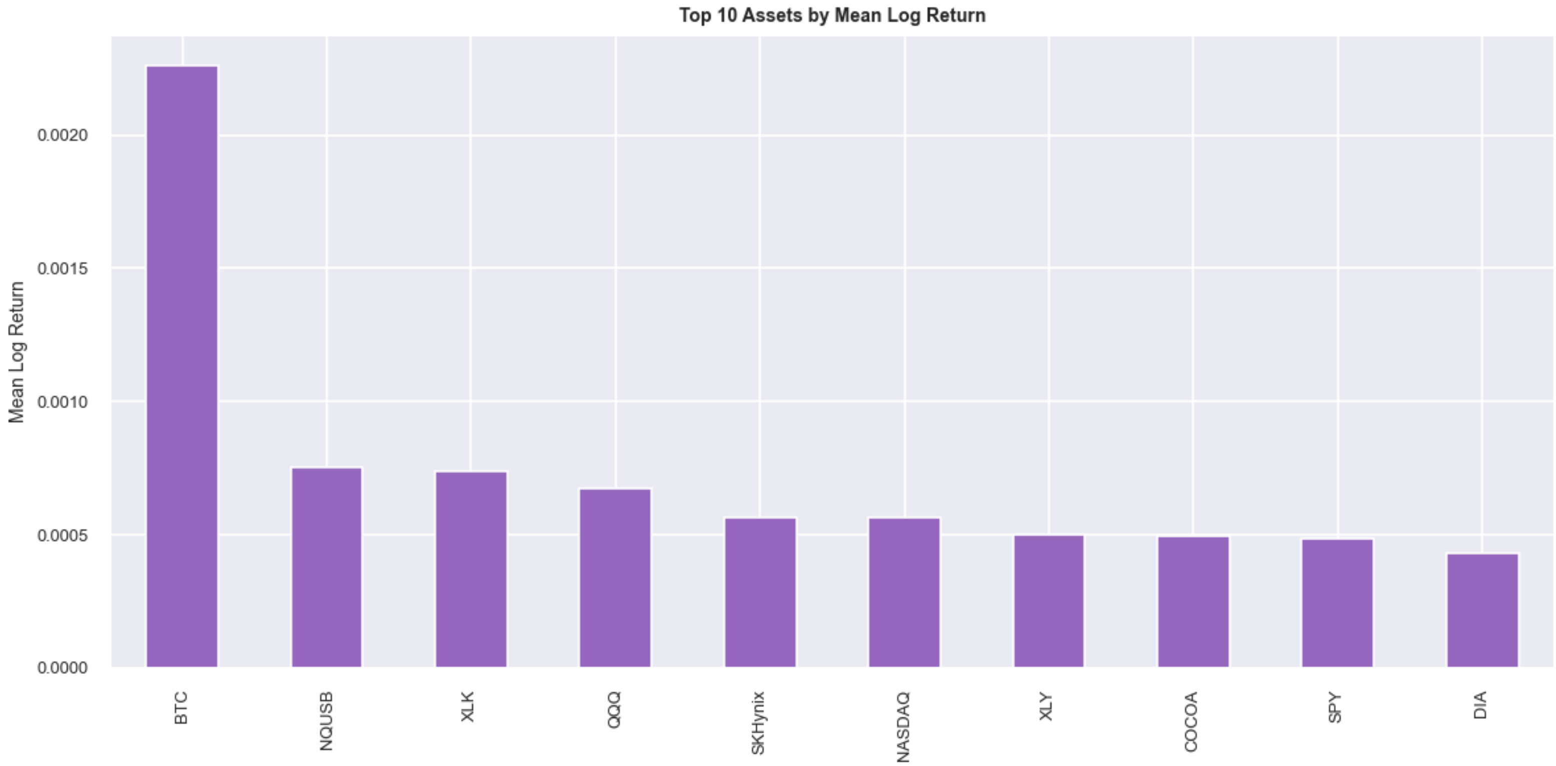


POSTECH

Explanatory Data Analysis

Mean Log Return

- **BTC shows the highest mean log return**, significantly outperforming all other assets, reflecting its high volatility and long-term upward trend.
- **Tech-related assets** such as NQUSB, XLK, and QQQ follow, indicating strong performance from the digital and technology sectors.
- **Traditional indices (SPY, DIA)** and **commodities (COCOA)** exhibit lower mean returns, consistent with their relatively stable nature.
- Overall, the pattern highlights a **clear risk–return trade-off**, where higher-risk assets yield higher average log returns.

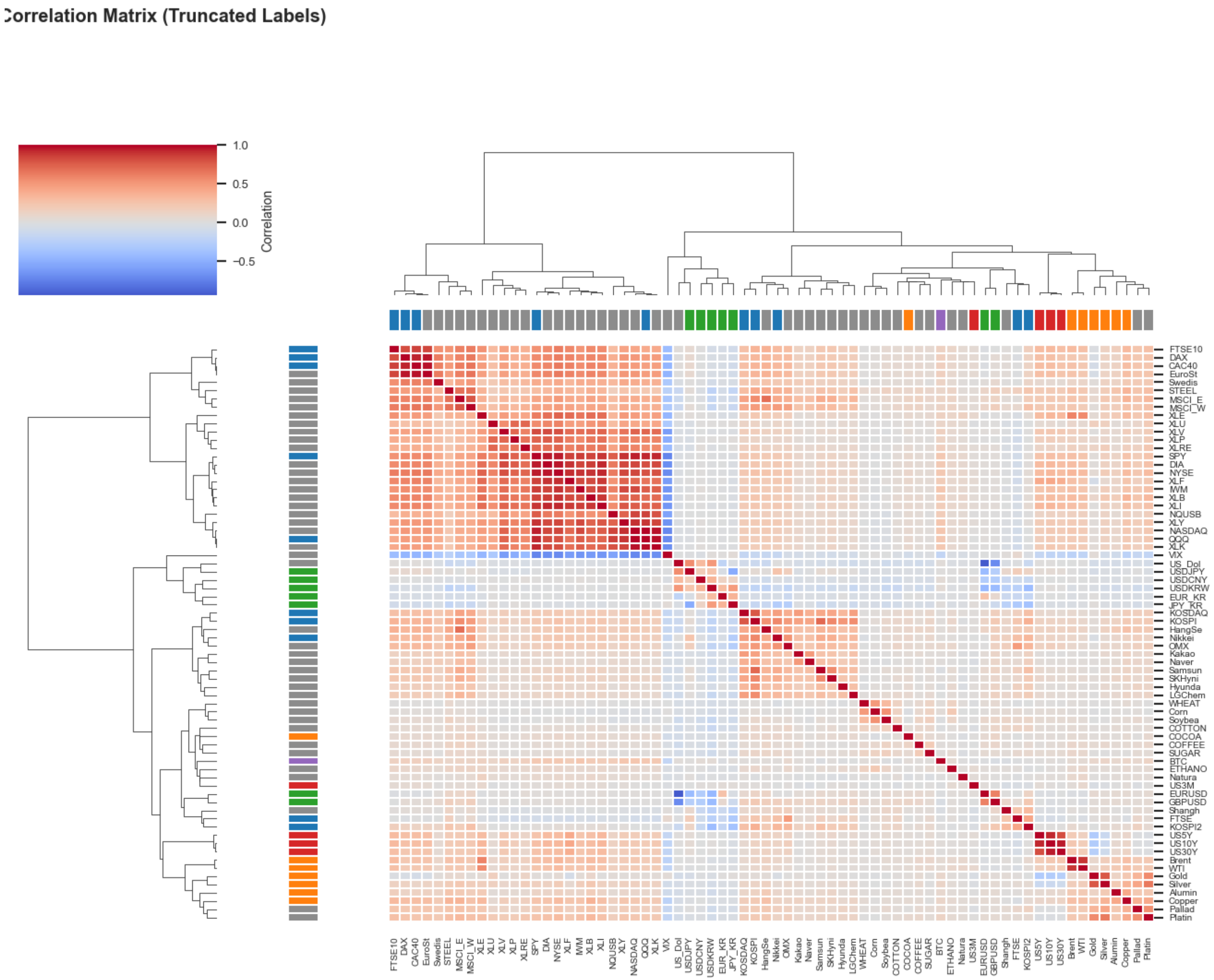


2 Data

Explanatory Data Analysis

Correlation Matrix

- The **correlation heatmap with hierarchical clustering** reveals distinct asset groupings based on return co-movements.
- **Equity indices and ETFs** (e.g., SPY, QQQ, NASDAQ, NQUSB) form a tight cluster with **strong positive correlations**, shown in deep red.
- **Commodities and currencies** exhibit **weaker or negative correlations**, suggesting diversification benefits across asset classes.
- The **blue patches** indicate negatively correlated pairs, mainly between **volatility indices (e.g., VIX)** and **risk-on assets**, consistent with market stress dynamics.

