TALOS

ANALYSIS

Data-Driven Crypto Risk Factors

April 2022

DATA-DRIVEN CRYPTO RISK FACTORS

Boris Skorodumov and Ilya Kulyatin, Cloudwall* Research

ABSTRACT: Usually, risk factor literature assumes that these factors driving returns are observable (at least theoretically) and are usually obtained by a regression model. An alternative approach is to assume that factors are in fact unobservable and generate these out of pure statistical methodology. One such technique is Principal Component Analysis (PCA), which can be used to generate orthogonal (or independent) factors that drive variability of returns. In this paper, we apply PCA to extract factors out of a top 1000 (by market cap) crypto asset universe, after data cleaning and processing, leaving a net 103 assets studied. We find that the first 10 factors explain about 60% of variability in returns, which is much lower than the usual explain ability with mature markets like equities, possibly suggesting additional factors in the crypto universe or hinting at a yet amateur market. We also see the effect of survival bias at play by observing a notable increase in total variability when we reduce the data coverage constraints. Lastly, we construct eigen portfolios and compare performance of these with an equally weighted benchmark.

INTRODUCTION

In traditional finance we often look at performance from the risk premia perspective, with a large body of literature developed on the topic, from the CAPM draft by Treynor (1961), passing through the Arbitrage Pricing Theory (APT) by Ross (1976), and finally getting to the Factor Investing direction started with Fama and French (1993). **Risk Premia** are also known as discount rates or expected returns. and when we model returns based on these premia we want to explain differences in returns on assets due to their exposure to systematic risk factors and the rewards associated with these factors. In the Risk Factor literature, these factors are treated as observable and they get estimated, most of the time, through some sort of a regression model.

An alternative approach considers these risk factors as latent variables and uses techniques like PCA to simultaneously estimate factor returns and factor exposures, also from asset returns. PCA is a dimensionality reduction algorithm, and its objective is to give us Principal Components, i.e. orthogonal drivers of the variability of returns.

*Cloudwall and the technology behind its Serenity System were acquired by Talos in April 2024.

Orthogonality comes from linear algebra and it is closely related to uncorrelatedness (if *X* and *Y* are uncorrelated, then *X-E[X]* is orthogonal to *Y-E[Y]*). If compared to the Factor approach, one advantage of using something like PCA is the minimal data requirement, as you don't need to come up with proxies for these risks (i.e. it's not model-based). The disadvantage is the lack of interpretability, as we don't know what the extracted Principal Components represent in terms of risk. In the coming future, we will share our Factor Risk model research, but in this piece we'll just start with a simple PCA study, just to get a sense of whether there are any risk factors to be found. For your reference on the limitations of PCA, check THIS out.

