Spectral Properties of Sample Covariance Matrices

Relationship Between
Variance Concentration & Average Correlation

September 2024

Decoding the Matrix

High-dimensional, low-sample-size scenarios (e.g., financial datasets, machine learning) pose unique statistical challenges and exhibit distinct properties for covariance matrices.

Assume a few key drivers dominate market covariance. Spectral decomposition of the sample return data's covariance matrix yields:

- Eigenvalues and eigenvectors representing market structure

Two key metrics derived from this process:

1. Fraction of variance explained by the leading eigenvector.
2. Average pairwise correlation.

Key question:

What is the relationship between these metrics, and why is it important?

Empirical Test

MARKETS

Daily returns for constituent stocks of *the US S&P 500 and China CSI 300.*

DATE RANGE

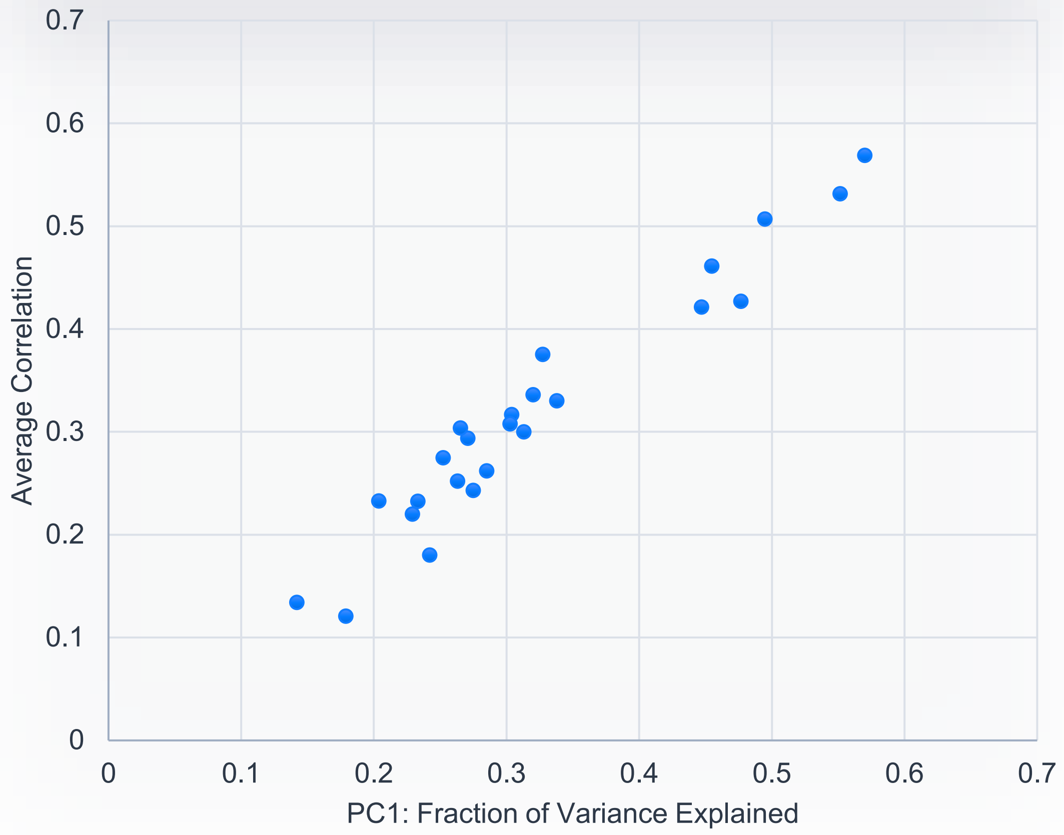
2000/01/01 – 2023/12/31

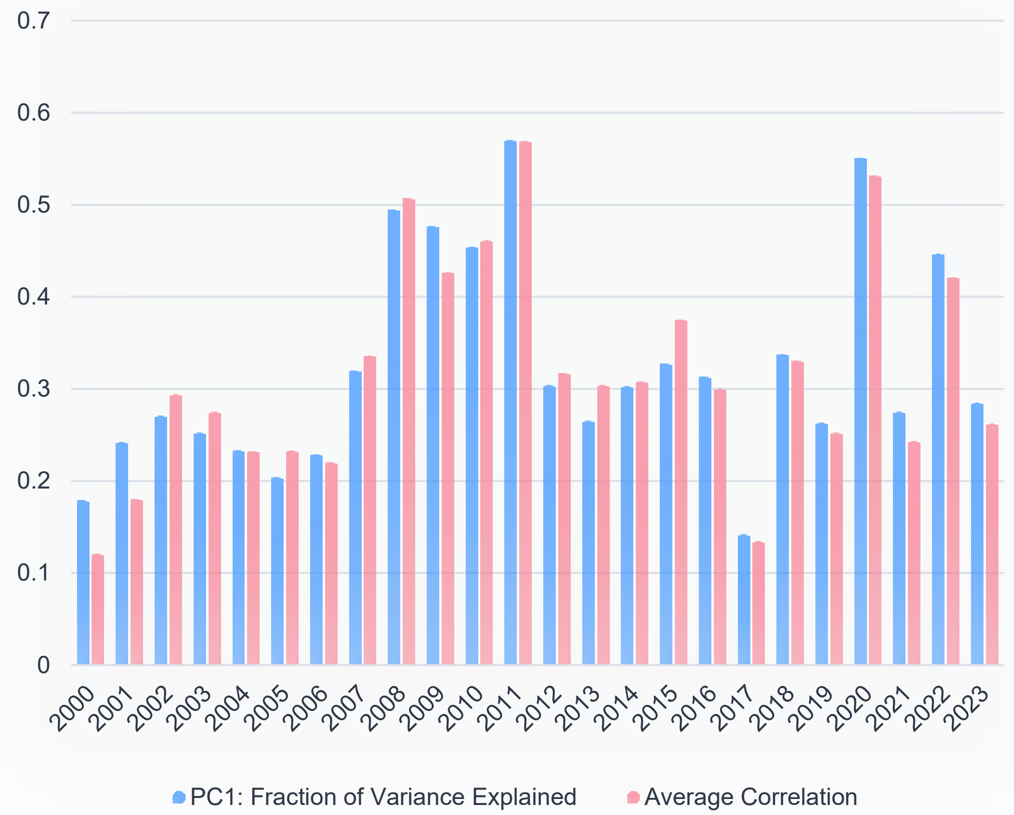
STEPS

1. One year's worth of daily returns were used to estimate covariance.
2. The fraction of variance explained by the leading eigenvector was calculated.
3. The average correlation among all pairs of constituent stock returns was computed.
4. This process was repeated for each subsequent year, comparing the fraction of variance explained by the leading eigenvector with the average correlation for each year.