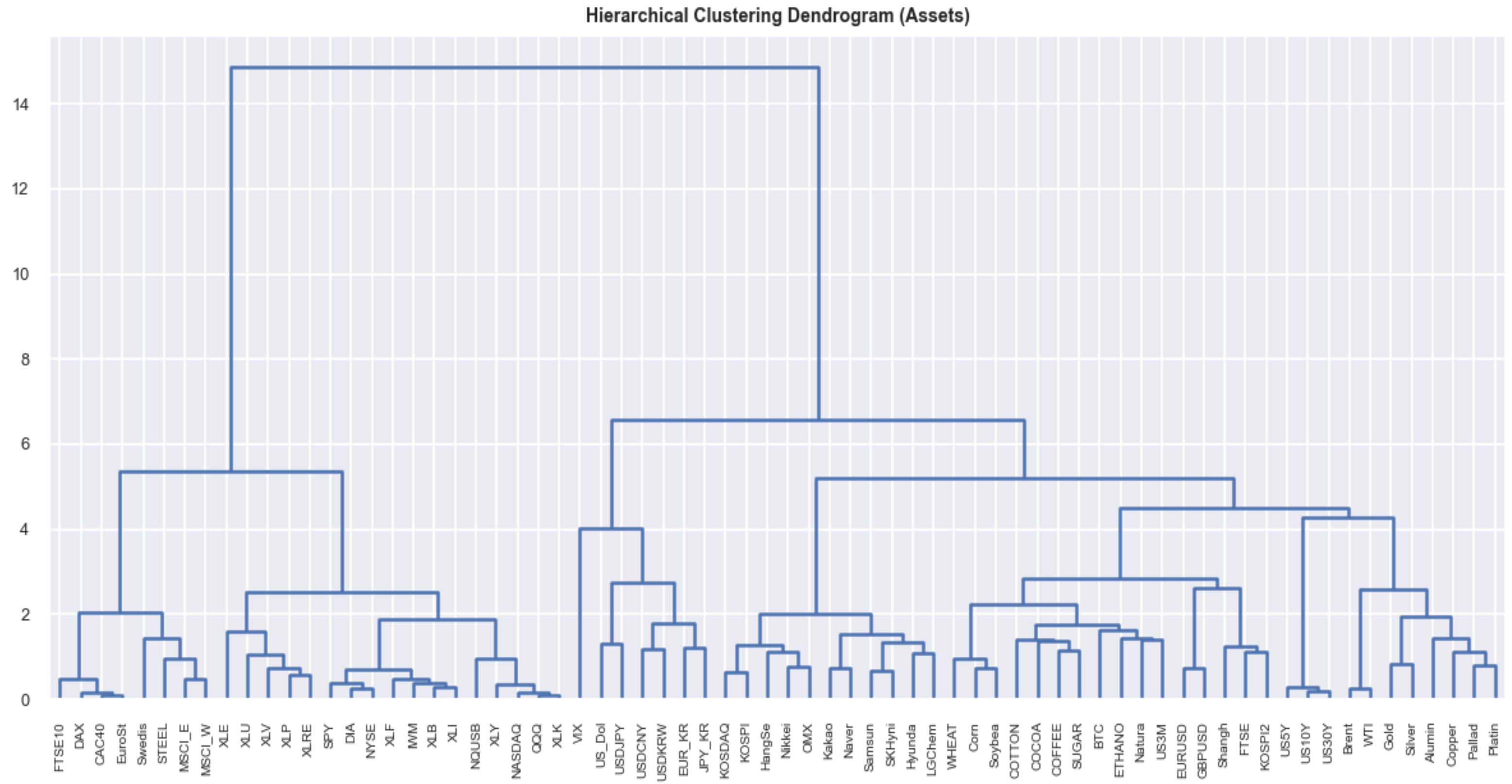


2 Data

Explanatory Data Analysis

Dendrogram Analyis

- The **hierarchical clustering dendrogram** visualizes the structural relationships among assets, grouping those with similar return dynamics.
- Unlike the baseline EDA clustering, our **integrated model** captures **clearer and more economically consistent clusters**, separating equities, commodities, and currencies more distinctly.
- This suggests that the integrated framework enhances **cross-asset structure recognition**, aligning with intuitive market linkages.
- Overall, it demonstrates **improved cluster interpretability and coherence** compared to standard correlation-based grouping.

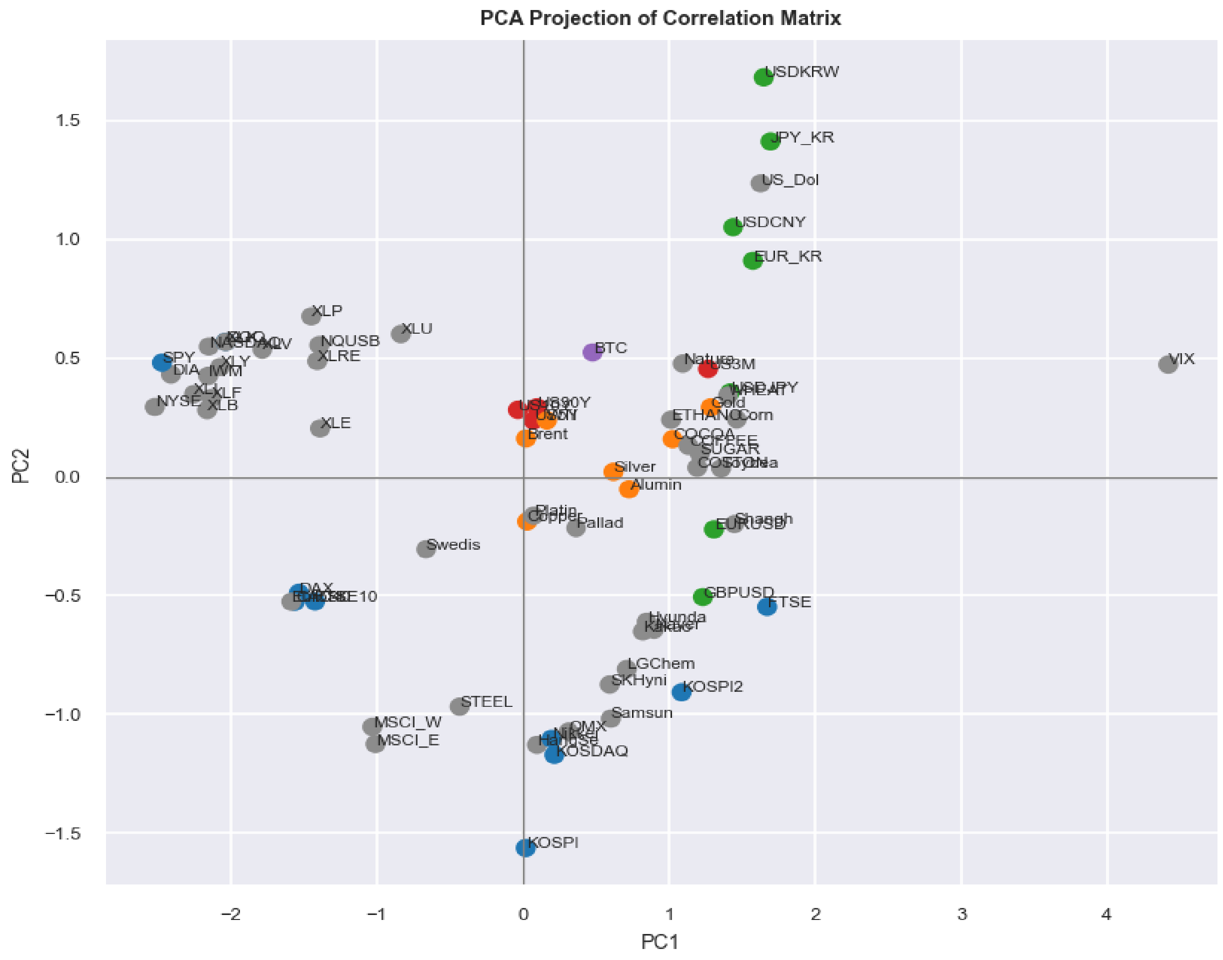


2 Data

Explanatory Data Analysis

PCA Projection Analyis

- Even with four principal components, PCA explains only **43% of total variance**, revealing limited ability to capture complex market dynamics.
- This suggests that a large portion of asset-specific or nonlinear variation remains unexplained.
- Our **integrated model** combines multiple covariance estimators to better capture both systematic and idiosyncratic risk factors.
- We expect it to explain a **substantially larger share of market variance** compared to PCA.

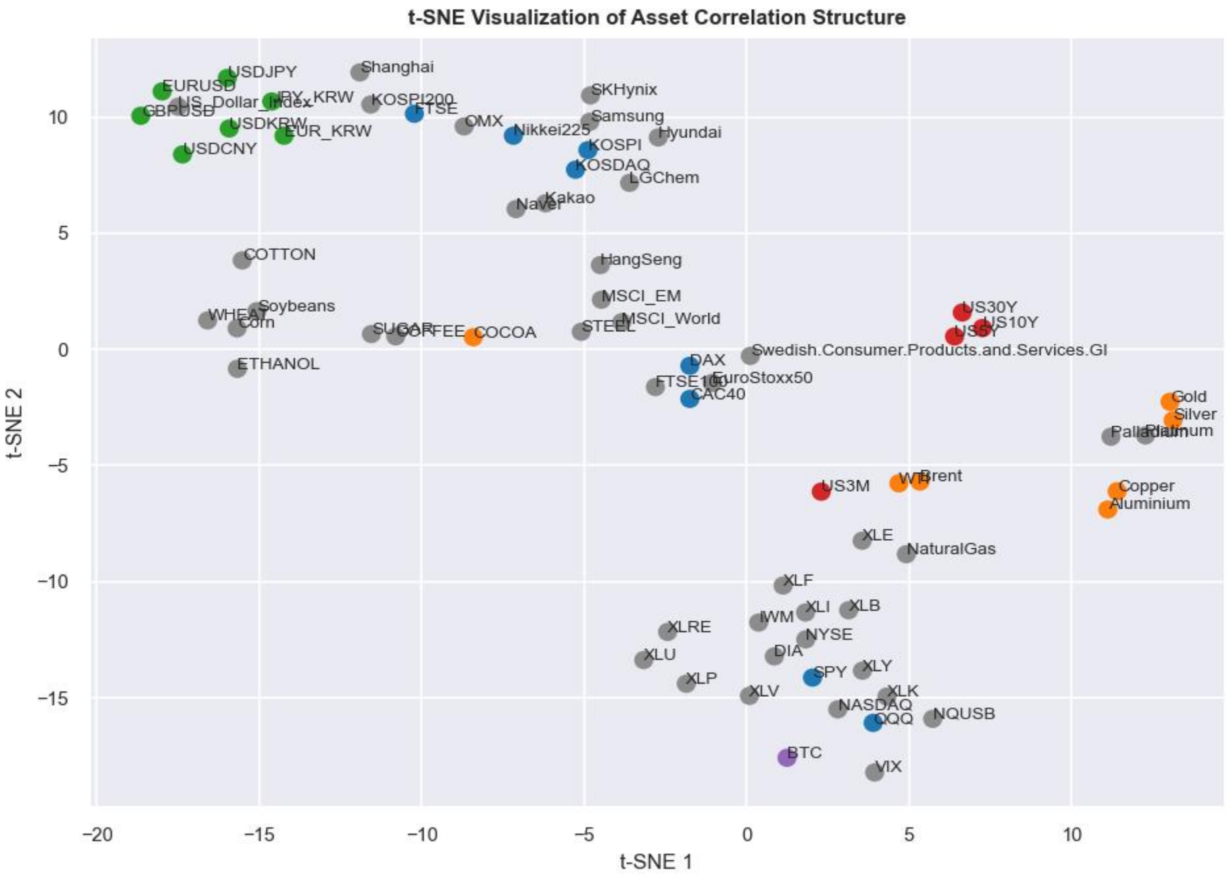


2 Data

Explanatory Data Analysis

t-SNE Visualization

- The **t-SNE visualization** maps high-dimensional correlations into a 2D space, revealing clear clusters across asset classes.
- **Equities (e.g., QQQ, XLK, NASDAQ)** form a distinct group, while **commodities (e.g., Gold, Copper, Aluminium)** and **FX pairs (e.g., USDKRW, EURUSD)** occupy separate regions.
- **BTC and VIX** are positioned far from other assets, highlighting their **unique, uncorrelated behavior**.
- Compared to raw data clustering, the **integrated model produces a more coherent and interpretable structure**, capturing both global and local dependencies among assets.



