Empirical Findings on Spectral Properties of Return Covariance Matrices

Dayi (Darwin) Yao, Jingyuan Chen

Special thanks to: Prof. Goldberg, Prof. Shkolnik, Rahul, Rahul Pothi Vinoth, Harrison Selwitz, and Jacob Lan

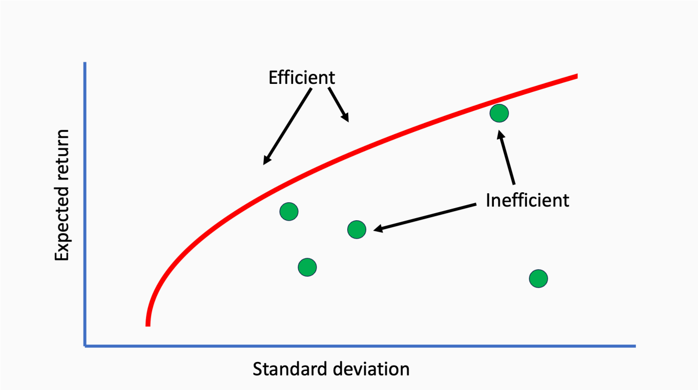
University of California, Berkeley

December 6, 2024

Contents

Introduction

In 1952, Harry Markowitz launched modern finance by framing portfolio construction a tradeoff between portfolio expected return and risk, and providing a mathematical mechanism to optimize portolios



Evidently realizing that classical statistics would not provide what he needed, Markowitz considered alternative ways to estimate optimization inputs

Perhaps there are ways, by combining statistical techniques and the judgment of experts, to form reasonable probability beliefs μ_{ij}, σ_{ij} One suggestion as to tentative μ_{ij}, σ_{ij} is to use the observed μ_{ij}, σ_{ij} for some period of the past. I believe that better methods, which take into account more information, can be found. I believe that what is needed is essentially a "probabilistic" reformulation of security analysis. I will not pursue this subject here, for this is "another story." It is a story of which I have read only the first page of the first chapter.

Harry Markowitz (1952)

Big Idea #1: Since they conform to empirically observed properties of financial data and reduce dimension, factor models are used almost universally to generate inputs to mean-variance optimization

The return generating process

$$r = \beta f + \epsilon$$

implies the expected returns:

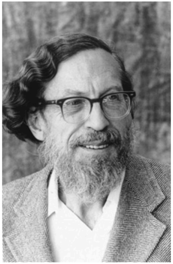
$$E[r] = \beta E[f] + E[\epsilon]$$

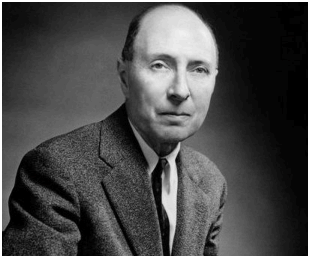
and covariance matrix:

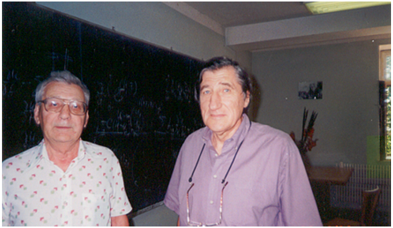
$$\Sigma = \beta F \beta^\top + \Delta$$

Returns or excess returns r are the sum of factor returns f scaled by exposures β and specific returns ϵ, which are pairwise uncorrelated and uncorrelated with factor returns. Returns are observable but the factor and specific components are not. The factor and (diagonal) specific return matrices are denoted by F and Δ.

Big Idea #2: Random matrix theory provides tools to estimate factor model parameters in high dimensions when data are scarce.







Spectral decomposition and covariance matrix estimation

Eigenvalues and eigenvectors of noisy sample return covariance matrices for large cap equities provide the components of factor-based covariance matrices used in Markowitz optimization

We identify salient characteristics of sample return covariance matrices, and provide some answers to these questions:

- How many spiked eigenvectors (factors) do we typically see, and how does that number vary over time.
- How do spiked eigenvalues depend on the number of securities in the estimation universe?
- How are the entries of spiked eigenvectors distributed?