The US S&P 500 Constituents

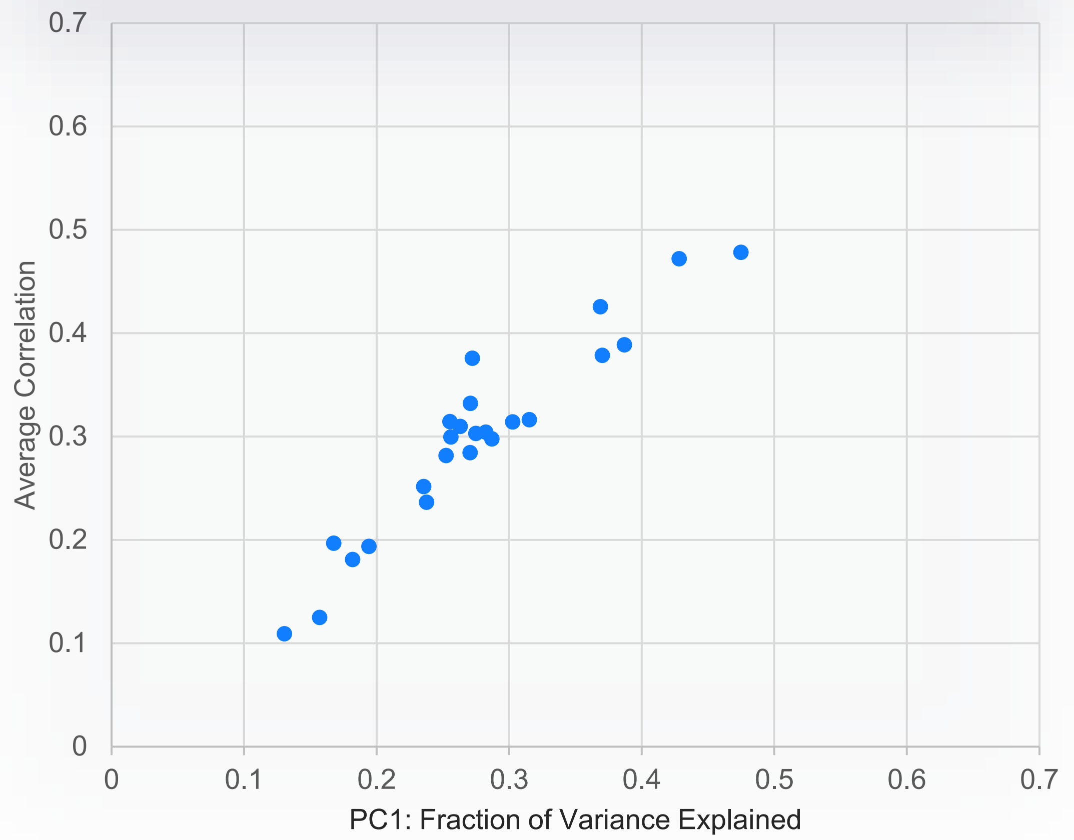
A Linear Relationship between the Two Metrics

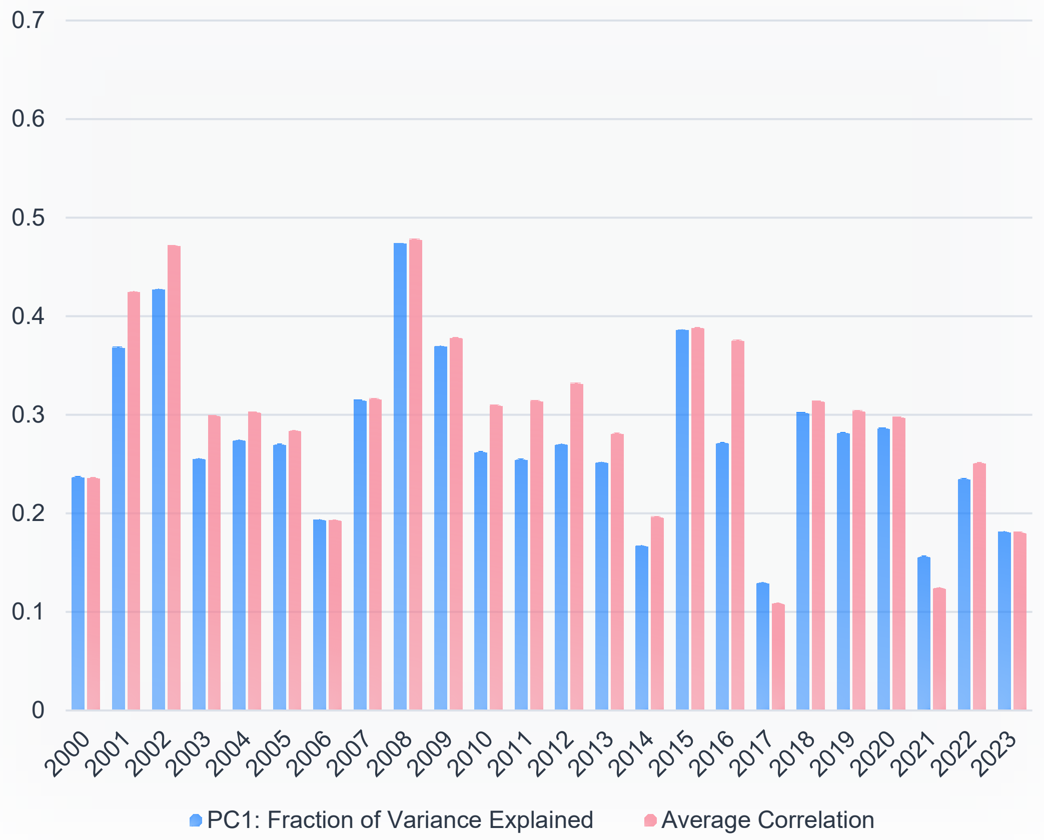




China CSI 300 Constituents

A Linear Relationship between the Two Metrics





Simulation Test

Simulation Test

A strong relationship between the fraction of variance explained by the leading eigenvector and the average correlation has been observed. An analysis on this is, in one-factor model:

$$r_i = \beta_i f + \epsilon_i$$

Under assumptions: $E(\epsilon_i) = 0, E(\epsilon_i f) = 0 \; and \; E(\epsilon_i \epsilon_j) = 0$, the formula for correlation $\rho(i,j)$ between securities i and j becomes:

$$\rho(i,j) = \frac{\beta_i \beta_j \sigma^2}{\sqrt{\beta_i^2 \sigma^2 + \delta_i{}^2} \sqrt{\beta_j^2 \sigma^2 + \delta_j{}^2}}$$