3 Methodological Framework

POSTECH

Step 0. Notation & Setup

- Let the time index be $t = 1, \dots, T$, the number of assets be N, and the rolling window length be N, and the rolling window length be W.
- The return matrix is denoted by $R \in \mathbb{R}^{T \times N}$, where each element $R_{t,i}$ represents the log return of asset i at time t.
- The **test period** is defined as $\mathcal{T}_{test} \subset \{1, \dots, T\}$, and the forecasting horizon is $h \in \mathbb{N}$ $(\text{in code: } h = 2)$.
- The **gross exposure constraint** is controlled by $G \in [1, \infty)$, where $G = 1$ corresponds to a long-only portfolio and $G > 1$ allows long-short positions.
- The **composite macro factor** y_t is constructed as the average of three major U.S. market indices: $y_t = \frac{1}{3}(R_{t,SPY} + R_{t,NASDAQ} + R_{t,DIA})$, representing the aggregate market-wide movement
- The **rolling window** used for estimation at time t is defined as : $\mathcal{W}_t = \{t - W, \dots, t - 1\}$, which corresponds to the past W trading days immediately preceding t.

Step 1 : LASSO-Based Asset Selection

- To identify assets that move consistently with the overall market factor and reduce dimensionality and remove noisy, uninformative assets.
- Regress the composite macro factor y_t on asset returns X_t:

$$\min_{\beta} \frac{1}{W} \sum_{s \in \mathcal{W}_t} (y_s - X_s^T \beta)^2 + \lambda ||\beta||_1$$

- Rolling window of $W = 250$ days.
- λ determined by cross-validation (LassoCV).
- Each day's regression uses the most-recent window → adaptive over time.

LASSO - Regularized regression (Tibshirani, 1996)

- To avoid overfitting and spurious correlations, we apply LASSO regression : $\hat{\beta} = \arg\min_{\beta} ||y - X\beta||_2^2 + \lambda ||\beta||_1$ where y is the asset return or factor proxy, and X is the predictor matrix.
- LASSO selects a sparse subject of variables, forming the observed macro-finance block f_t^{Macro}.

Methodological Framework

Step 2. Hierarchical Clustering of Selected Assets

- To group the selected assets $\mathcal{A}_t$ into structurally similar clusters.
- To capture sector, style, or co-movement patterns in return-behavior.

1) Correlation Matrix

- Using the selected assets' returns within the same rolling window:

$$\rho_{ij} = \mathrm{Corr}(R_i, R_j)$$

- Convert to a distance measure (Highly correlated assets have similar dist.):

$$D_{ij} = 1 - \rho_{ij}$$

2) Hierarchical Linkage

- Apply Ward linkage, which merges clusters to minimize within-cluster variance.
- Iteratively combine the most similar pairs until all assets from a hierarchy.
- Determine the final number of clusters using a distance threshold.

3) Cluster Assignment

- Each assets receives a cluster label : $c_t(i) \in \{1,2,\ldots,K_t\}$ where K_t = # of clusters at time t.

Step 3 : Factor Decomposition by PCA (POET Framework)

- To decompose asset returns into **systematic (factor)** and **idiosyncratic** components.
- Form the foundation for **POET (Principal Orthogonal complEment Thresholding)** covariance estimation.

1) PCA Decomposition

- On the selected asset returns $R^{(t)} \in \mathbb{R}^{W \times N_t}$: $R^{(t)} \approx F_t L_t^T$ where F_t denotes factor return matrix (common components), and L_t : factor loading matrix(exposures of each asset).

2) Covariance Components (Fan, Liao & Mincheva, 2013)

- Compute two covariance parts (Factor, Idiosyncratic Part) :

$$\Sigma_t^{factor} = L_t \mathrm{Cov}(F_t) L_t^T , \qquad \Sigma_t^{id,raw} = Cov(R^{(t)} - F_t L_t^T)$$

3) POET Integration

- Combine both parts after applying thresholding on the idiosyncratic covariance. (Next step).
- Produces a low-rank + sparse covariance estimator.

3 Methodological Framework

POSTECH

Step 4. Integrated Covariance Construction (POET–Based)

- To construct a stable covariance estimator combining **low-rank factor structure (POET)** and **block-sparse residuals and incorporate clustering information** to reflect market structure.

1) Cluster-based Hard Thresholding

- Build a binary mask using hierarchical clustering results:

$$M_{ij} = \begin{cases} 1, & c_t(i) = c_t(j) \\ 0, & \text{otherwise} \end{cases}$$

- Apply thresholding to the idiosyncratic covariance:

$$\Sigma_t^{id,blk} = \Sigma_t^{id,raw} \odot M_t$$

- It retains correlations within the same cluster, sets cross-cluster elements to zero.

2) Integrated Covariance Estimator

$$\hat{\Sigma}_t^{integrated} = \Sigma_t^{factor} + \Sigma_t^{id,blk}$$

- Low-rank factor component from POET
- Block-sparse residual component from clustering

Step 5 : Benchmark Covariance Models

- Construct baseline covariance estimators for comparison against the proposed integrated model.

A) LASSO-only Model

$$\Sigma_t^{LASSO} = Cov(R_t^{(\mathcal{A}_t)})$$

- Uses only asset selected by LASSO (no factor or clustering).
- Simple Empirical covariance within the active asset set.
- Measures how well pure statistical selection performs.

B) POET-only Model

$$\Sigma_t^{POET} = L_t Cov(F_t) L_t^T + diag\left(Cov\left(R^{(t)} - F_t L_t^T\right)\right)$$

- Implements the **standard POET framework** (Fan et al., 2013).
- Uses low-rank factor structure + diagonal residuals.
- Captures global systematic risk, but ignores cross-asset structure.

C) OLS(Shrinkage) Model

$$\Sigma_t^{Shrinkage} = (1-\alpha)\Sigma_t^{sample} + \alpha\Sigma_t^{target}$$

- Estimated using **Ledoit–Wolf shrinkage** method.
- Shrinks noisy sample covariance toward a well-conditioned target.

POSTECH

Step 6. Portfolio Optimization & Performance Evaluation

- Evaluate each covariance estimator (Integrated & Benchmarks) through **Global Minimum Variance (GMV)** portfolio backtesting.

1) Global Minimum Variance (GMV) Optimization Problem

- For each covariance matrix Σ_t, find portfolio weights w_t minimizing portfolio variance: $\min_w w^T \Sigma_t w$ subject to $\sum_i w_i = 1, \sum_i |w_i| = G$

- G : Gross exposure constraint (1.0, 1.25, …, 3.0)
- w_i : Portfolio weight of asset i

- Solved via SLSQP (Sequential Least Squares Quadratic Programming)

2) Backtesting Framework

- Rolling-window optimization over years **2022-2024.**

- Forecast horizon : $h = 2$ days ahead

- Compute realized returns using future data:

$$r_{p,t+1:t+h} = w_t^T R_{t+1:t+h}$$

3) Metrics

Metric	Definition	Interpretation
Annualized Return (%)	$\bar{r_p} \times 252$	Profitability
Annualized Risk (%)	$\sigma_p \times \sqrt{252}$	Volatility
Sharpe Ratio	$\frac{\bar{r_p}}{\sigma_p}\sqrt{252}$	Risk-adjusted return
Frobenius Loss	$\frac{\left\|\Sigma_t^{real} - \Sigma_t^{forecast}\right\|_F}{N}$	Covariance prediction error
KL Divergence	$\frac{1}{2}\left[\mathrm{tr}\left(\left(\Sigma^{forecast}\right)^{-1}\Sigma^{real}\right) - n + \log\frac{\lvert\Sigma^{forecast}\rvert}{\lvert\Sigma^{real}\rvert}\right]$	Distributional gap
Risk Gap	$\lvert\sigma_{ex-post} - \sigma_{ex-ante}\rvert/\sigma_{ex-post}$	Accuracy of Risk Prediction

3 Methodological Framework

POSTECH

Step 4. Integrated Covariance Construction (POET–Based)

- To construct a stable covariance estimator combining **low-rank factor structure (POET)** and **block-sparse residuals and incorporate clustering information** to reflect market structure.

1) Cluster-based Hard Thresholding

- Build a binary mask using hierarchical clustering results:

$$M_{ij} = \begin{cases} 1, & c_t(i) = c_t(j) \\ 0, & \text{otherwise} \end{cases}$$

- Apply thresholding to the idiosyncratic covariance:

$$\Sigma_t^{id,blk} = \Sigma_t^{id,raw} \odot M_t$$

- It retains correlations within the same cluster, sets cross-cluster elements to zero.