Data and Methodology

Data

- Our dataset consists of an approximation of the Russell 3000 constituents based on the BlackRock iShares ETF IWV (Oct 09, 2024)
- Using WRDS Center for Research in Security Prices data, we retrieved the Daily Total Return (DlyRet) and Daily Market Capitalization (DlyCap) from January of 2003 to December of 2023 for each Ticker obtained from the above ETF.
- Securities without complete history are dropped, we have an effective maximum of 2340 securities.

Methodology Overview

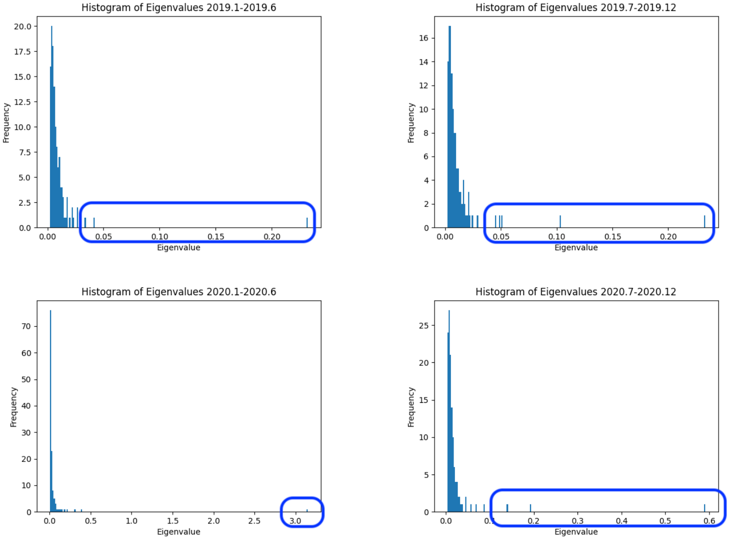
- Covariance Matrix Estimation:
  - Look-back window of 126 trading days
  - Sample covariance matrix computed for each window
- Spectral Analysis:
  - Spectral decomposition of sample covariance matrices
  - Focus on leading eigenvalues and corresponding eigenvectors
  - Comparison between market-cap sorted and "randomly" selected stocks
- Time Period Analysis:
  - Eight distinct 126-day periods (2019-2022)
  - Attention paid to market stress periods (e.g., 2020 pandemic)

How many factors?

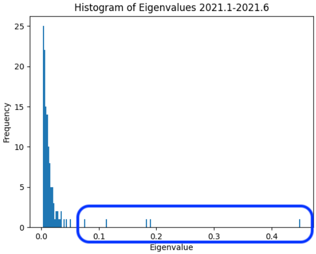
How many factors

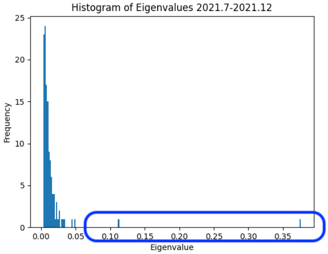
- How many spiked eigenvectors (factors) do we typically see
- How does that number vary over time
- Eight time periods of 126 days each are chosen (2019 - 2022)
- The spectrum of eigenvalues is used to identify the number of factors

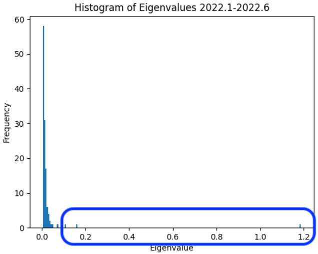
Observation over 8 time periods (Part I)

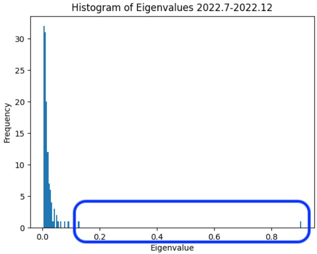


Observation over 8 time periods (Part II)