When ① exposures to the factor, β have low dispersion and are equal to $1/\sqrt{p}$

② specific variances are identical

$$\rho(i,j) \approx \frac{\sigma^2/p}{\frac{\sigma^2}{p} + \delta^2}$$

$$= \frac{\sigma^2}{\sigma^2 + p\delta^2}$$

p - number of securities

Next Pages:
Simulate scenarios ① and ② to test effect on relationship between average correlation $\bar{\boldsymbol{\rho}}(\boldsymbol{i}, \boldsymbol{j})$ and Fraction of Variance Explained by the leading eigenvector.

One-factor Simulation Setup

In one-factor model:

$r_i = \beta_i f + \epsilon_i$

Simulate 500 securities with 252 returns,

Simulate f in normal distribution, shape 1 x 252, μ_f = 0, σ_f = $0.16/\sqrt{252}$

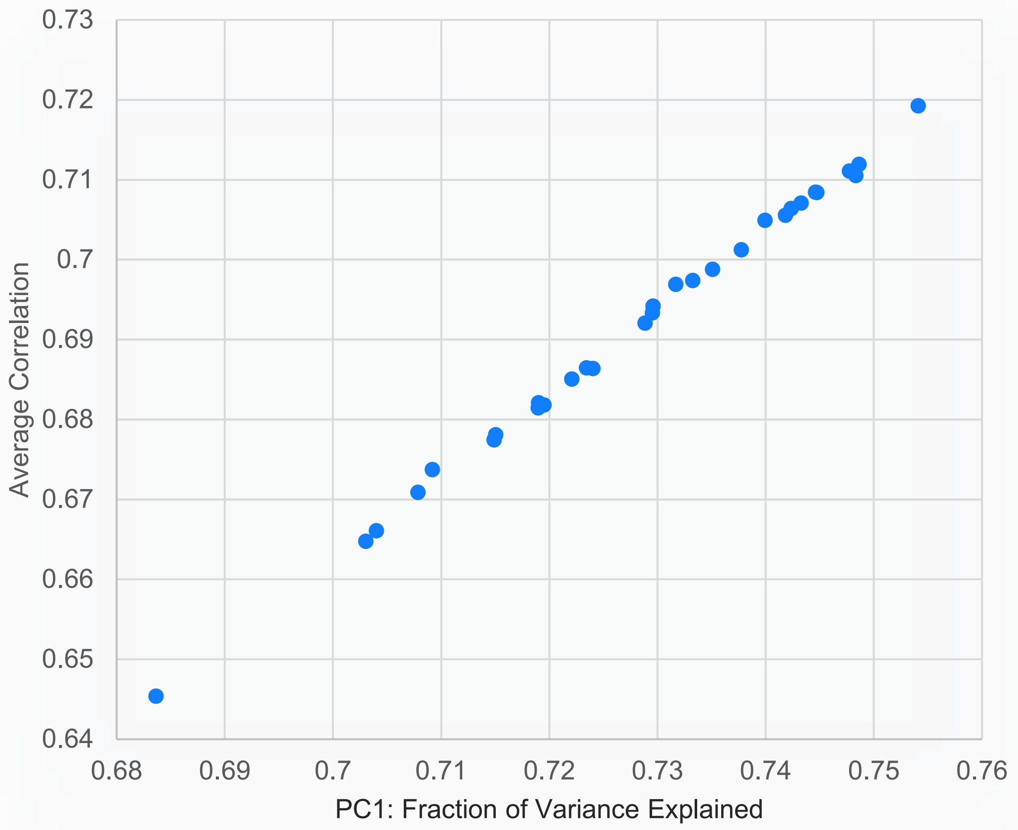
① Simulate β in normal distribution, shape 500 x 1, μ_β = 1, σ_β from 0.25 to 0.05, β becoming less dispersed.

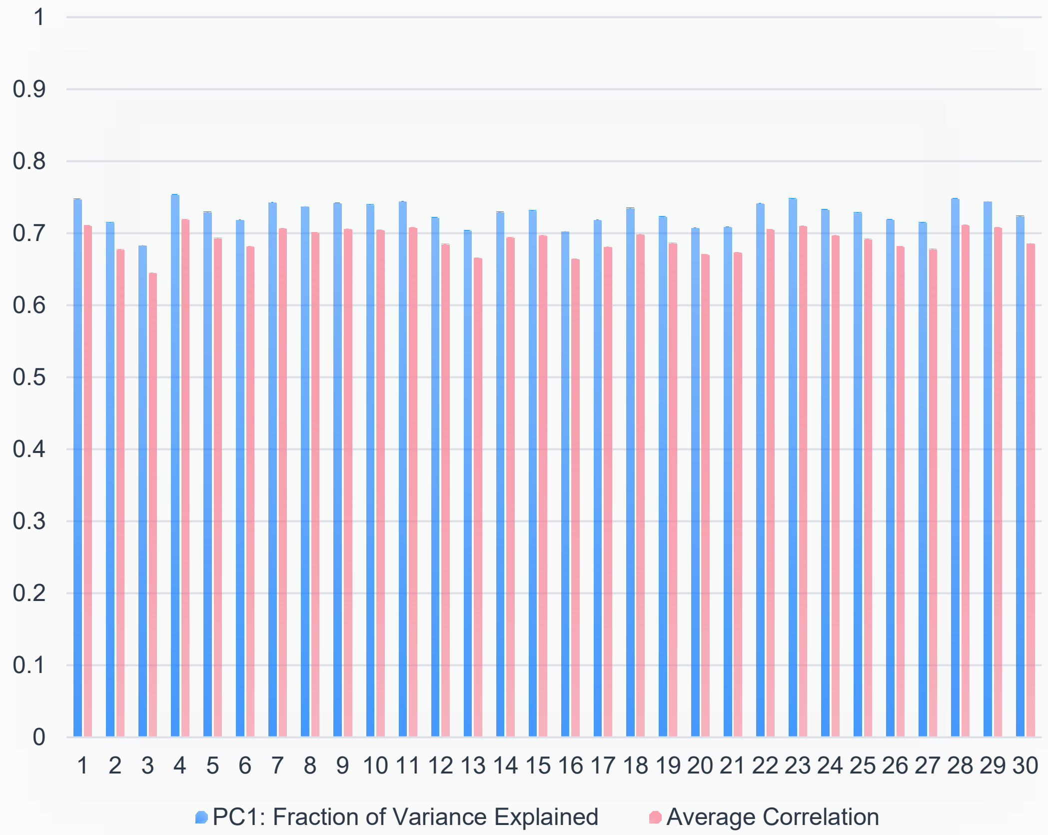
② Simulate ϵ in normal distribution, shape 500 x 252, μ_ϵ = 0, σ_ϵ from $0.5/\sqrt{252}$ to $0.1/\sqrt{252}$, e becoming less dispersed, δ^2 becoming more identical.

