Statistical Arbitrage within Crypto Markets using PCA

Isekai Quant

February 2025

Abstract

This paper explores the application of principal component analysis (PCA) in statistical arbitrage trading. The methodology involves constructing eigenportfolios, estimating the Ornstein-Uhlenbeck process for residual mean reversion, and implementing a walk-forward validation framework. Performance metrics indicate that the approach may not be robust in the cryptocurrency market, but room for improvement definitely exists.

Contents

1 Introduction

The main objective of this study is to focus on the implementation details of the pairs trading strategy presented by Avellaneda (2008), discussing the nuances that arise in practical use. Although there are several studies extending this framework to ETFs, almost none explore eigenportfolio-based factor proxies in depth. One reason for this gap may be the additional complexity eigenportfolios introduce, making their interpretation and execution less straightforward. This paper attempts to clarify those concepts while simultaneously applying the strategy to cryptocurrency markets, a relatively new and evolving area.

Before delving into the core methodology, we offer an introductory overview of certain concepts and terminologies relevant to this discussion.

1.1 Market Neutral Strategies

Market neutral strategies aim to neutralize exposure to risk factors within a portfolio. This is typically desired by portfolio managers who may or may not wish to take active views on particular factors. By neutralizing unwanted risks, managers can isolate and capture only the "alpha," or idiosyncratic returns, and avoid exposure to broader market factors they do not wish to bet on.

In practice, market neutral strategies are often implemented either through (1) dollar neutrality or (2) beta neutrality. Dollar neutrality means taking equal dollar amounts in long and short positions. Beta neutrality, on the other hand, neutralizes exposure to specific market factors. While beta neutrality theoretically aligns better with the goal of isolating idiosyncratic risk, in practice it can require awkward or concentrated exposures. For instance, beta neutrality might force one to hold a large short position in a high-beta security while balancing it with multiple long positions in lower-beta securities, thereby creating a different set of concentration risks.

In this paper, we adopt the perspective of beta neutrality, focusing on neutralizing exposure to certain factors:

$$\sum_i Q_i \beta_{ij} = 0 \quad \forall j \in \{1, 2, \ldots, q\},$$

where Q_i is the dollar exposure in asset i, and β_{ij} is the loading of asset i on factor j. The total number of factors is q.

1.2 Statistical Arbitrage

Statistical arbitrage is a category of market neutral strategies, where statistical relationships of constituents of a basket are exploited to make idiosyncratic bets on mean reversion of relative performances. Another important feature of this type of strategy would be that the bulk of returns should be factor neutral. This paper focuses on a specific subset of statistical arbitrage known as pairs trading. Pairs trading is simpler since we restrict ourself to the price / returns space when looking for statistical relationships to exploit among pairs.

1.3 Objective of Study and Value Added

We delve into the practical side of implementing the pairs trading method laid out by Avellaneda (2008). While multiple studies exist that apply this approach to ETFs, few explore the intricacies of using eigenportfolios as factor proxies. As noted above, eigenportfolios typically demand additional care in their interpretation and execution. This paper clarifies those challenges and applies the methodology to crypto markets, which are both novel and rapidly evolving. Our hope is that this contribution helps researchers and practitioners gain deeper insight into the strategy's performance and associated complexities.

2 Literature Survey

We begin with a concise overview of several standard pairs trading strategies commonly referenced in literature. Although these approaches do not necessarily represent the entire space of possible strategies or the most up-to-date real-world practices, they serve as a valuable point of reference for those beginning with pairs trading.

2.1 Distance Based Approach

The distance based approach, introduced by Gatev et al. (2006), is both intuitive and straightforward. It identifies pairs of assets whose sum of squared differences (SSD) over a formation period is minimal, implying close historical movements. Practitioners typically use Euclidean distance, though Manhattan or Chebyshev distances are also possible. Although once successful, this method reportedly became less effective after the early 2000s, likely due to increased popularity and market efficiency.

2.2 Cointegration Approach

First outlined by Vidyamurthy (2004), the cointegration approach checks whether two (log) price series P_i and P_j, though individually $I(1)$, can be linearly combined (e.g., $P_i - \beta P_j$) to form an $I(0)$ stationary series. One commonly verifies this using a Cointegrated Augmented Dickey-Fuller (CADF) test.

A notable subtlety is that if you regress the first variable on the second to estimate β, it is not simply the inverse of the coefficient from the reversed regression. Thus, practical cointegration strategies often run regressions in both directions, selecting the one that yields the more negative ADF test statistic.