Observation over 8 time periods

Period	Factor Number
2019.1 - 2019.6	3
2019.7 - 2019.12	5
2020.1 - 2020.6	1
2020.7 - 2020.12	3

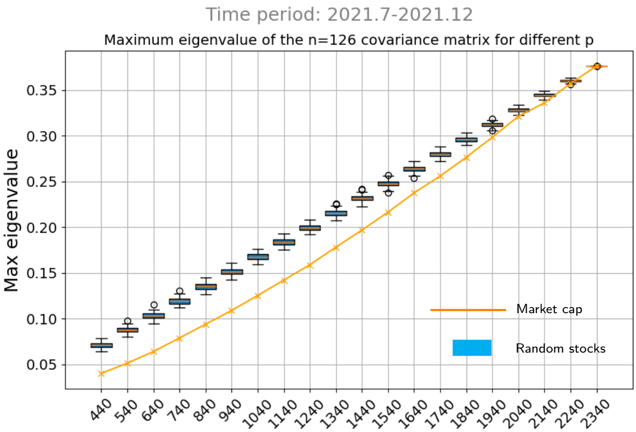
Period	Factor Number
2021.1 - 2021.6	5
2021.7 - 2021.12	2
2022.1 - 2022.6	3
2022.7 - 2022.12	2

- In normal cases, the number of outstanding factors is around 4
- In a financial crisis, the number of factor concentrates to one
- e.g. the pandemic (2020.1 - 2020.6)

Question: Is this a good way to count factors? Can we have a more systematic approach?

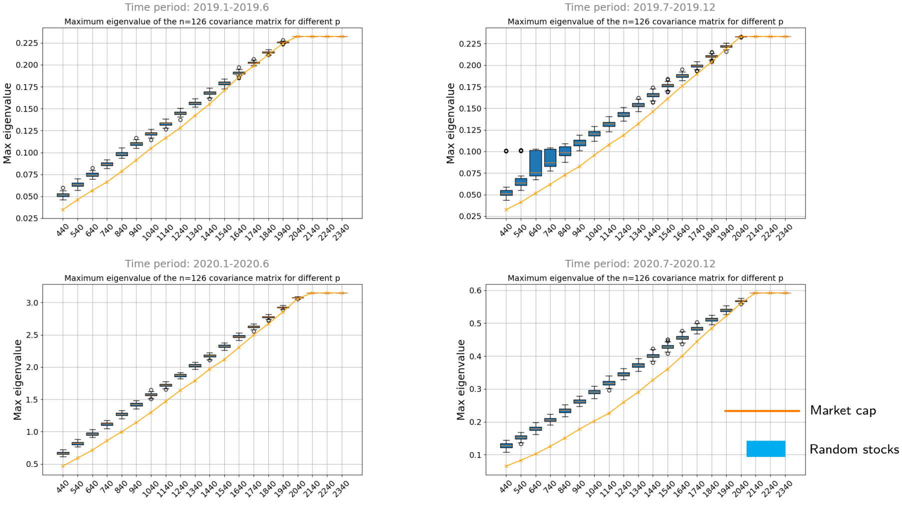
How do Eigenvalues grow with respect to the number of securities

The leading eigenvalue shows roughly affine dependence on the number of securities p



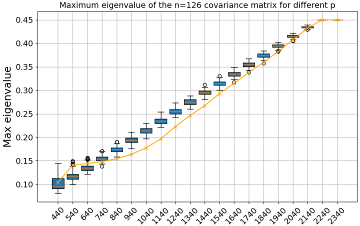
Covariance matrix estimated EOY 2021 using trailing 126 days of data. Stocks are sorted by market capitalization (orange line) or randomly drawn for each p (blue box plots).