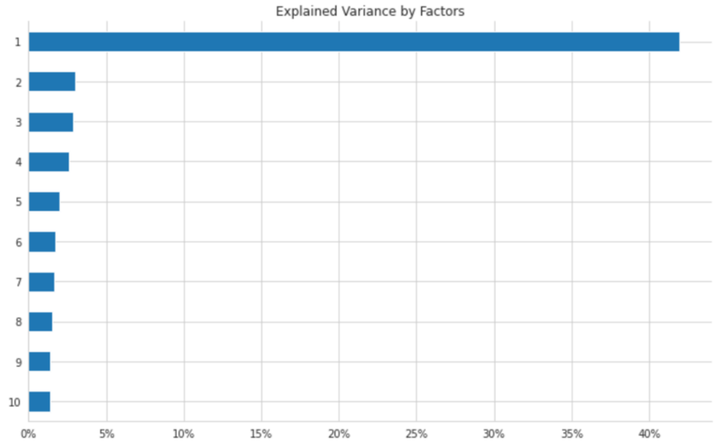
DATA PREPARATION

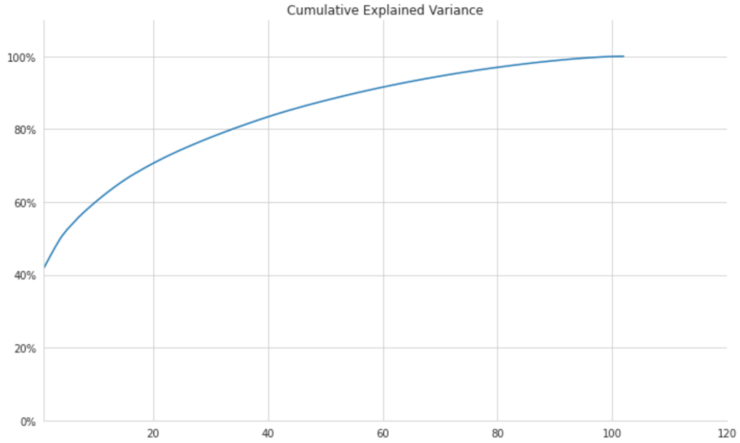
For the sake of reproducibility, we won't use our proprietary clean datasets, but go with whatever is available in public: price and volume data from Coinmarketcap.com (thank you for the free data, CMC!). The frequency is daily, aggregated across several exchanges, see their methodology HERE. Next, we select the top 1000 crypto assets with the largest market capitalization as of 28 April 2022 and get the historical series for the period between 2018 and 2022.

Since in a PCA decomposition we minimize the quadratic norm, this means it's sensitive to outliers, helping them dominate the total norm and thus "driving" the PCA components. To deal with this here we winsorized the daily returns at 2.5% and 97.5%. We've also removed any crypto asset that doesn't have data for at least 95% of the considered time period and any day that doesn't have observations on at least 95% percent of the remaining crypto assets. For any remaining missing values, we input the average return across all crypto assets available on that particular day. At the end of this second cleanup, we have 103 crypto assets and 1492 time periods. That was a bloodbath.

DRIVERS OF RETURNS

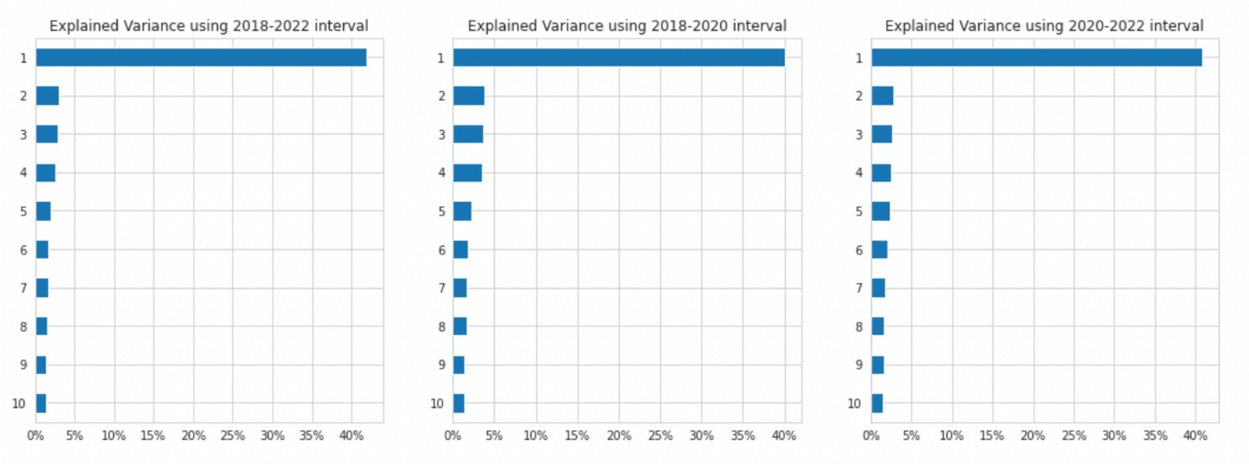
Next, we do the PCA transformation through a Singular Value Decomposition (SVD). Here we find that the most important factor explains around 45% percent of the daily return variations. In analyzing market data, the dominant factor is usually interpreted as the *market factor*, while the remaining factors are usually interpreted as style factors (to be in line with traditional finance). This is not particularly correct, as with Principal Components (PCs) we assume the features have a simple linear relationship and that's rarely the case: take, for instance, size, value, momentum, and carry factors — we can't expect them to be orthogonal! In any case, this serves as a useful indication of the number of potential risk factors. When we look at the cumulative explained variance, we see that 10 factors explain around 60% percent of the returns.