2) Integrated Covariance Estimator

$$\hat{\Sigma}_t^{integrated} = \Sigma_t^{factor} + \Sigma_t^{id,blk}$$

- Low-rank factor component from POET
- Block-sparse residual component from clustering

Step 5 : Benchmark Covariance Models

- Construct baseline covariance estimators for comparison against the proposed integrated model.

A) LASSO-only Model

$$\Sigma_t^{LASSO} = Cov(R_t^{(\mathcal{A}_t)})$$

- Uses only asset selected by LASSO (no factor or clustering).
- Simple Empirical covariance within the active asset set.
- Measures how well pure statistical selection performs.

B) POET-only Model

$$\Sigma_t^{POET} = L_t Cov(F_t) L_t^T + diag\left(Cov\left(R^{(t)} - F_t L_t^T\right)\right)$$

- Implements the **standard POET framework** (Fan et al., 2013).
- Uses low-rank factor structure + diagonal residuals.
- Captures global systematic risk, but ignores cross-asset structure.

C) OLS(Shrinkage) Model

$$\Sigma_t^{Shrinkage} = (1-\alpha)\Sigma_t^{sample} + \alpha\Sigma_t^{target}$$

- Estimated using **Ledoit–Wolf shrinkage** method.
- Shrinks noisy sample covariance toward a well-conditioned target.

Methodological Framework

POSTECH

Summary of Methodology

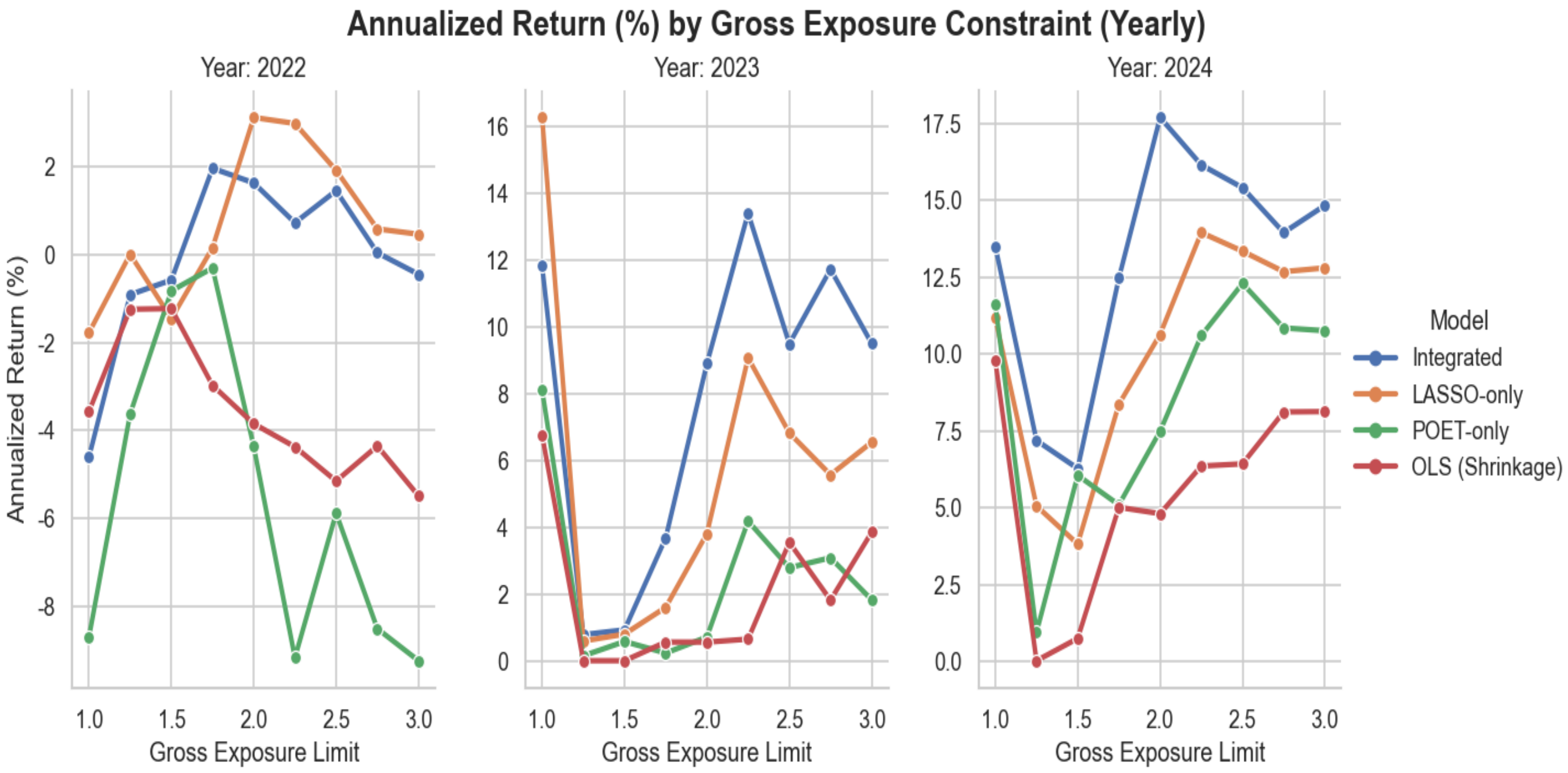
Step	Method	Mathematical Role	Expected Benefit
1	**LASSO**	Sparse regression	Select market-linked assets
2	**Clustering**	Structural grouping	Reflect sector/style patterns
3	**POET (PCA)**	Factor decomposition	Capture common systematic risk
4	**Hard-Thresholding**	Block sparsity	Preserve intra-cluster correlation
5	**GMV Optimization**	Quadratic programming	Stable portfolio weights
6	**Evaluation Metrics**	Frobenius / KL / Risk Gap	Validate predictive accuracy

4 Result Analysis

POSTECH

Annualized Return

- **Integrated model** shows the most consistent and robust performance across years, maintaining **positive and stable returns** even at high exposure levels.
- In **2022 (down market)**, it effectively **suppressed noise** and avoided overfitting through block-sparse covariance.
- During **2023–2024 (recovery and expansion)**, Integrated model achieved **higher responsiveness** and **efficient risk–return balance**, outperforming benchmarks.
- **LASSO-only** fluctuates under regime shifts, **POET-only** underfits, and **OLS (Shrinkage)** remains **overly conservative** with limited upside.



4 Result Analysis

POSTECH

Annualized Risk

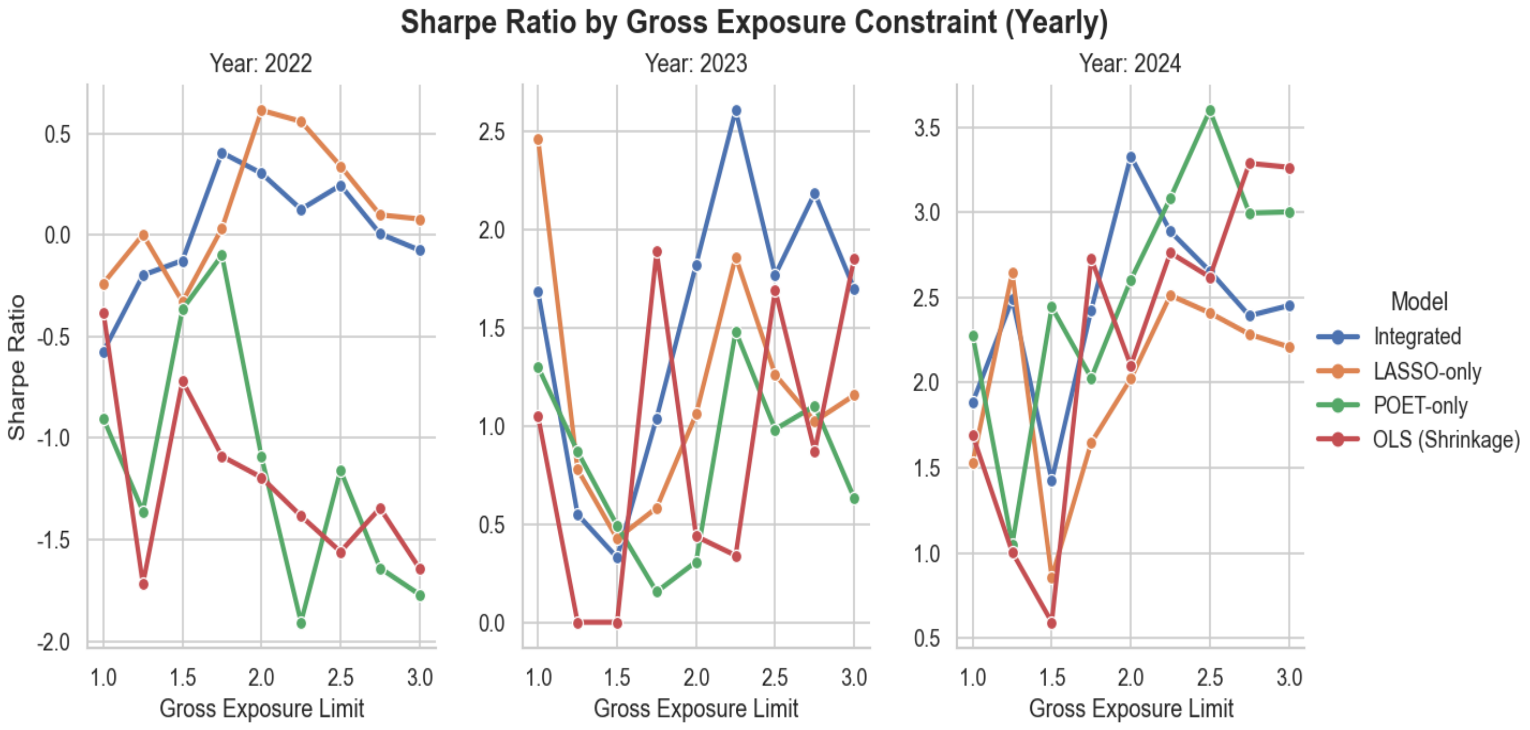
- Across all years, the **Integrated model** maintains a **moderate and controlled risk profile**, showing smooth increases with exposure and **no abrupt volatility spikes**.
- In low-exposure regimes (G ≤ 1.5), all models experience a **sharp risk drop**, but Integrated stabilizes faster than others, indicating **better covariance regularization**.
- **POET-only** and **OLS (Shrinkage)** display the **lowest absolute risk**, but at the cost of **under-exposure and limited returns**, implying over-conservatism.
- **LASSO-only** becomes unstable at higher exposures, while Integrated sustains **consistent risk scaling** aligned with expected leverage effects.