Over 8 time periods (Part I)

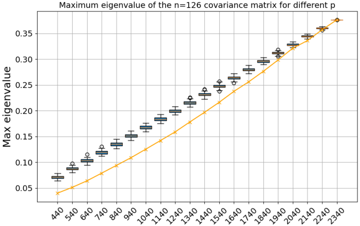


Over 8 time periods (Part II)

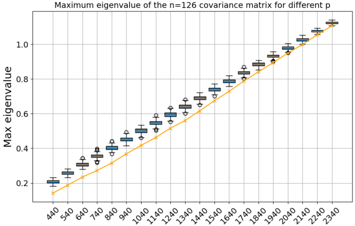
Time period: 2021.1-2021.6



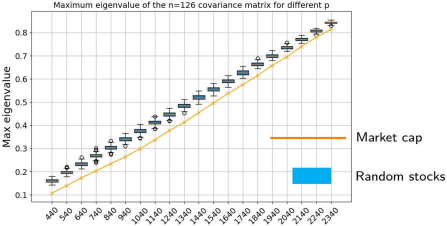
Time period: 2021.7-2021.12



Time period: 2022.1-2022.6



Time period: 2022.7-2022.12



Fit into a one-factor model

- Suppose returns follow a one-factor, homogeneous specific risk model:

$$r = \beta f + \epsilon$$

- β is a p-vector of exposures
- f is the factor return
- ϵ is a p-vector of mean 0 specific returns
- The population covariance matrix of r is given by:

$$\Sigma = \sigma^2 \beta \beta^\top + \delta^2 I$$

- σ and δ are factor and specific volatility and I is the p $\times$ p identity matrix

Fit into a one-factor model

- We draw βs from a normal distribution with mean 1 and standard deviation τ
- β is the leading eigenvector of Σ and the eigenvalue is given by:

$$\lambda^2 = \sigma^2|\beta|^2 + \delta^2 \approx \sigma^2 p(1+\tau^2) + \delta^2$$

- The approximation should improve as p grows
- If we fit this to the previous plot of EOY 2021 we'll have:

$$\sigma^2(1+\tau^2) = Slope * 252 = 0.040$$

$$\delta^2 = Intercept * 252 = 0.160 \quad \delta = 0.399$$

- τ can be calculated from the first eigenvector (normalized to mean 1)

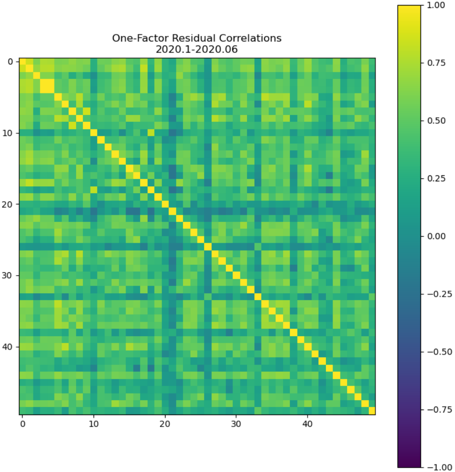
$$\tau^2 = 0.27 \quad \text{Therefore,} \quad \sigma^2 = 0.0317 \quad \sigma = 0.178$$

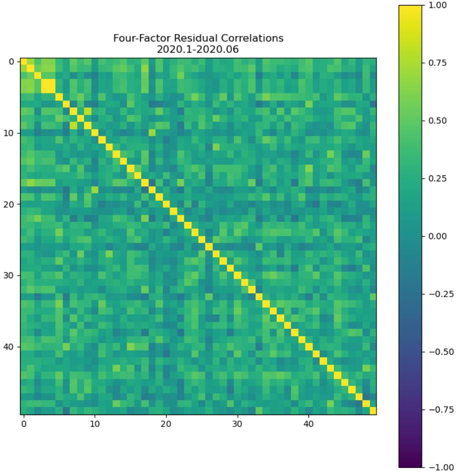
How do σ and δ change over time

Period	σ	δ
2019.1 - 2019.6	0.138	1.778
2019.7 - 2019.12	0.129	2.123
2020.1 - 2020.6	0.545	4.982
2020.7 - 2020.12	0.196	2.008
2021.1 - 2021.6	0.187	1.825
2021.7 - 2021.12	0.178	0.399
2022.1 - 2022.6	0.297	0.269
2022.7 - 2022.12	0.270	0.830