Each setup is experimented 30 times to create box plots

Simulation Result

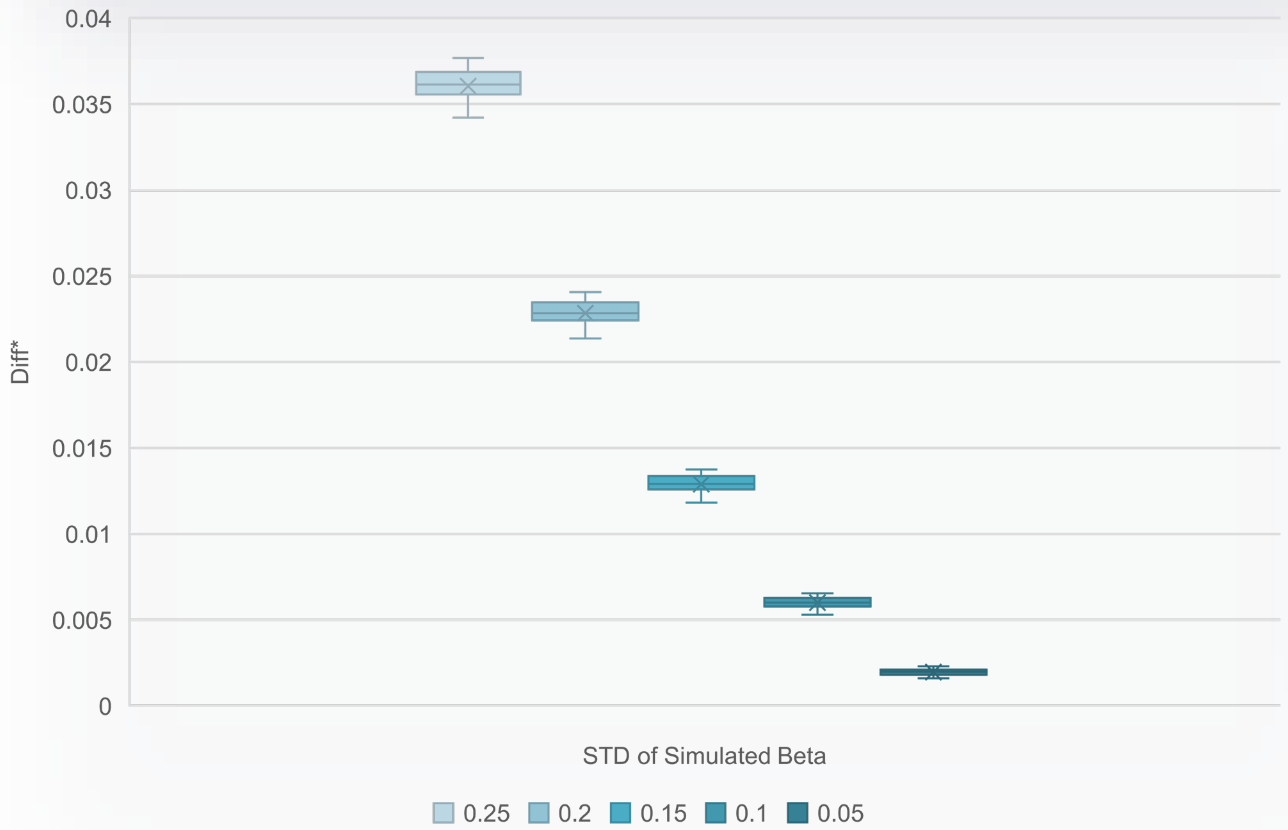
Relationship between Fraction of Variance Explained by the Leading Eigenvector and Average Correlation in a controlled environment