Does that mean that in crypto there are more risk factors to get exposure to? Or maybe the market is not mature enough for a clear set of risk factors driving token performance?

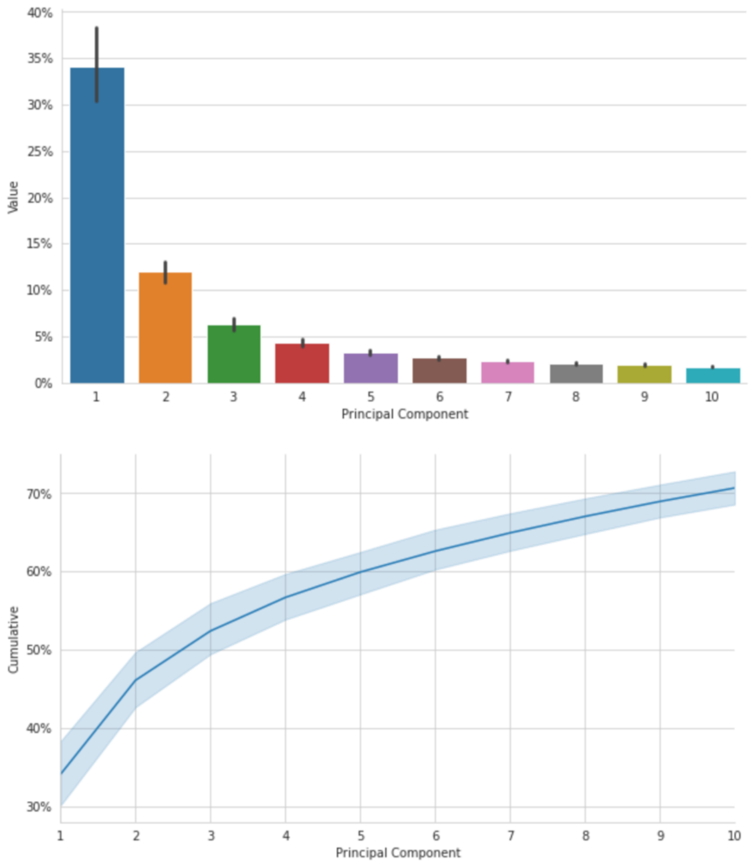
We do notice similar results using three different time-frames: 2018 to 2020, 2020 to 2022, and 2018 to 2022. Interestingly, in the most recent period, it seems like there are even more factors required to explain the same level of returns variability. What’s happening here? To be honest, the PCA might be even breaking here in proper identification of orthogonal components if there’s no pronounced direction in the lower order PCs (is that the case here?).



RANDOM SAMPLING

To torture the data a bit more and get some additional insights we randomly resample 100 tokens over 100 trials from the initial universe of 1000 tokens, relaxing constraints for the data coverage (i.e. we remove tokens and days not having at least 75% of coverage). Within each resampling, we run our PCA algorithm and get the average explained variance by components, as well as the cumulative explained variance. This picture looks interesting, with less “weight” on the first PC, and a clearer decay for the subsequent PCs. Now with 10 PCs, we pass 70% of returns variation explainability. Might show that our survival bias is playing some nasty tricks. We also think there are insights to be gained by grouping tokens, e.g. by sectors. We will be publishing something on that topic in the coming weeks.

Explained Variance of Top 10 Principal Components with 100 Trials



Overall, this analysis suggests it's worth exploring which Risk Factors might be driving this set of tokens. This will be part of our future publications, give us a couple of weeks. Next, let's see what can we do with PCA for portfolio selection.