- Fitted σs are in a reasonable range and are close to the ones used in reality
- Fitted δs are larger than expected (due to the hidden factors)
- σ and δ explodes in financial crisis (also true for 2008)

We need more factors





σ and δ implied by a four factor model.

- Consider the four-factor model where we have

$$r = Bf + \epsilon$$

- $B = [\beta_1, \beta_2, \beta_4, \beta_4]$, our matrix of factor exposures
- $f = [f_1, f_2, f_3, f_4]$ our 4-vector of factor returns
- With covariance matrix

$$\Sigma = B\Phi B^\top + \delta^2 I$$

- where Φ is our 4×4 covariance matrix of factor returns.

How do σ and δ change over time (4 factor)

Period	τ^2	σ_1	δ
2019.1 - 2019.6	0.2753	0.1517	1.3722
2019.7 - 2019.12	0.3846	0.1442	1.4655
2020.1 - 2020.6	0.1557	0.5710	2.4866
2020.7 - 2020.12	0.5942	0.2097	1.8080
2021.1 - 2021.6	0.3527	0.1945	1.7617
2021.7 - 2021.12	0.2759	0.1782	1.5867
2022.1 - 2022.6	0.4117	0.2927	1.8245
2022.7 - 2022.12	0.2720	0.2670	1.9680

Table: Four-Factor Model (Factor 1) Parameters by Period

How do σ and δ change over time (4 factor) cont.

Period	σ_1	σ_2	σ_3	σ_4	δ
2019.1 - 2019.6	0.1517	0.0729	0.0652	0.0584	1.3722
2019.7 - 2019.12	0.1442	0.1129	0.0798	0.0775	1.4655
2020.1 - 2020.6	0.5710	0.2160	0.1913	0.1627	2.4866
2020.7 - 2020.12	0.2097	0.1516	0.1280	0.1015	1.8080
2021.1 - 2021.6	0.1945	0.1470	0.1438	0.1140	1.7617
2021.7 - 2021.12	0.1782	0.1102	0.0723	0.0694	1.5867
2022.1 - 2022.6	0.2927	0.1280	0.1052	0.0856	1.8245
2022.7 - 2022.12	0.2670	0.1123	0.0944	0.0873	1.9680

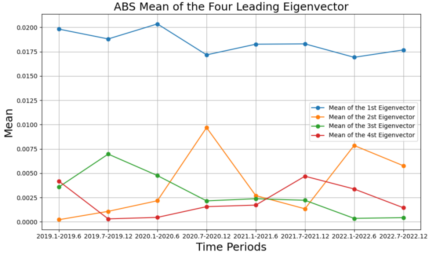
Table: Four-Factor Model: Factor σ Values and δ by Period

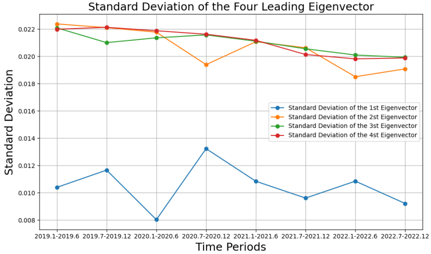
How are the entries of spiked eigenvectors distributed

How are the entries of spiked eigenvectors distributed

- Look at how the mean and variance of spiked eigenvectors change over time (before normalization)
- Normalize the fist factor to mean 1
- Z-score normalize the rest factors