Simulation Result: Reducing Beta Dispersion

When reducing beta dispersion, diff decreases



*Diff: FracVar - AvgCorr

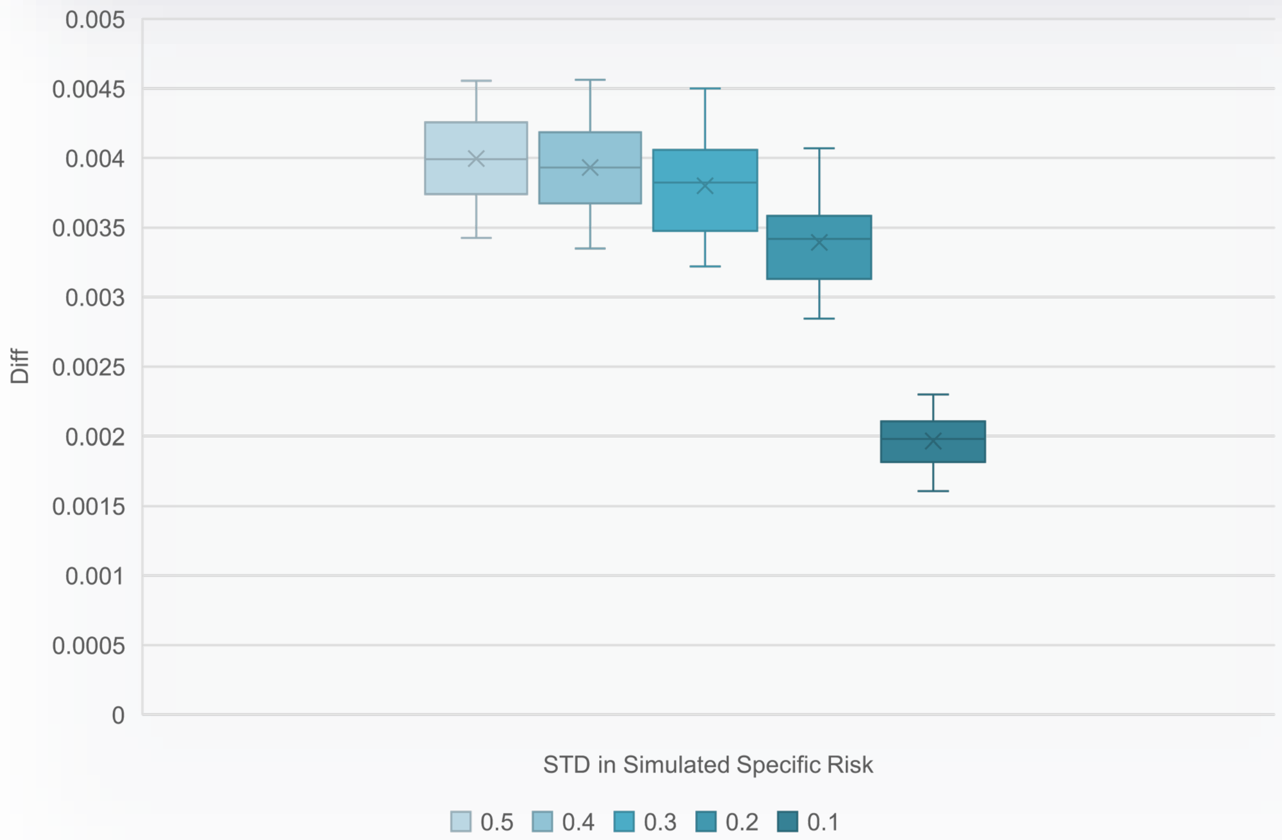
① when exposures to the factor, β have low dispersion

$$\rho(i,j) = \frac{\boldsymbol{\beta}_i \boldsymbol{\beta}_j \sigma^2}{\sqrt{\boldsymbol{\beta}_i^2 \sigma^2 + {\delta_i}^2}\sqrt{\boldsymbol{\beta}_j^2 \sigma^2 + {\delta_j}^2}}$$

$\sigma_\beta \downarrow \rightarrow \text{diff\%} \downarrow$

Simulation Result: Reducing Beta Dispersion

When reducing specific risk dispersion, diff decreases



② when specific variances are identical

$$\rho(i,j) = \frac{\beta_i \beta_j \sigma^2}{\sqrt{\beta_i^2 \sigma^2 + \boldsymbol{\delta_i}^2}\sqrt{\beta_j^2 \sigma^2 + \boldsymbol{\delta_j}^2}}$$

$\sigma_\epsilon \downarrow \rightarrow \text{diff\%} \downarrow$

What's Next:

Estimate Correlation Matrix with Different Numbers of Factors

Estimating Correlation Matrix

Note: Steps to estimate the sample correlation matrix

1. **Assuming that a few key drivers account for most of the market correlation, let's suppose the S&P 500 stock returns data follow a factor model.**
2. **Center returns data to mean zero and compute $p \times p$ sample covariance matrix S from daily returns data.**
3. **Spectral decomposition of the covariance matrix:**
   The sample covariance matrix S can be decomposed into its eigenvalues and eigenvectors:

$$S = \sum_{i=1}^{p} \lambda_i v_i v_i^{\top}$$

   where λ_i are the eigenvalues and v_i are the corresponding eigenvectors of S. These eigenvalues are sorted such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.
4. **Use k factors to estimate covariance and replace the small components with matrix g.**

$$S = \sum_{i=1}^{k} \lambda_i v_i v_i^{\top} + g$$

5. **Estimate diagonal terms on matrix g using a heterogeneous or a homogeneous specific variance matrix.**
   ① Heterogeneous specific variance estimation (credit to Alex Bernstein):

$$diag(g) = diag\left(S - \sum_{i=1}^{k} \lambda_i v_i v_i^{\top}\right)$$

   ② Homogeneous specific variance estimation:

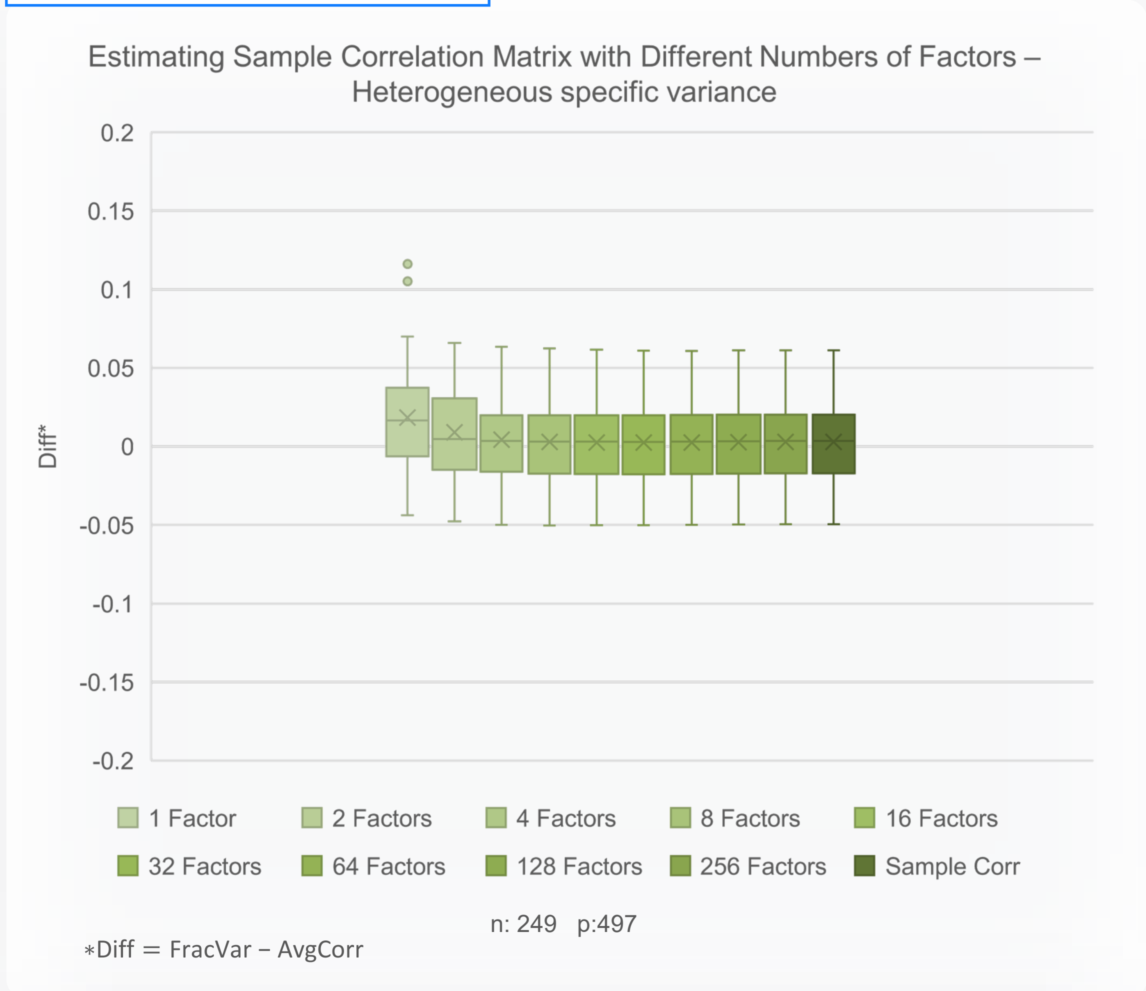
$$\delta^2 = \left(\frac{n}{p}\right) \ell^2$$