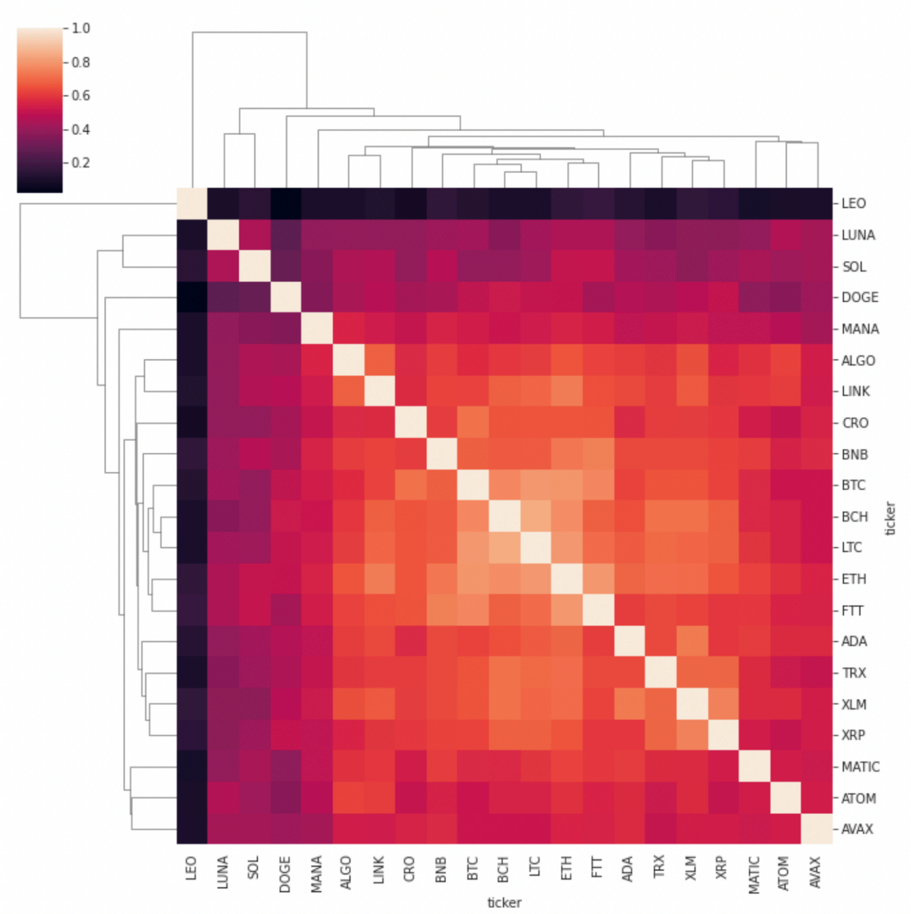
EIGEN PORTFOLIOS

There is a way to connect the concept of minimum risk portfolios and PCA decomposition: **eigenportfolios**. For each eigenvector, we can build a portfolio with weights proportional to the coefficient of each token and inversely proportional to its volatility. These portfolios can also be used as Risk Factors (pro: uncorrelated and only data-based; con: what do they represent?).

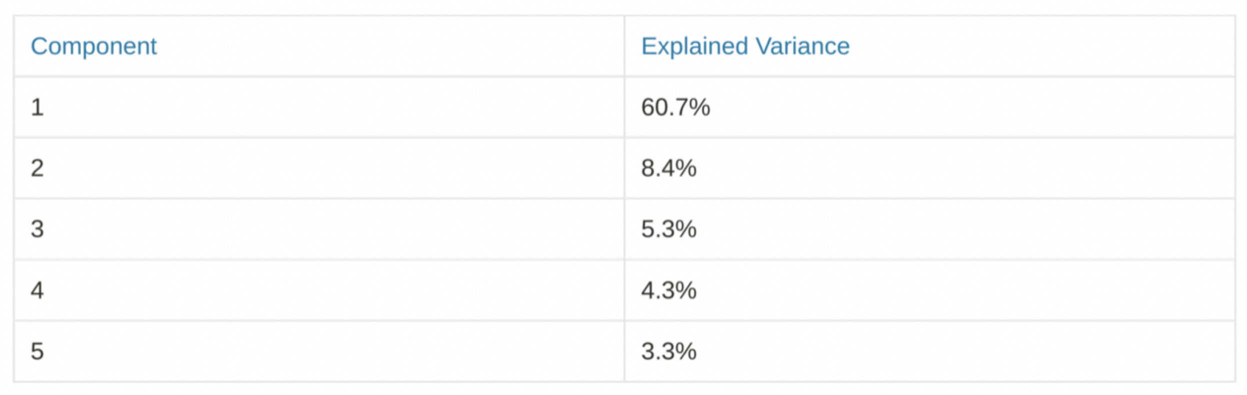
First, we pick the top 30 crypto assets by market capitalization (again, as of 28 April 2022) and remove stablecoins. Second, we calculated daily returns in the following way:

- Winsorize, clean based on token and day coverage and impute missing values just in the same way we did above.
- We also normalize daily returns cross-sectionally for every period.

Next, we get the covariance matrix for the selected tokens and show here the hierarchically-clustered heatmap. Notice the weird behaviour by LEO, way too uncorrelated to look normal, but here we will not explore the reasons behind this. Lesson for the next analysis: remove low-volume tokens.



Now that we have the covariance matrix of the normalized returns, we run PCA and find that the three largest components explain 75% of returns variability, while the first 5 explain 83%. That’s interesting, and we take a note for the next analysis to do the same on token buckets created based on the market cap (small vs mid vs large), given that now we started with just 30 tokens (before the cleaning brought them down to 21).

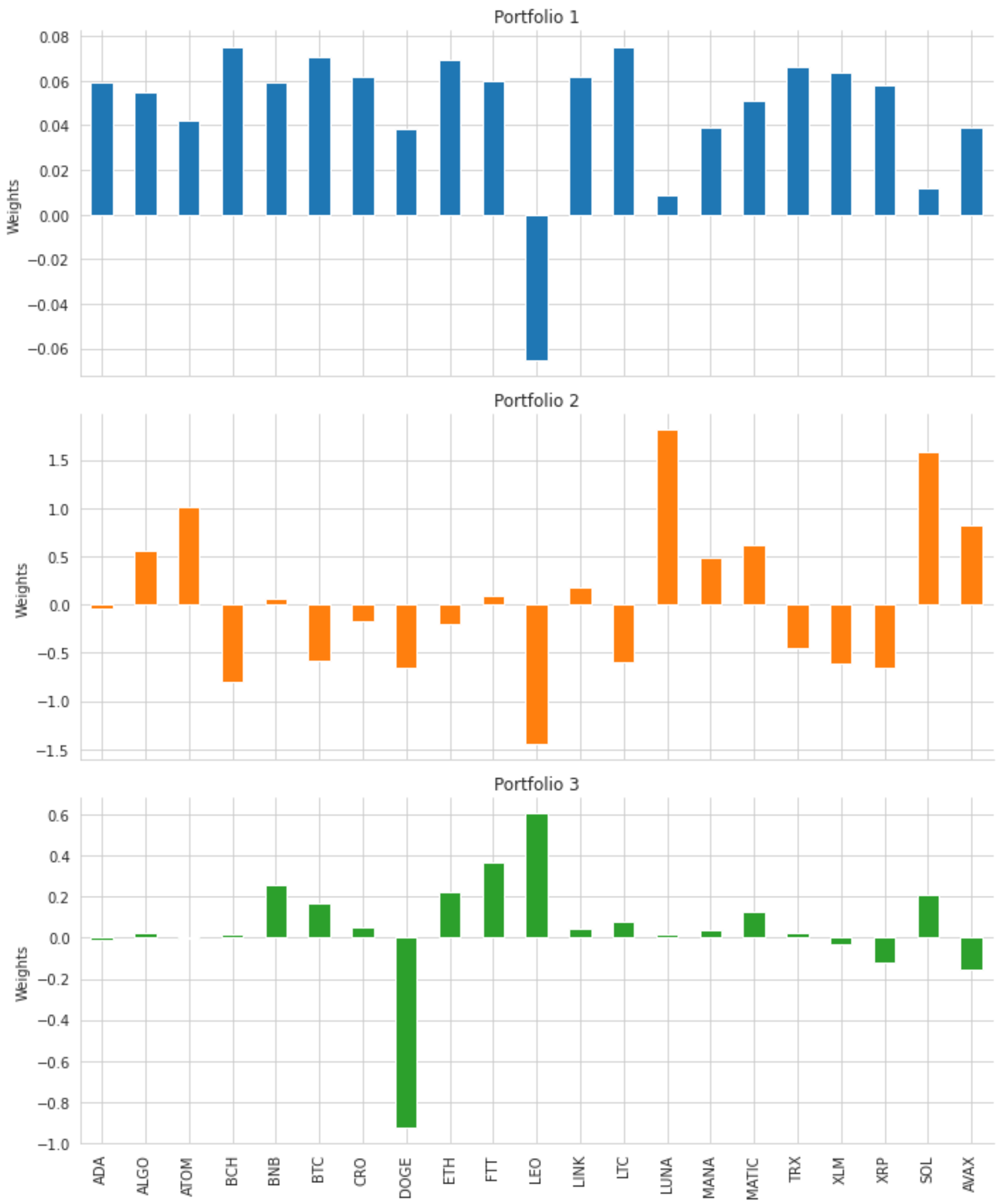


Component	Explained Variance
1	60.7%
2	8.4%
3	5.3%
4	4.3%
5	3.3%

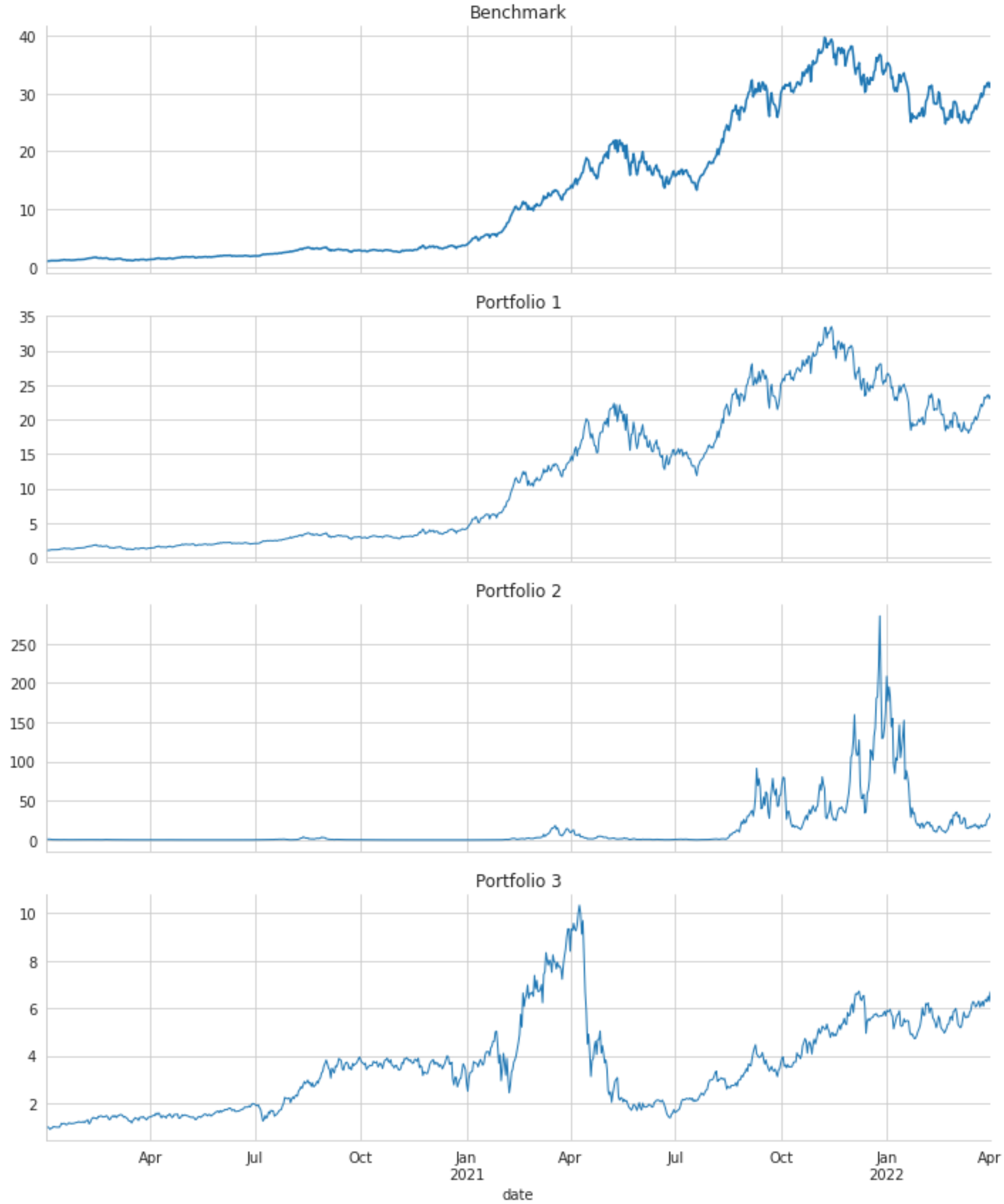
Finally, we select the three largest components and normalize them to sum to 100%. These will be our portfolio weights which we will compare to an equal-weighted portfolio formed using the same tokens.