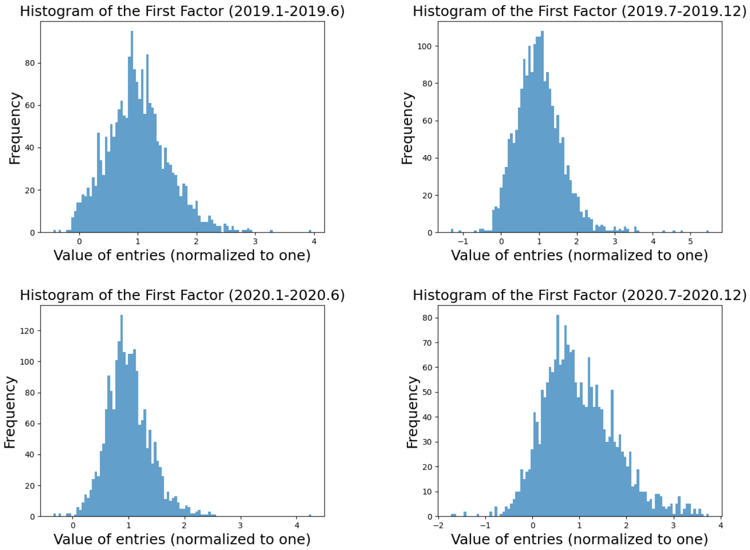
A look at how the mean and variance of spiked eigenvectors change over time (before normalization)



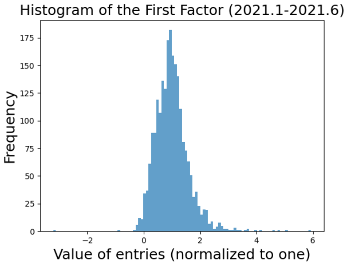


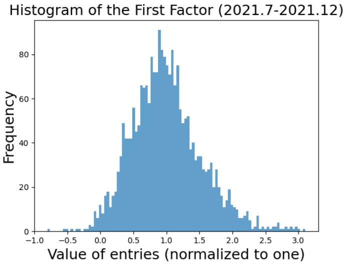
- The first factor has an outstanding mean, while the rest are rather close to 0
- The standard deviation of the first factor is lower and responses more to market changes (e.g. 2020.1 - 2020.6)

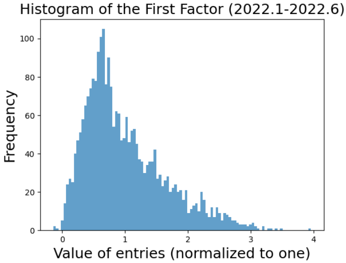
Normalize the first factor to mean 1 (Part I)

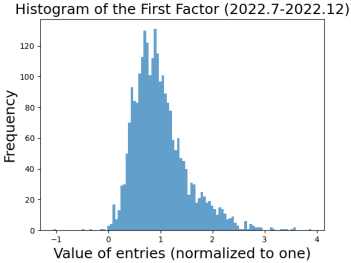


Normalize the first factor to mean 1 (Part II)