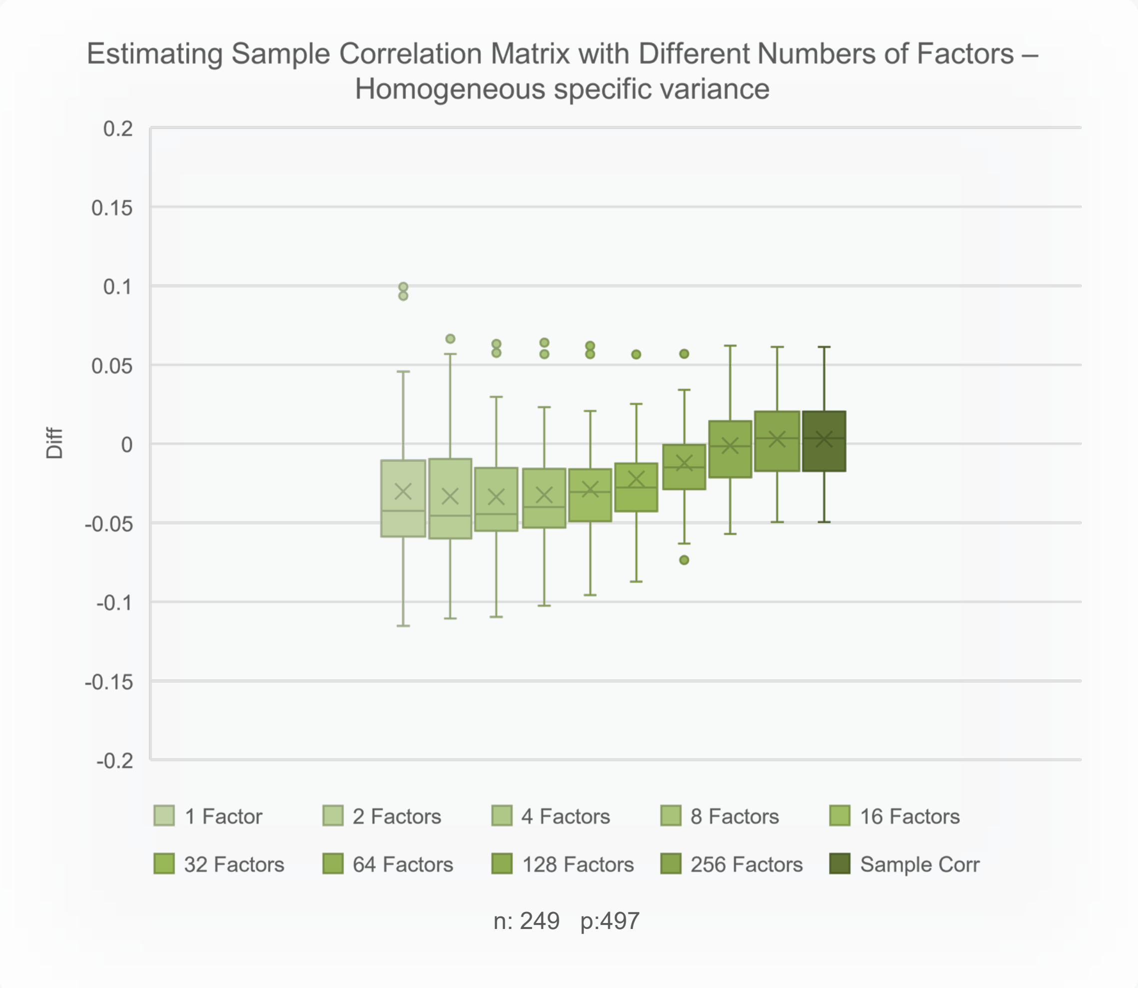
$$diag(g) = \delta^2 I$$

   ℓ^2 – Average of remaining non−zero eigenvalues
   I – Identity matrix
6. **Convert the estimated covariance matrix to a correlation matrix by dividing means of variances.**