The weights of portfolios show distinct patterns. For example, in portfolio 1 we have a smoother distribution between components, with all positive weights except LEO. At the same time, the other two portfolios show us much more dispersion in allocations.

We then compare the performance of each portfolio vs the equally weighted benchmark, over the sample period between 2020 and 2022. Here we limit ourselves to visual analysis, and it’s interesting to note how the portfolio based on the first component looks similar to the market portfolio. Does that help us connect the first PC to the market risk factor? Whether that’s the case or not, this 1st PC similarity to the benchmark, but with fewer assets, also tells us we might be able to create cheaper benchmarks, with fewer tokens to rebalance (eigenportfolios can be used to add sparsity and reduce transaction costs). It’s also interesting that the other two portfolios show their distinct behavior, maybe suggesting there’s space for active portfolio management.



For the quant portfolio managers out there, the next step would be to create eigenportfolios optimized on metrics such as the Sharpe ratio. Again, this is not investment advice, we are just playing with some basic linear algebra, and we don't even show you the performance metrics. The objective of this type of analysis in Cloudwall is just to make sure the risk, valuation, and optimization models we are building for our users are sound not only theoretically, but also empirically.



Overall, we see that PCA might be useful in 1) identifying some components driving token returns and 2) in giving us an interesting covariance matrix to use in portfolio construction. In general, our experience tells us there are some similarities with PCA analysis in Equities, where the first PC is also covering the majority of the explainability. At the same time, in Equities we usually see only the first 3 PCs explaining +95% of the variability of returns.

EDITS

- Our friend Joel Guglietta on LinkedIn mentioned it could be more interesting to look at ICA vs PCA. We usually think about PCA vs ICA as compression vs (independent) separation, with the latter more dependent on "composed" Gaussianity (i.e. it works best if our features are Gaussian, but composed of non-Gaussian underlying data-generating processes). Even though the PCA decomposition doesn't require Gaussianity, it's still based on Pearson correlations, so it benefits from a symmetric distribution. We did a bit of cleaning with winsorization, but I'm sure we are not that close to Gaussianity. For ICA we do require non-Gaussian sources, and it works if the features are Gaussian (through a mixture of non-Gaussians). In other words, it's similar to a reverse of the Central Limit Theorem. In our previous post (on the stylized facts) we have indeed noticed that grouping several tokens gets us closer to Gaussianity, so ICA might indeed be the way to go, especially if we carry out this analysis on sectoral indices. Some more goodies on the topic in THIS paper. Thanks for the insight, Joel!
- Talking about winsorization, several readers mentioned it's a dangerous technique, as it might be removing useful information. Ioan suggested going with Sparse PCA in our experiment. He also shared interesting recent developments in Autoencoders, such as Beta-TCVAE, Factor-VAE, PCAAE, and the "uncertainty autoencoder").