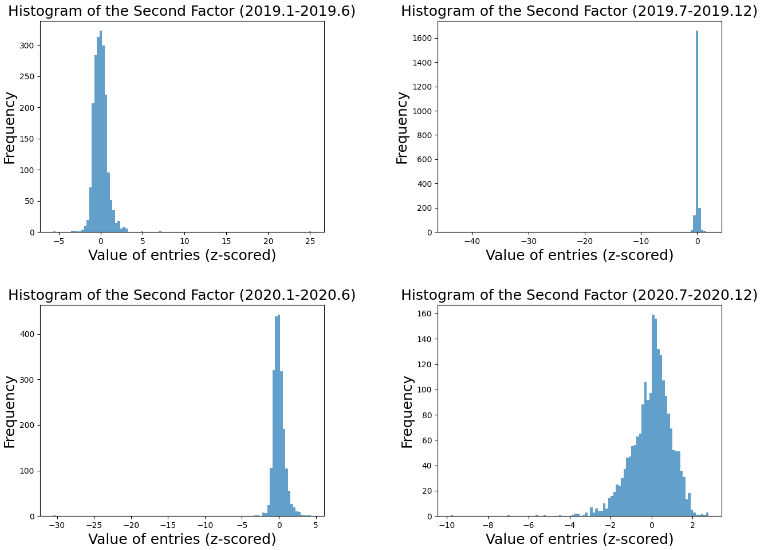




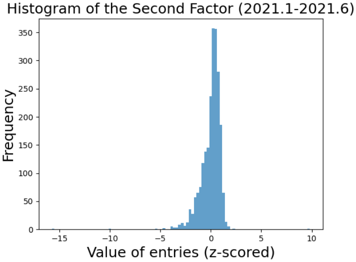


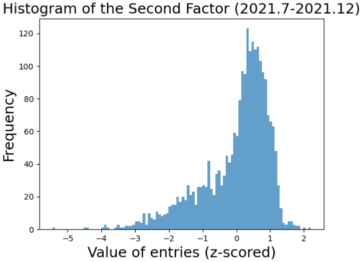


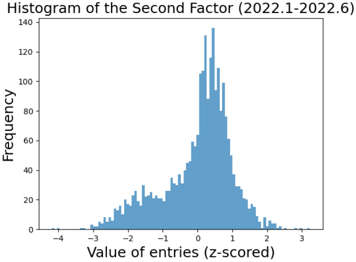
Z-score the second factor (Part I)

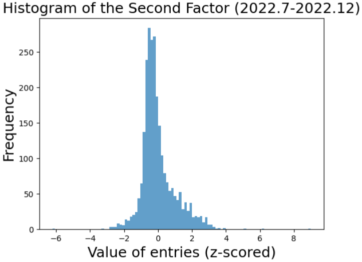


Z-score the second factor (Part II)

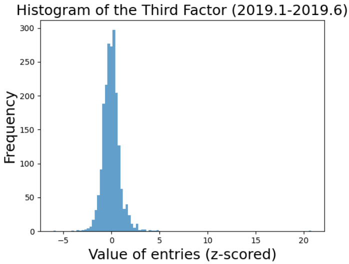


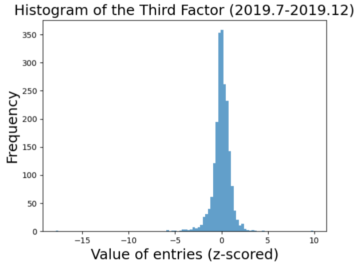


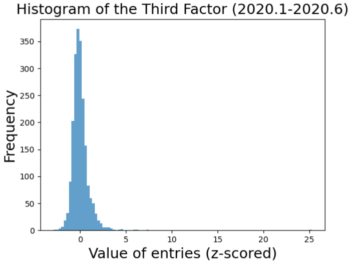


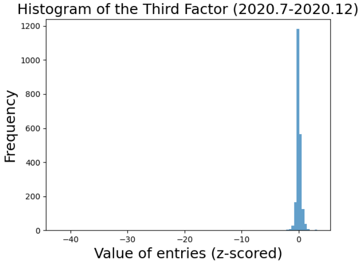


Z-score the third factor (Part I)

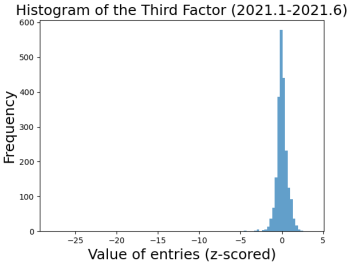


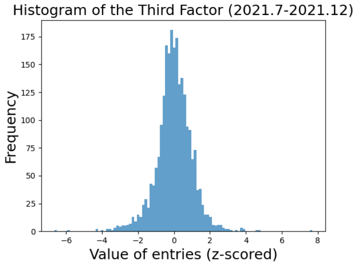


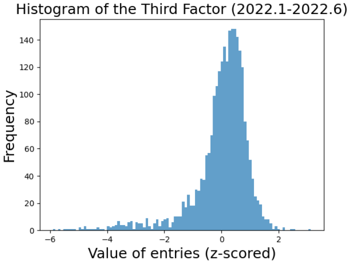


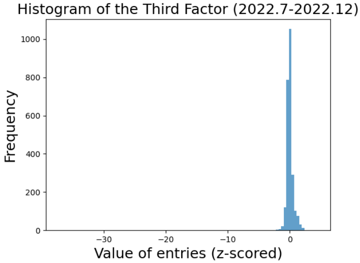


Z-score the third factor (Part II)