Cointegration can also be extended beyond pairs to baskets of assets. For instance, the Johansen procedure allows you to identify one or more cointegrating relationships among multiple time series. In this approach, one estimates a VAR(p) system:

$$\mathbf{y}_t = \boldsymbol{\mu} + A_1\mathbf{y}_{t-1} + A_2\mathbf{y}_{t-2} + \cdots + A_p\mathbf{y}_{t-p} + \boldsymbol{\epsilon}_t \tag{2.1}$$

- $\mathbf{y}_t \in \mathbb{R}^{d\times 1}$: Vector of endogenous variables at time t
- $\boldsymbol{\mu} \in \mathbb{R}^{d\times 1}$: Intercept (drift) vector
- $A_i \in \mathbb{R}^{d\times d}$: Coefficient matrices for lag i
- $\boldsymbol{\epsilon}_t \in \mathbb{R}^{d\times 1}$: White noise error term

Applying the difference operator Δ and with some algebra we can get the error correction form (VECM).

$$\Delta\mathbf{y}_t = \boldsymbol{\mu} + \Pi\mathbf{y}_{t-1} + \sum_{i=1}^{p-1}\Gamma_i\Delta\mathbf{y}_{t-i} + \boldsymbol{\epsilon}_t \tag{2.2}$$

$$\begin{aligned}\Pi &= -(I - A_1 - A_2 - \cdots - A_p)\\ \Gamma_i &= -(A_{i+1} + \cdots + A_p)\end{aligned}$$

The matrix Π is called the *error correction matrix* and its rank tells us the number of cointegrating relationships that we can form from linearly combining the d variables. If a cointegrating relationship exists, we would like to see $0 < rank(\Pi) = r < d$.

To determine the rank of Π, the Johansen cointegration test is employed. This test involves sequentially evaluating the null hypothesis:

$$H_0 : \text{rank}(\Pi) \leq c \quad \text{for} \quad c = 0, 1, \ldots, d-1$$

against the alternative hypothesis $H_1 : \text{rank}(\Pi) > c$. The Johansen test can be carried out with either the eigen statistic or the trace statistic. Once one verifies that the rank is in between *0* and d, then we decompose the error correction matrix as below:

$$\Pi = \alpha\beta^\top \tag{2.3}$$

- $\alpha \in \mathbb{R}^{d\times r}$: Loading matrix (adjustment coefficients)
- $\beta^\top \in \mathbb{R}^{r\times d}$: Cointegrating matrix (long-run equilibrium relationships)
- r: Rank of Π, indicating the number of cointegrating relationships

The columns of matrix β contain the hedge ratios that produce the stationary portfolio. Assuming the columns are arranged in descending order of eigenvalues, we take the first column vector to form our stationary portfolio as it is generally considered "the most stationary" portfolio (Chan, 2013).

2.3 Stochastic Residual Approach

Note that the aforementioned strategies work with *price spreads*. However, one may argue that mean reversion is not only restricted to long-run price levels but also applies to the relative performance of two securities. In other words, valid mean reversion strategies can be formulated by studying the *return spreads*.

Do et al. (2006) were among the first researchers to formalize the theory of mean reversion in return spreads. They proposed adopting a state-space model of the return spread, where the true spread is treated as a latent state governed by a mean-reverting process such as the Ornstein-Uhlenbeck (OU) process, while the observed mispricing contains Gaussian white noise.

An interesting aspect of the method suggested by Do et al. (2006) is their attempt to integrate the Arbitrage Pricing Theory (APT) framework to isolate the residual, which serves as a proxy for mispricing or the disequilibrium component. We illustrate this with a simple example. Consider the return series of stock A and stock B, modeled using a one-factor model (market):

$$
\begin{aligned}
R^A &= R_f + \beta_A R^M + e_A \\
R^B &= R_f + \beta_B R^M + e_B
\end{aligned}
$$

- R^A, R^B: Returns of stocks A and B
- R_f: Risk-free rate
- β_A, β_B: Beta coefficients for stocks A and B, measuring sensitivity to the market factor
- R^M: Excess return of the market portfolio over the risk-free rate
- e_A, e_B: Error terms (idiosyncratic risks) associated with assets A and B, representing the part of the returns not explained by the market

One can then define the *relative APT* as:

$$R^A = R^B + \Gamma R^M + e \tag{2.4}$$

where $\Gamma = \beta_A - \beta_B$ is the vector of exposure differentials, and e is the residual noise term.