- By using the Pearson correlation coefficient, we test for linear relationships, while most likely there are non-linear effects dominating asset returns. A way to deal with that is to use rank-based correlation methods such as Kendall Tau or Spearman correlation. This is also known to provide covariance matrices more robust to outliers.
- As Ben Steiner suggested, we should also look at how stable are the eigenvectors over time. And if they are not stable, when do we recalibrate these models? This is directly related to the problem of concept drift. Ben also suggested using "real" risk factors (what we've mentioned in the first paragraph of this post), while applying PCA to the residuals we can't explain.
- On the stability of PCA results, Giorgio Borelli referred us to THIS literature review paper, especially highlighting the need for PCA based on iterative procedures such as in THIS Bilokon and Finlestein (2021) paper (authors also made available their library at THIS link).
- Someone asked about ways to de-noise correlation matrices. A simple way to check for correlation matrix stability is to use a rolling analysis, but our dataset is too short for this to give any insight. I'm almost sure these correlations are not stable, but the last part of our analysis does PCA on the correlation matrix to de-noise it. In general, from the analysis of crypto stylized facts, we might be able to get a hint of what to use to get a better-behaved correlation matrix. From our previous post, we can think maybe to use a multivariate GARCH to get the conditional correlation matrix, it should be better-behaved if our dataset shows conditional heavy tails (and it does!). There are other methods that could be better than GARCH, like the shrunk Wishart stochastic volatility (SWSV) model and probably many others.

AUTHORS

- Boris Skorodumov, Quantitative Researcher @ Cloudwall Capital.
- Ilya Kulyatin, Head of Research @ Cloudwall Capital.

DISCLAIMER

Not a financial advice, solicitation, or sale of any investment product. *The information provided to you is for illustrative purposes and is not binding on Cloudwall Capital. This does not constitute financial advice or form any recommendation, or solicitation to purchase any financial product. The information should not be relied upon as a replacement from your financial advisor. You should seek advice from your independent financial advisor at all times. We do not assume any fiduciary responsibility or liability for any consequences financial or otherwise arising from the reliance on such information.*

You may view this for information purposes only. Copy, distribution, or reproduction of all or any portion of this article without explicit written consent from Cloudwall is not allowed.

[*] *Cloudwall and the technology behind its Serenity System were acquired by Talos in April 2024.*

talos.com

Disclaimer: Talos Global, Inc. and its affiliates ("Talos") offer software-as-a-service products that provide connectivity tools for institutional clients. Talos does not provide clients with any pre-negotiated arrangements with liquidity providers or other parties. Clients are required to independently negotiate arrangements with liquidity providers and other parties bilaterally. Talos is not party to any of these arrangements. Services and venues may not be available in all jurisdictions. For information about which services are available in your jurisdiction, please reach out to your sales representative. Talos is not is not an investment advisor or broker/dealer. This document and information do not constitute an offer to buy or sell, or a promotion or recommendation of, any digital asset, security, derivative, commodity, financial instrument or product or trading strategy. This document and information are not intended to constitute investment advice or a recommendation to make (or refrain from making) any kind of investment decision and may not be relied on as such. This document and information are subject to change without notice. It is provided only for general informational, illustrative, and/or marketing purposes, or in connection with exploratory conversations with institutional investors and is not intended for retail clients. The information provided was obtained from sources believed to be reliable at the time of preparation, however Talos makes no representation as to its accuracy, suitability, non-infringement of third-party rights, or otherwise. Talos disclaims all liability, expenses, or costs arising from or connected with the information provided.