Changing factor number to estimate sample correlation matrix

Sample Correlation Matrix

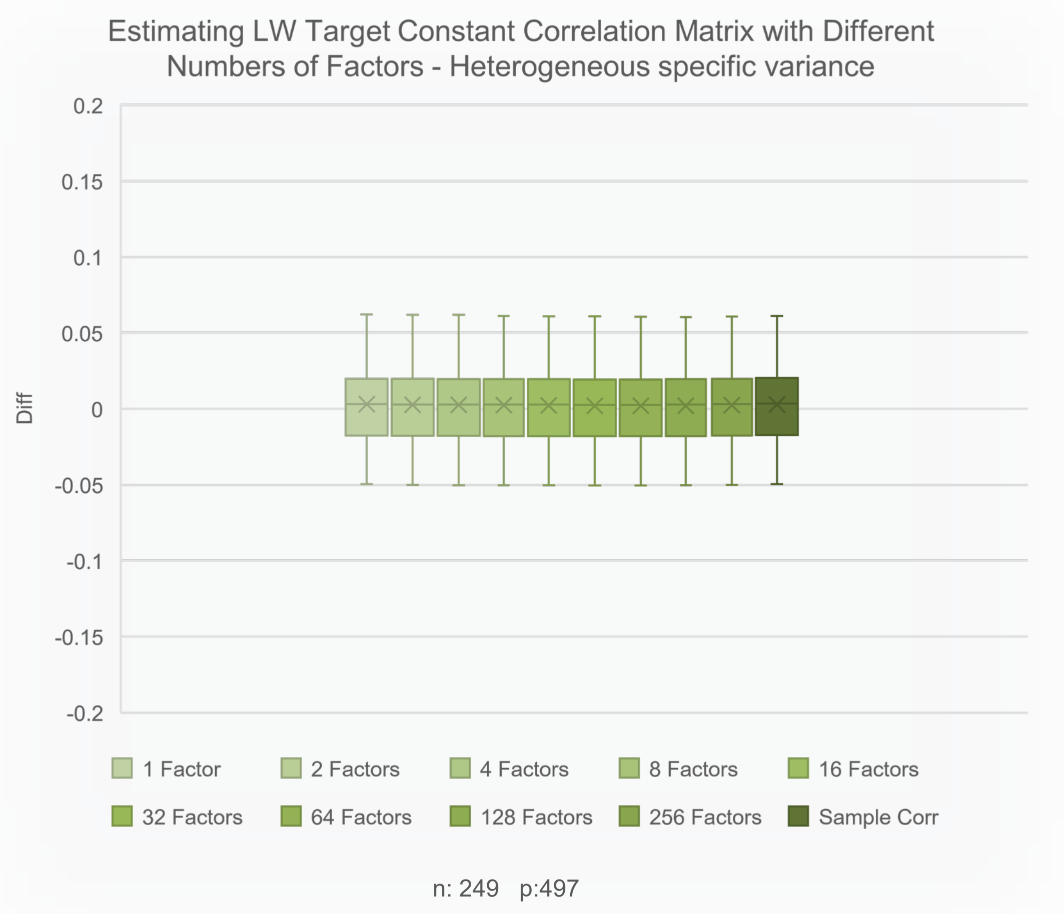


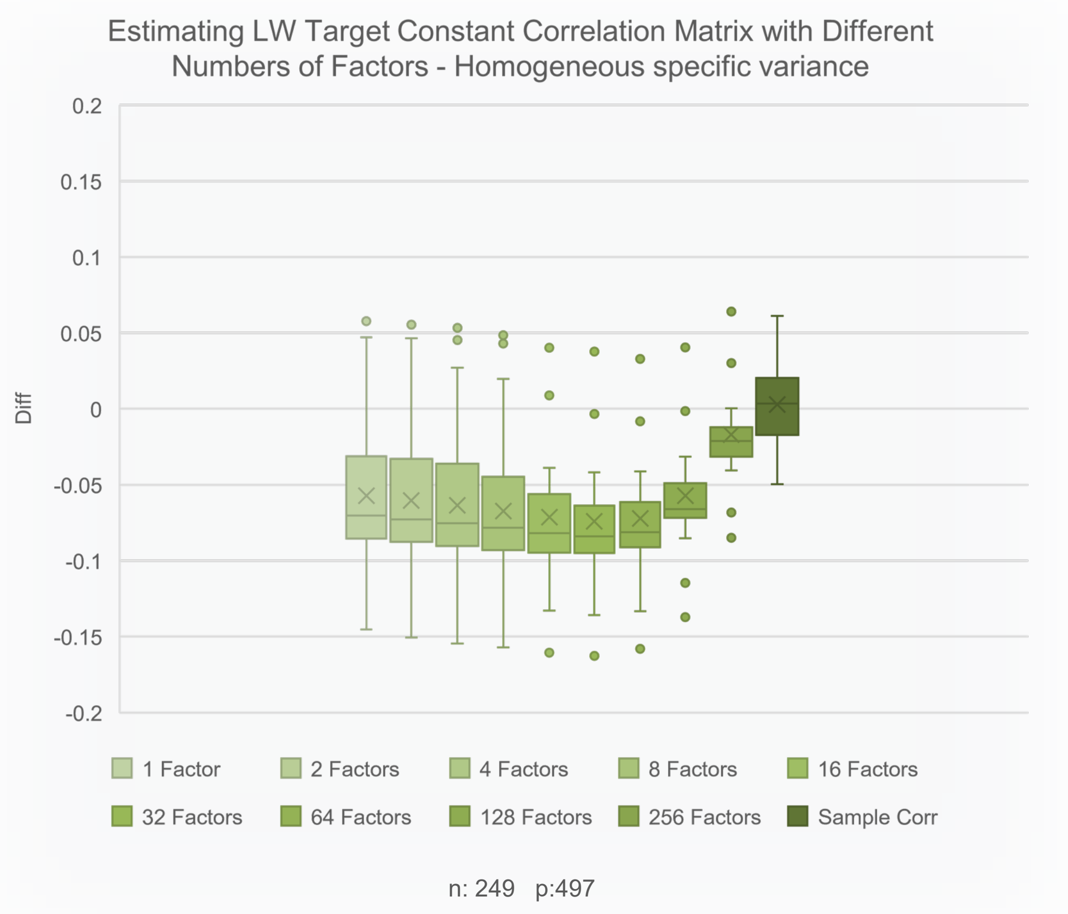


The average correlation remains largely unchanged after estimating the correlation matrix with 4 factors