Result Analysis

POSTECH

Sharpe Ratio

- **Integrated model** consistently achieves **the highest or near-highest Sharpe ratios**, showing **balanced risk–return efficiency** across all years.
- In **2022 (volatile market)**, it maintains positive Sharpe while others fluctuate, reflecting **effective noise suppression and stable covariance estimation**.
- During **2023–2024**, Integrated and OLS models both improve sharply, but Integrated remains more **responsive to exposure scaling** and market recovery.
- **LASSO-only** delivers short-lived peaks, and **POET-only** exhibits delayed improvements, confirming that **hybrid integration yields superior risk-adjusted performance**.

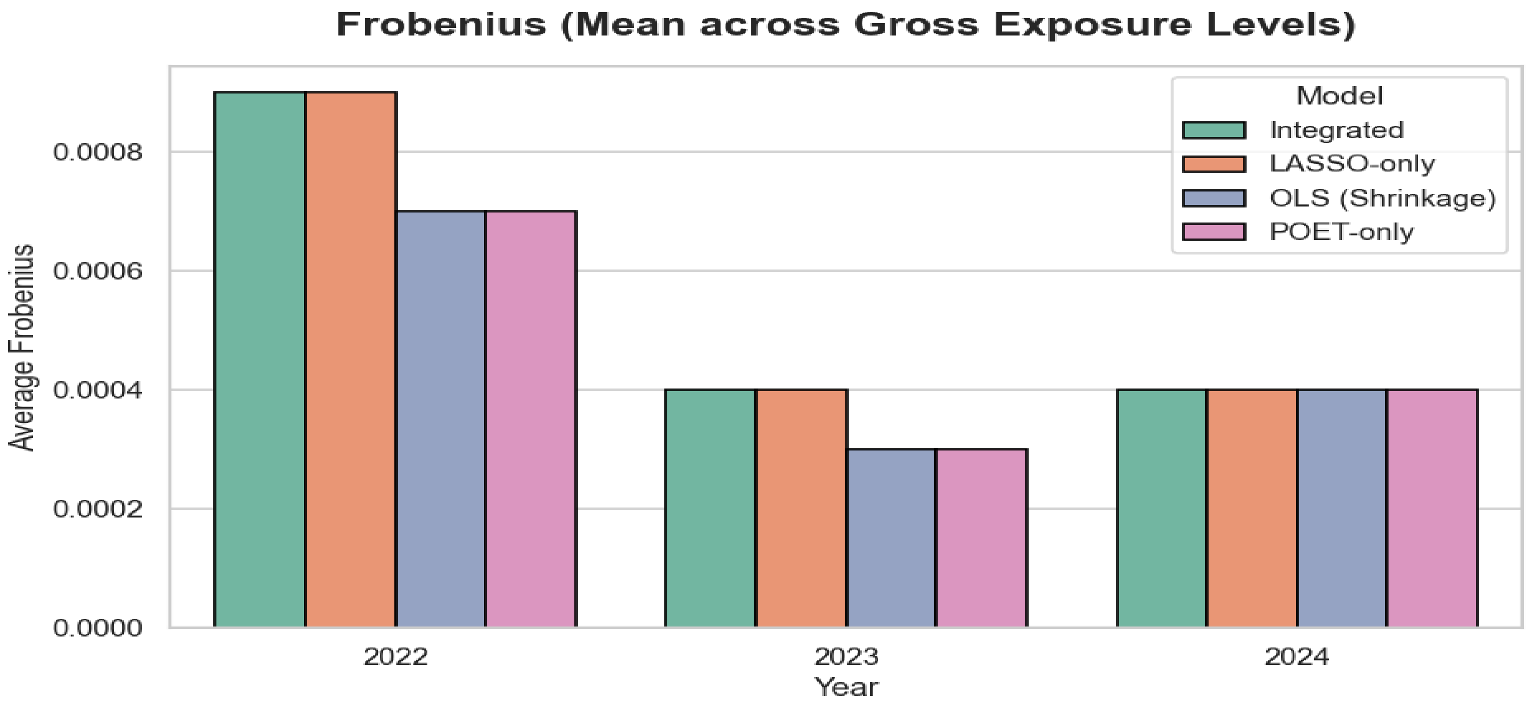


4 Result Analysis

POSTECH

Mean Frobenius Dist. (Cross-sectional accuracy)

- **Integrated model** and **LASSO-only** show **higher errors in 2022**, likely due to unstable covariance dynamics under market stress.
- From **2023 onward**, all models converge to **lower Frobenius norms**, indicating stabilization and improved estimation consistency.
- **OLS (Shrinkage)** and **POET-only** consistently maintain **the smallest deviations**, highlighting their **strong baseline stability** but limited adaptability.
- **Integrated model's error reduction over time** suggests that its **hybrid structure learns and regularizes better** across regimes.

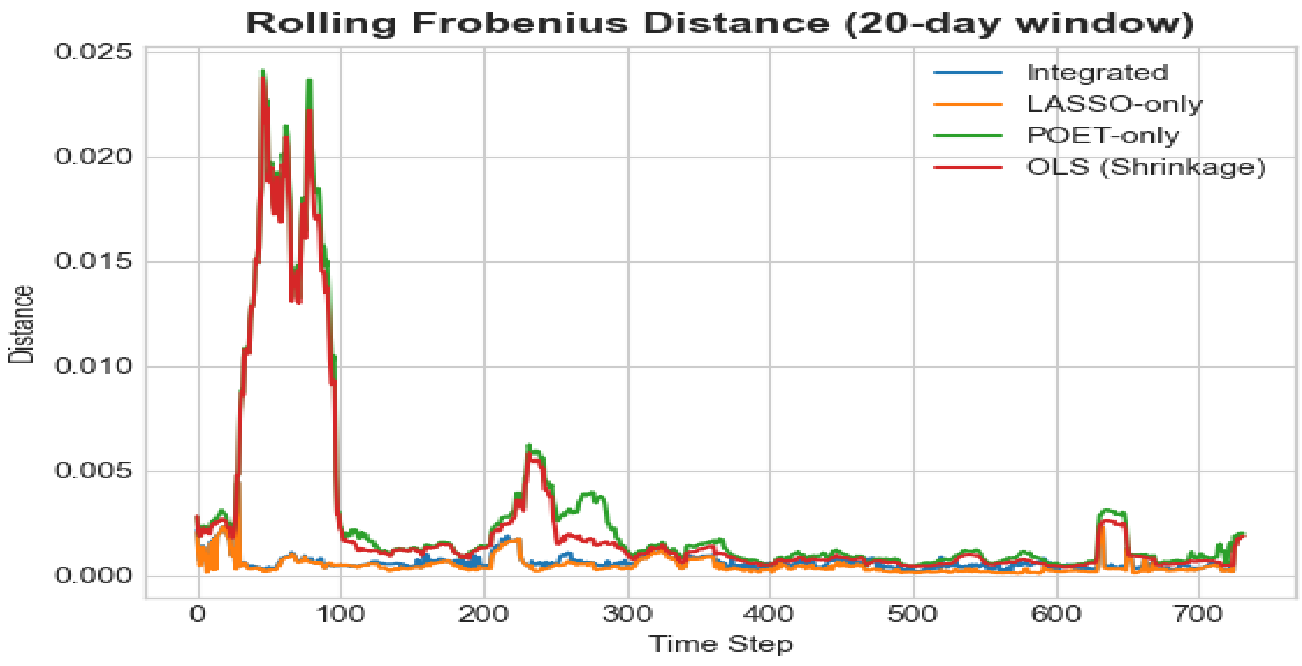


4 Result Analysis

POSTECH

Rolling Frobenius Dist. (Temporal Stability)

- Despite higher average Frobenius distance in annual means, the **Integrated model maintains the lowest rolling error** across time.
- It exhibits **remarkable temporal consistency**, showing minimal spikes even during early high-volatility periods.
- **OLS** and **POET-only** display large transient deviations, indicating sensitivity to regime shifts and local covariance shocks.
- The Integrated estimator thus demonstrates **superior stability** and **robust adaptation** under evolving market conditions.