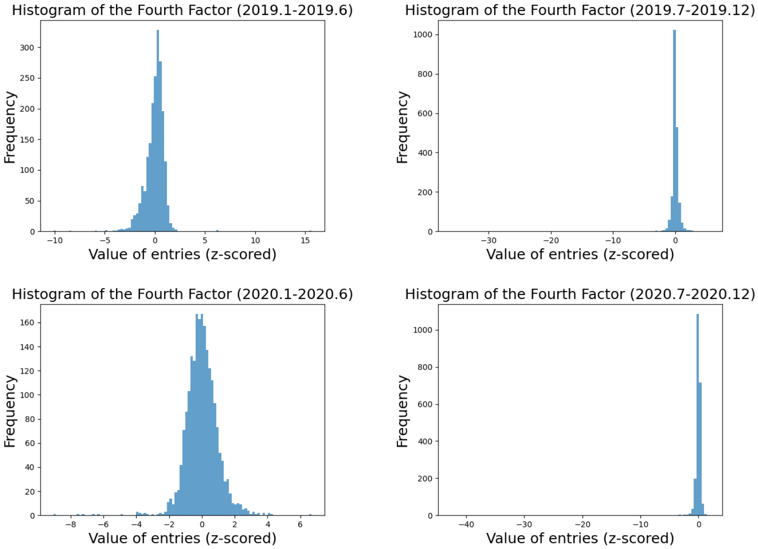




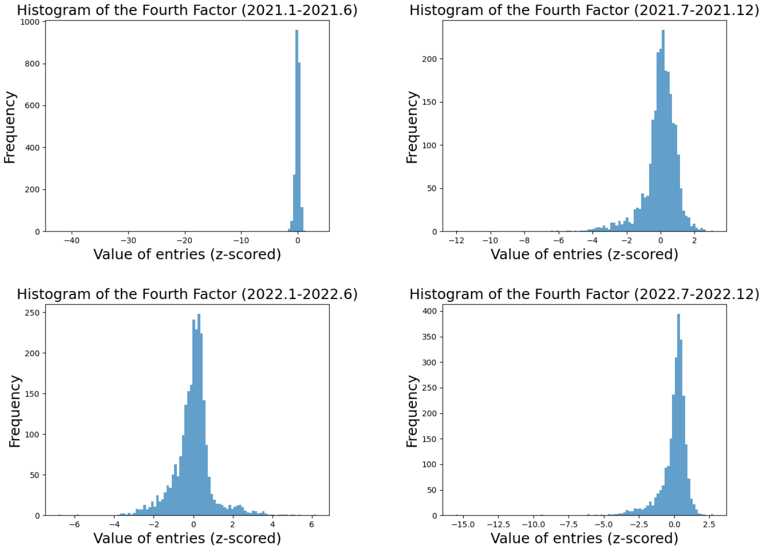




Z-score the fourth factor (Part I)



Z-score the fourth factor (Part II)



Limitations and work in progress

- Further interpreting the histograms of eigenvectors
- Siamak idea: Instead of only looking at how the mean and variance of spiked eigenvectors change over time, can we use ML (say random forest) to make predictions from the histograms? I guess this may entail computing metrics from the histograms and training the ML alg on a bunch of histogram metrics. I'm not sure how much people have done this in this context.
- Include or exclude outliers in the dataset
- Look at the plots on the same horizontal scales and look at the four factor panels for one date at a time
- And much more to do...