Changing factor number to estimate LW target constant correlation matrix

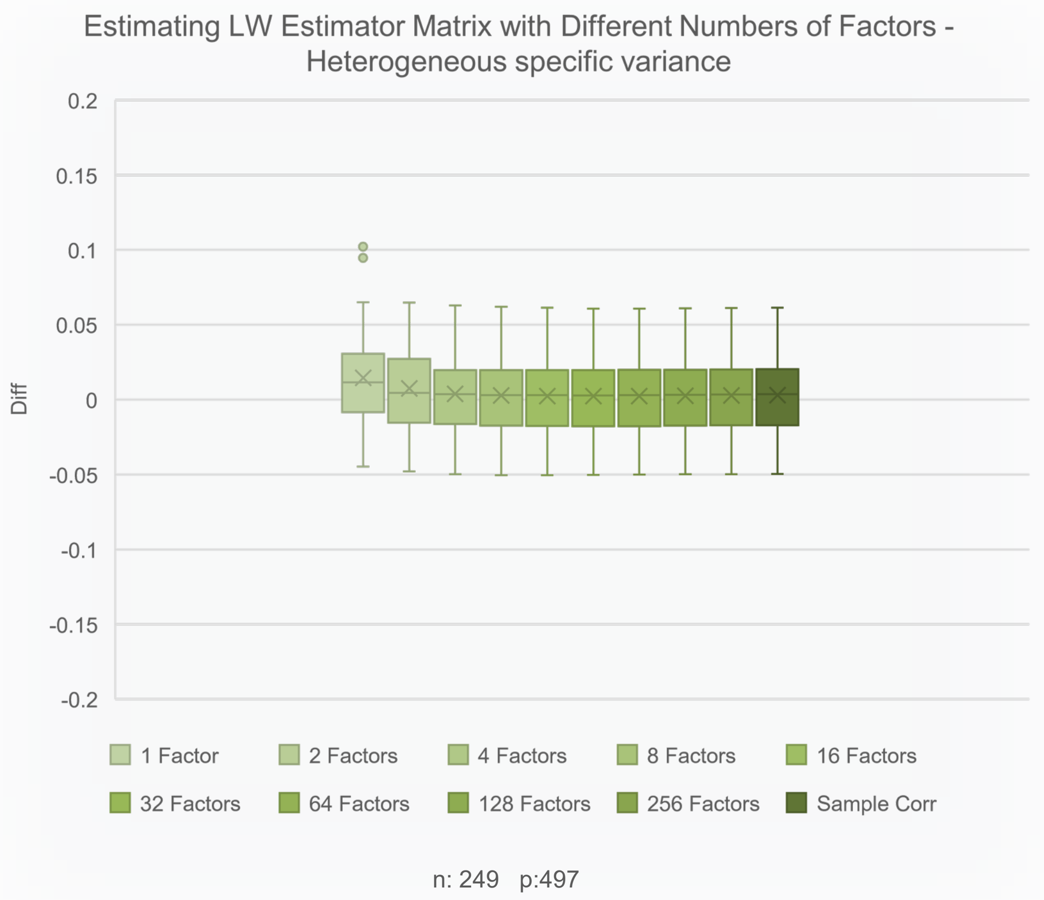
LW Target Constant Correlation Matrix

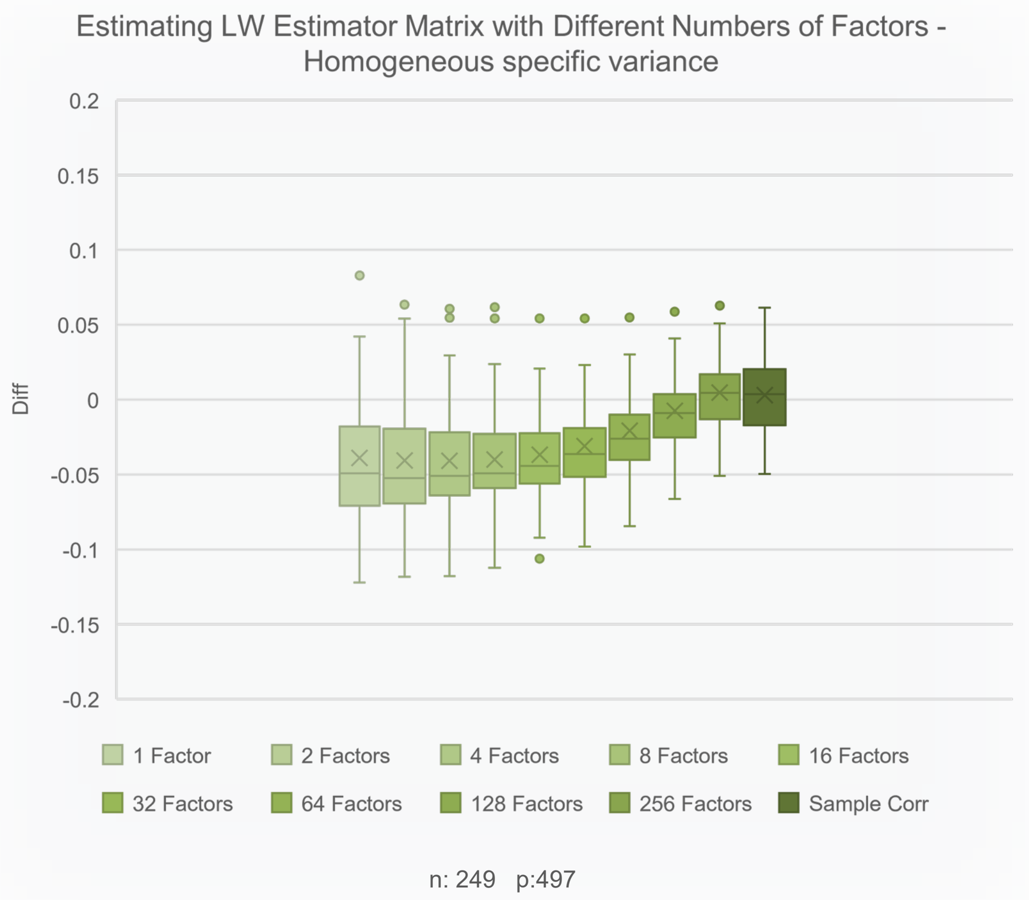




Changing factor number to estimate LW target constant correlation matrix

LW Estimator Matrix





References

Goldberg, L. R., Papanicolaou, A. & Shkolnik, A. (2022), ‘The dispersion bias’, *SIAM Journal on Financial Mathematics, 13* (2), 521–550.

Ledoit, O. & Wolf, M. (2004), ‘Honey, I shrunk the sample covariance matrix’, *The Journal of Portfolio Management, 30*, 110–119.