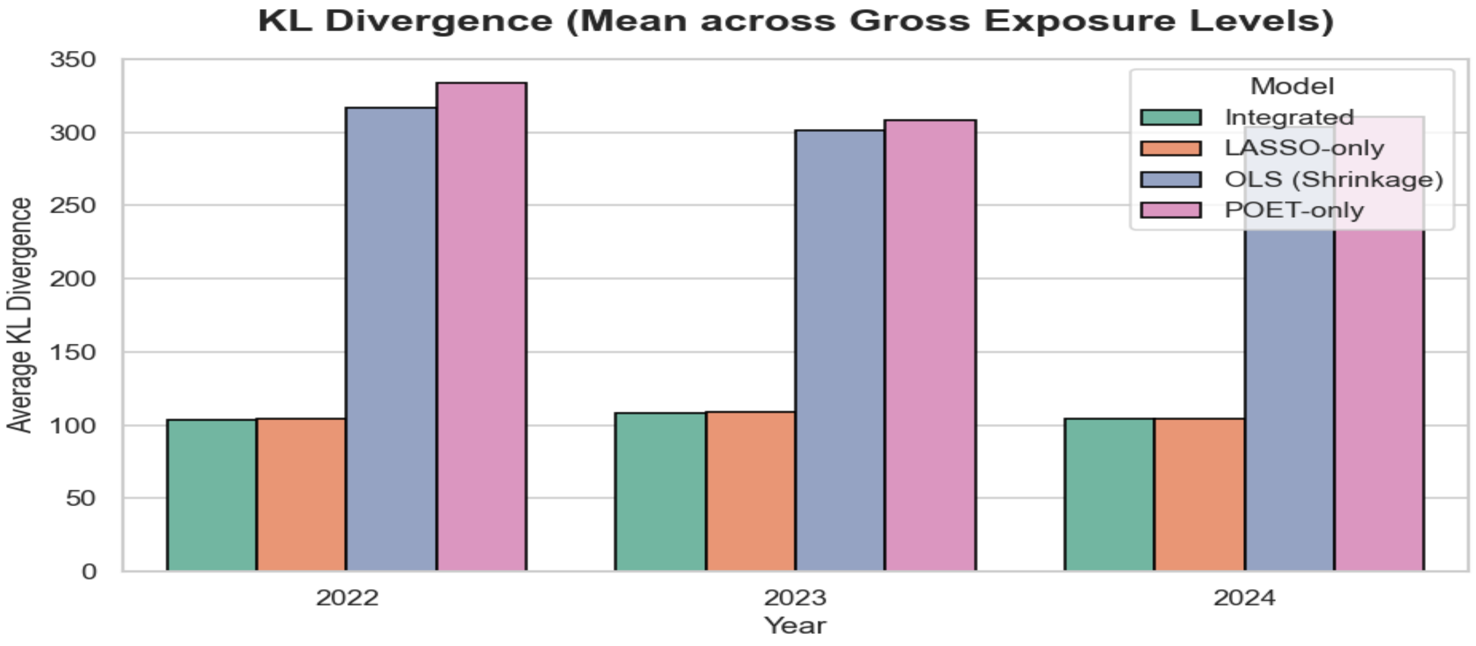


Result Analysis

POSTECH

KL Divergence

- **Integrated model** consistently records the **lowest KL divergence**,
  meaning it captures **true covariance structure with minimal information loss**.
- **OLS** and **POET-only** show extremely high divergences, reflecting **rigid shrinkage or over-simplified factor structures**.
- **LASSO-only** performs moderately well but lacks cross-cluster coherence, while the **Integrated model balances sparsity and dependency learning**.
- Overall, the Integrated estimator achieves **the most faithful approximation** of market risk distribution across all years.



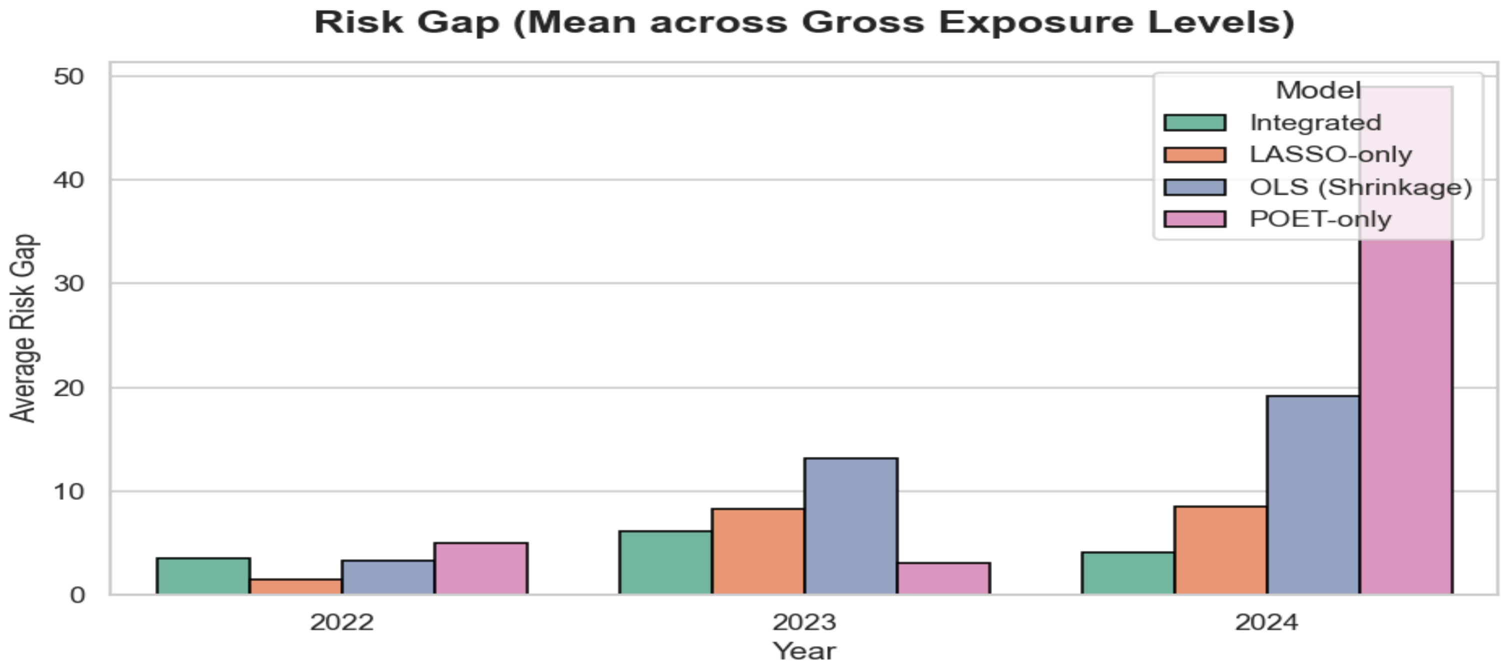
4

Result Analysis

POSTECH

Risk Gap

- **Integrated model** maintains **consistently low risk gaps** across all years, showing **strong alignment between predicted and realized volatility**.
- **OLS** and especially **POET-only** exhibit rapidly widening gaps in 2024, implying **systematic underestimation of true portfolio risk**.
- **LASSO-only** achieves moderate accuracy but fluctuates over time, lacking robustness under regime transitions.
- Overall, the **Integrated model provides the most reliable volatility forecasts**, balancing flexibility and structural stability in covariance updates.

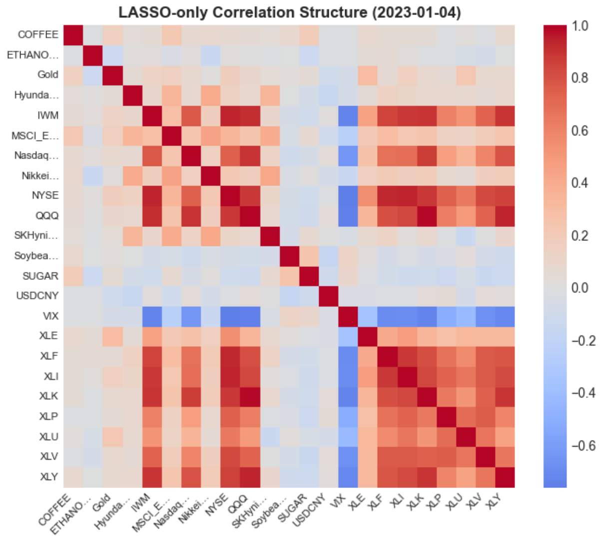


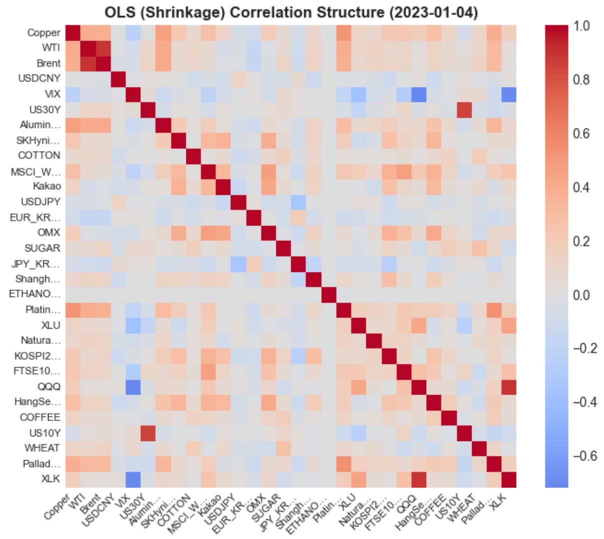
4 Result Analysis

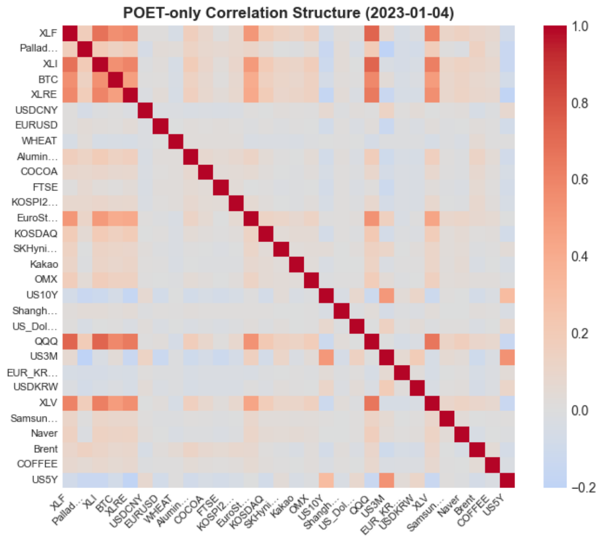
POSTECH

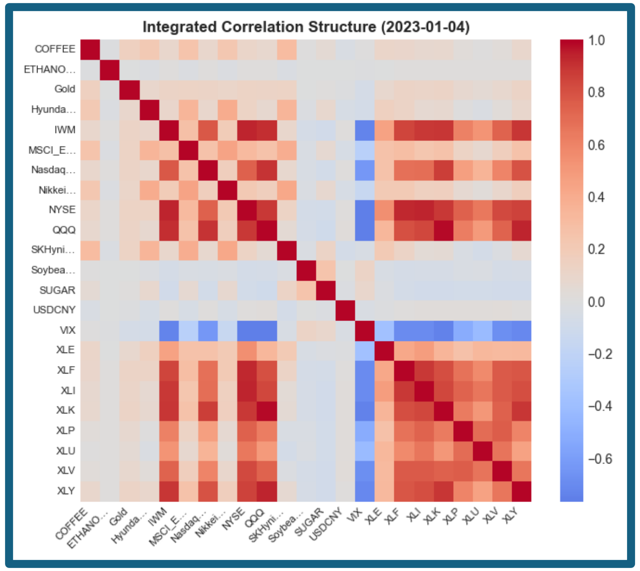
Correlation Structure Analysis

- **LASSO-only** captures **clear cluster blocks** (e.g., sector-wise correlations), but tends to **over-sparsify** and lose inter-cluster dependencies.
- **OLS** and **POET-only** produce **diffuse, unstructured correlations**, indicating **over-smoothed or noisy covariance patterns**.
- **Integrated model** preserves **the interpretable block structures** observed in LASSO while **suppressing spurious noise** — effectively balancing **sparsity and continuity** in correlation geometry.
- Visually, the Integrated estimator inherits **the structural clarity of LASSO** and **the global stability of shrinkage**, achieving the best of both approaches.







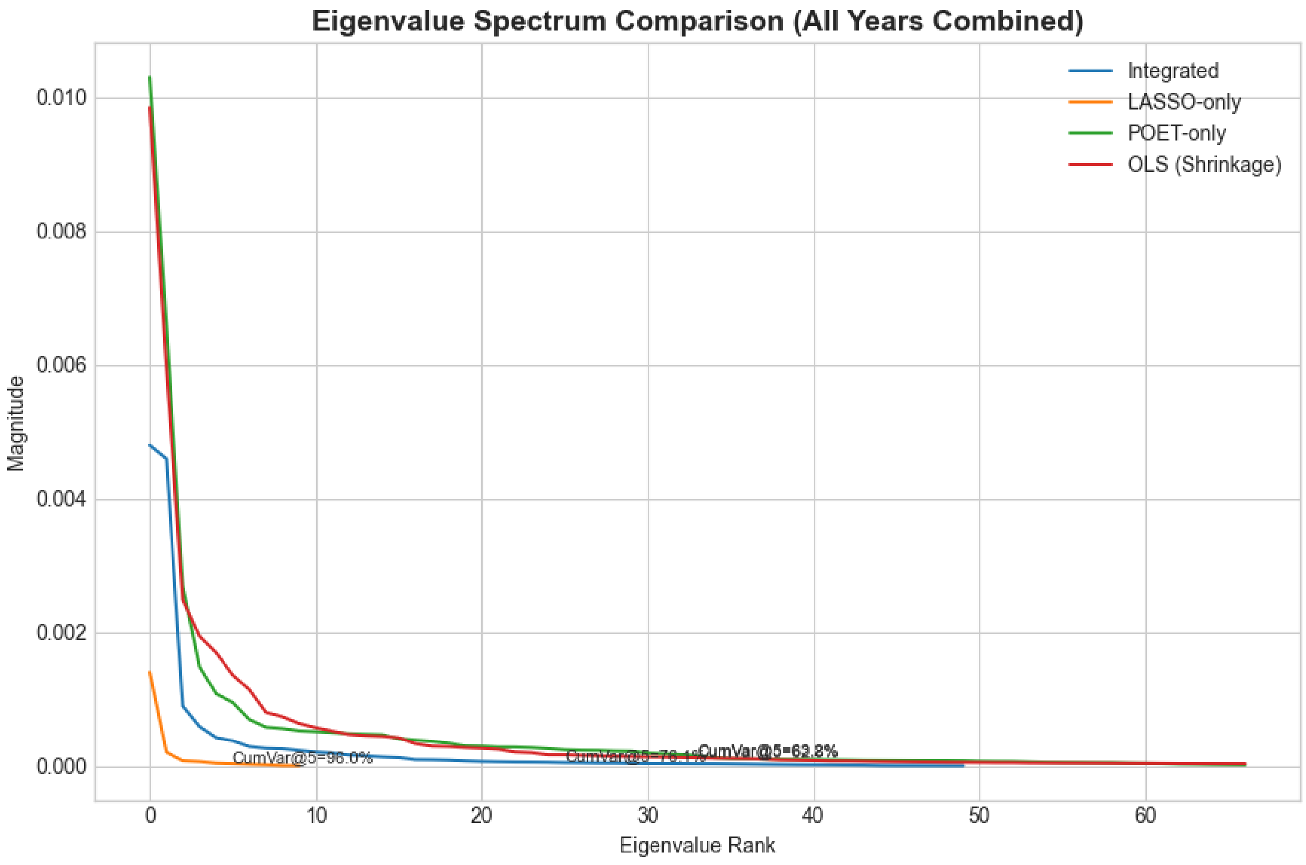


4 Result Analysis

POSTECH

Eigenvalue Spectrum Comparison

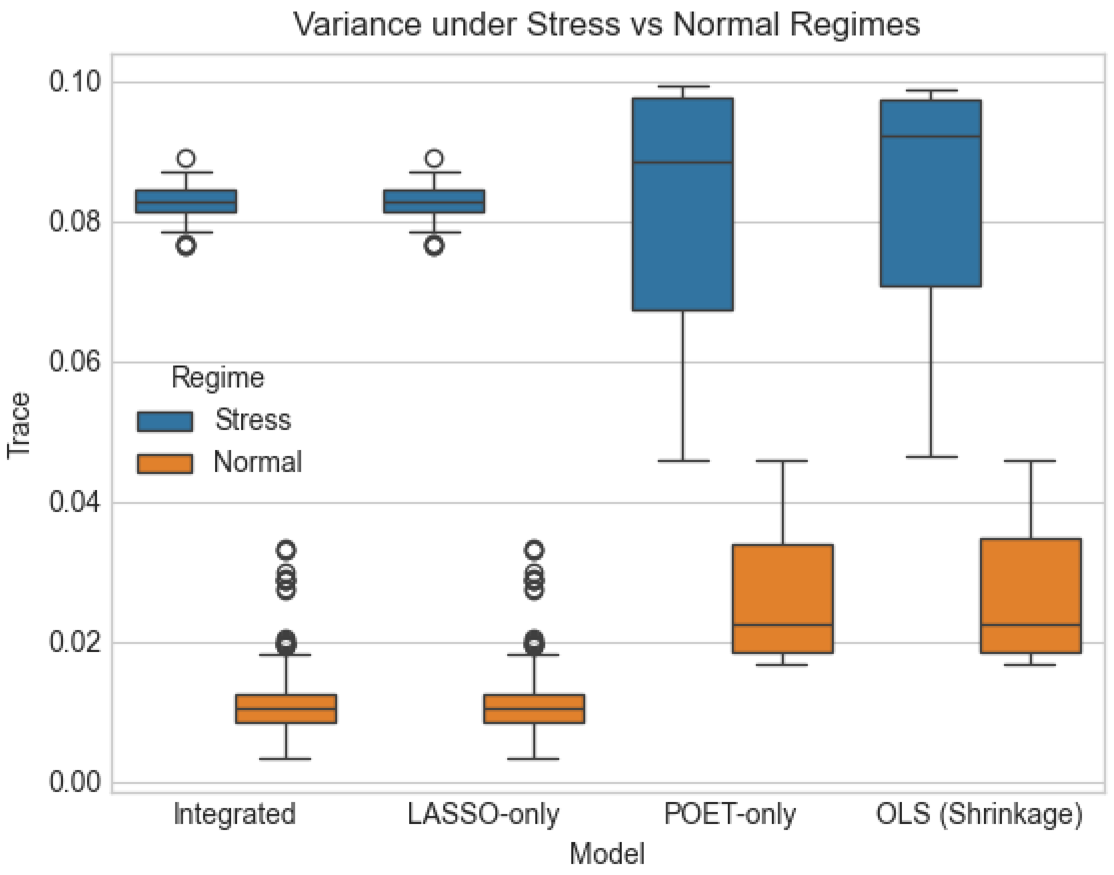
- **Integrated model** shows a **smooth and gradual eigenvalue decay**, indicating **balanced factor contributions** and improved diversification.
- **LASSO-only** collapses rapidly with few dominant eigenvalues, suggesting **over-sparsification** and loss of secondary risk structure.
- **OLS** and **POET-only** display heavier tails, implying **redundant or noisy factor components**.
- Overall, the **Integrated estimator captures richer latent structure**, preserving essential covariance geometry while suppressing noise.



Result Analysis

Variance under Stress vs. Normal Periods

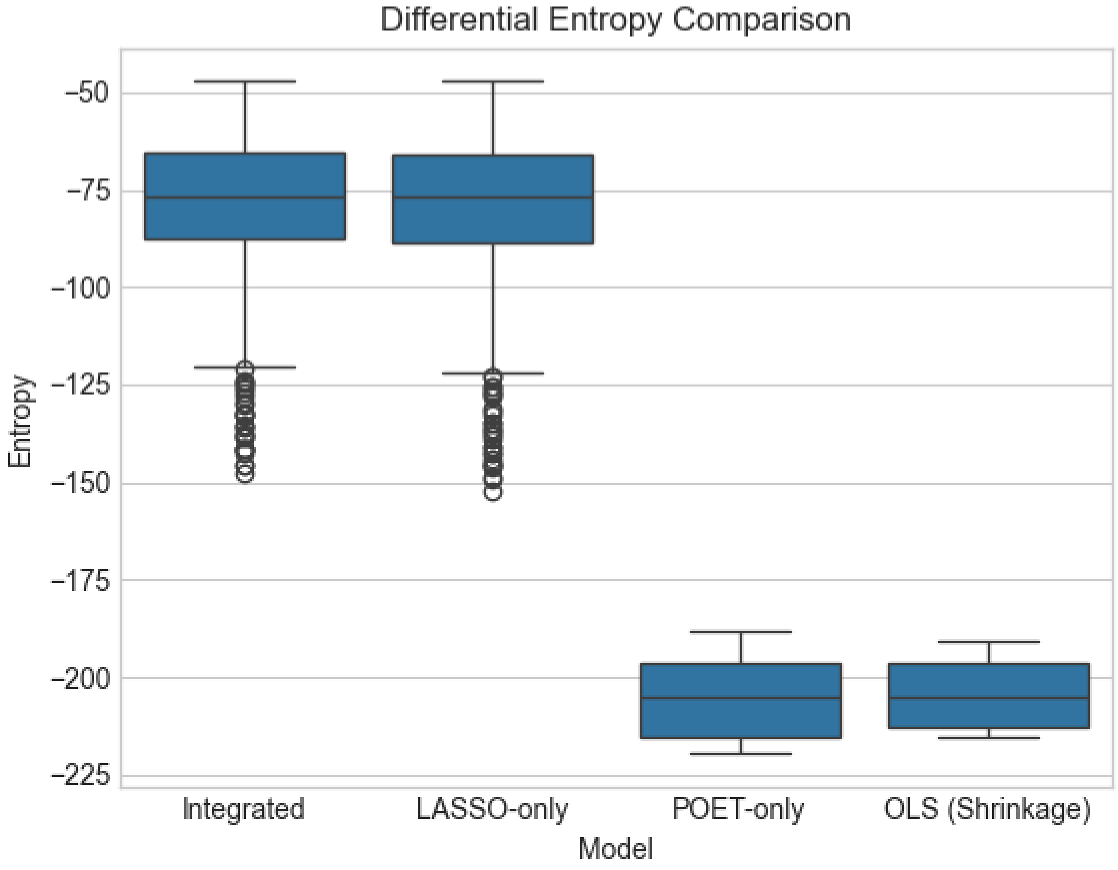
- **Integrated model** maintains the **most stable variance** across both stress and normal regimes, showing controlled volatility and strong adaptability under market shocks.
- Even during high-volatility periods, its total portfolio variance remains well-controlled, reflecting **robust adaptability** and **effective suppression of noise-driven covariance inflation**.
- **LASSO-only** remains relatively stable but **underestimates systemic risk** due to over-sparsification.
- **POET-only** and **OLS** exhibit **large variance jumps and dispersion**, indicating **instability and overreaction** during stress periods.
- Overall, the Integrated model achieves **balanced risk sensitivity**, reacting appropriately to market stress while preserving estimation stability.



4 Result Analysis

Differential Entropy Comparison

- **Integrated** and **LASSO-only** models exhibit **significantly higher entropy**, capturing **richer dependency structures** and a broader spread of risk factors.
- In contrast, **POET-only** and **OLS (Shrinkage)** show **very low entropy**, reflecting **over-compressed or oversmoothed covariance structures** that may ignore meaningful cross-asset variations.
- The **Integrated model** preserves LASSO's structural diversity while maintaining stability — balancing **information richness** and **robustness**.
- High entropy combined with stable variance (previous slide) suggests that the Integrated estimator **captures complexity without instability**.

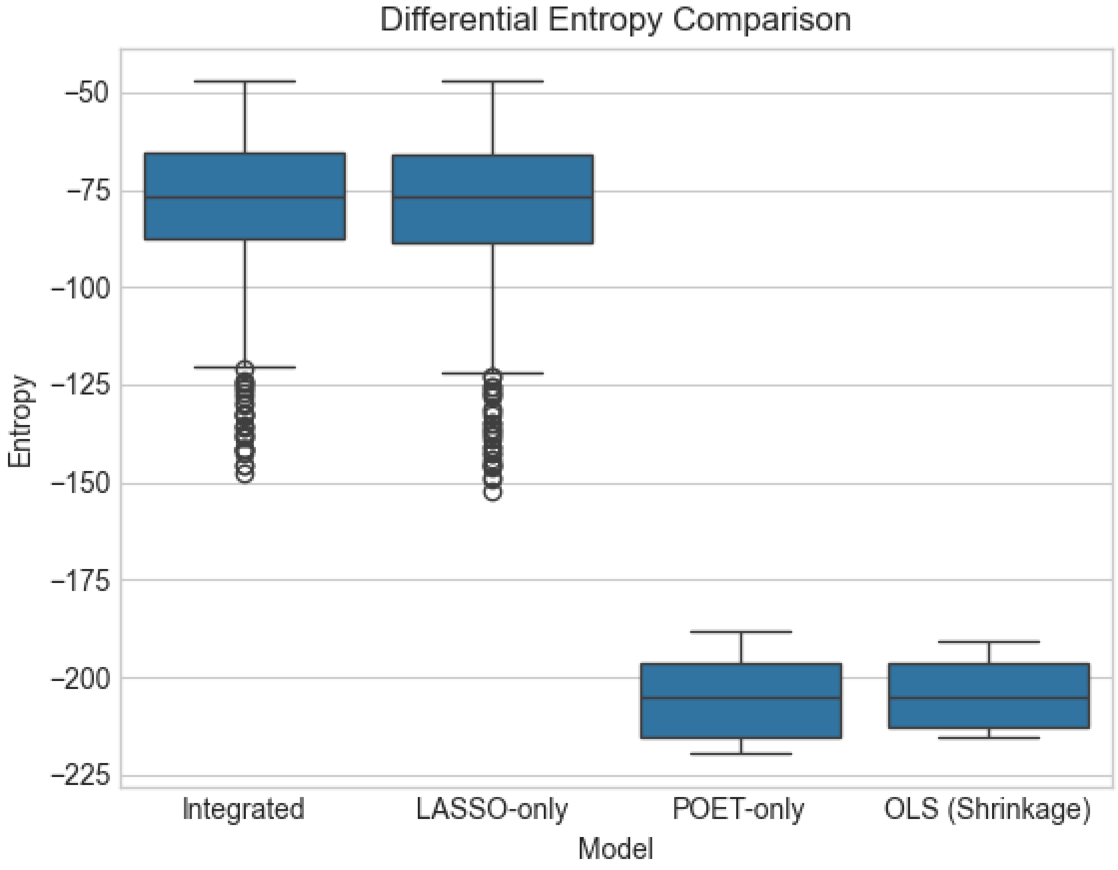


4 Result Analysis

POSTECH

Differential Entropy Comparison

- **Integrated** and **LASSO-only** models exhibit **significantly higher entropy**, capturing **richer dependency structures** and a broader spread of risk factors.
- In contrast, **POET-only** and **OLS (Shrinkage)** show **very low entropy**, reflecting **over-compressed or oversmoothed covariance structures** that may ignore meaningful cross-asset variations.
- The **Integrated model** preserves LASSO’s structural diversity while maintaining stability — balancing **information richness** and **robustness**.
- High entropy combined with stable variance (previous slide) suggests that the Integrated estimator **captures complexity without instability**.



5 Discussion / Interpretation

Why does the Integrated Model perform better?

Cluster-based Residual Sparsifiaction

- The Integrated estimator applies **hard-thresholding within correlation clusters**, retaining **intra-cluster dependencies** while suppressing **cross-cluster noise**.
- This yields a **block-sparse covariance structure** that mirrors real market segmentation — e.g., sectoral or macro-style partitions (Equity / Commodity / FX).
- As a result, estimation noise is reduced **without erasing genuine structural relationships**,
achieving a rare balance between **bias reduction** and **variance control**.

Statistical Implications : Bias-Variance Tradeoff

- LASSO-only induces **high bias (over-sparsification)**, while OLS and POET suffer from **high variance (overfitting / dense noise)**.
- The Integrated approach acts as a **structural regularizer** as shrinking noise-dominated correlations across clusters and preserving informative variance within clusters.
- This mechanism **reduces estimation error (Frobenius, KL)** and **stabilizes volatility forecasts (Risk Gap)** simultaneously.

5 Discussion / Interpretation

POSTECH

Why does the Integrated Model perform better?

Dynamic Robustness Under Market Regime Shifts

- Cluster structures evolve slowly compared to individual asset correlations — thus providing **natural regularization** across rolling windows.
- Under stress regimes, Integrated maintains **bounded covariance trace**, preventing blow-ups that appear in POET-only and OLS.
- The model **reacts to systemic shocks** (entropy remains high) but avoids unstable noise amplification — ensuring **regime-consistent sensitivity**.

Information-Theoretic Interpretation

- High **differential entropy** in Integrated and LASSO models implies **richer latent dependency capture**, whereas low-entropy OLS/POET estimates indicate **information loss via oversmoothing**.
- Integrated's moderate entropy dispersion shows it **retains structural complexity** while maintaining **predictive stability**.

The Integrated model succeeds because it embeds structural prior knowledge (cluster topology) into statistical regularization.

It exploits the **natural modularity of financial markets** — filtering noise *across* clusters while preserving dependence *within* clusters.

This structural sparsification translates into **lower estimation error**, **higher Sharpe efficiency**, and **superior stability across regimes.**

6 Conclusion

POSTECH

Model Design Adjustment

- In the original proposal, we intended to extend the framework with a **Spatial Autoregressive (SAR)** layer to explicitly model inter-asset dependencies within local subgroups.
- However, during implementation, we found that **the subgroup-based autoregressive structure in SAR** plays a similar role to **the latent factor extraction in POET** — both capture **cross-sectional dependency compression**.
- To avoid **redundant modeling of local dependency**, we **adopted only the subgrouping mechanism from SAR**, applying it **prior to POET decomposition** to form the **group masks used in idiosyncratic thresholding**.
- This modification ensures that **grouping precedes factor extraction**, aligning the cluster topology consistently across both systematic and idiosyncratic components.

Summary

- We proposed an **Integrated covariance estimation framework** combining **LASSO-based asset selection**, **cluster-wise residual sparsification**, and **POET-style factor decomposition**.
- The model achieves **lower estimation error**, **stronger regime stability**, and **higher Sharpe efficiency** by preserving meaningful intra-cluster dependencies while suppressing cross-cluster noise.
- Empirical results confirm that **cluster-aware structural regularization** improves both risk forecast reliability and distributional fidelity compared to standalone methods.