The authors further define the residual spread using the residual spread function G_t, given by:

$$G(R_t^A, R_t^B, U_t) = R_t^A - R_t^B - \Gamma R_t^M \tag{2.5}$$

Another point worth mentioning is that while the authors did not present a specific trading strategy based on their model, they acknowledged that the execution of a mean reversion strategy should depend on the *accumulation of residual spread.* That is, once a persistent level of relative mispricing is observed over consecutive periods, one can open a position to bet on the mean reversion of relative performance back to equilibrium. This concept will be revisited briefly in the next section.

For details on the estimation procedure, please refer to the original paper by Do et al. (2006).

2.4 Avellaneda and Lee (2008)

Avellaneda and Lee (2008) also employ a variant of the stochastic residual spread method. Specifically, they develop two core approaches:

- **Principal Component Analysis (PCA)-Based Strategies:** Using PCA to extract systematic risk factors from the correlation matrix of returns.
- **ETF-Based Strategies:** Using sector ETFs as proxies for systematic factors.

Both strategies aim to isolate the idiosyncratic (residual) components of stock returns, which are assumed to follow mean-reverting processes. Trading signals are generated when these residuals deviate significantly from their equilibrium levels, indicating potential contrarian trading opportunities.

Unlike Do et al. (2006), Avellaneda and Lee suggest using ETFs or statistical factors to isolate the *residual return.* Moreover, they introduce what they call an *auxiliary series*, defined as the cumulative sum of the residual returns:

$$X_i(t) = X_{t,i} = \sum_{s=1}^{t} \epsilon_{s,i} \tag{2.6}$$

for asset i at time t. They proceed to model this *auxiliary series* using an Ornstein-Uhlenbeck (OU) process. This differs from the approach of Do et al. (2006), where the *residual returns* themselves are modeled as OU processes.

Trading signals are generated based on the *s-score*, a dimensionless variable defined by the authors as:

$$s_{t,i} = \frac{X_i(t) - m_i}{\sigma_{\text{eq,i}}} \tag{2.7}$$

where m_i represents the long-run mean of auxiliary series for coin i, and $\sigma_{\text{eq,i}}$ is the equilibrium standard deviation of the auxiliary series. Intuitively, the s-score measures how far the auxiliary series has deviated from its long-run mean in terms of standard deviations. The authors use this measure to construct simple threshold-based trading rules, as indicated below. The subscripts denote buy-to-open (bo), sell-to-close (sc), sell-to-open (so), and buy-to-close (bc) signals, respectively:

$$s_{bo} = -1.25, \quad s_{sc} = -0.5, \quad s_{so} = 1.25, \quad s_{bc} = 0.5$$

Rationale of strategy is straightforward; if the dimensionless s-score deviates from the long run mean m_i, we either short or long the coin i and long or short the corresponding eigenportfolios by their respective hedge ratios. We close the position once the s-score reverts back to the long-run mean.

Performance results suggest that the strategy generated non-trivial excess returns over the simulation period, regardless of whether ETFs or statistical factors were used.

However, the more recent performance of the strategy, particularly when applied to eigenportfolios, has not been well-documented. While some studies report strong in-sample performance, we believe that such results do not accurately reflect the true efficacy of the strategy. Therefore, our aim is to provide a robust evaluation of this strategy using more rigorous methodologies.

2.4.1 Application in Crypto

Furthermore, this paper contributes to the literature by documenting the performance of the strategy in the cryptocurrency markets. As a relatively nascent asset class, cryptocurrencies are characterized by high volatility and market inefficiencies, which create abundant trading opportunities compared to the more mature public equity markets. As we will see, both the methodology and the results are heavily influenced by the inherent characteristics of cryptocurrencies, yielding unexpected yet interesting findings.

3 Methodology

In this section, we outline the specific methodology used to implement the mean reversion strategy proposed by Avellaneda (2008). We also highlight key differences between our approach and the original, which arise primarily from variations in market condition and experimental choice.

3.1 Dynamic Trading Universe

Avellaneda's original paper (2008) sets the trading universe to be the S&P 500, which makes intuitive sense because the index has consistently captured a large share (at least 60%) of the U.S. equity market. However, for crypto assets, the situation is more nuanced. The cryptocurrency market is highly concentrated around a few major coins (notably Bitcoin and Ethereum), with Bitcoin alone capturing 40–60% of total market capitalization. In many instances, the top ten coins collectively represent 60–80% of the market cap at any given time.

Unlike the relatively stable composition of the S&P 500, the top coins in cryptocurrency can change rapidly. A coin that ranked among the top ten last month may be displaced within weeks, and new listings can quickly ascend to a top-ten position if they attract sufficient inflows. This poses a challenge: our trading universe should be representative of current market conditions (so that statistical factors remain relevant) but should not be updated so frequently as to incur excessive rebalancing costs.

With this in mind, we developed a daily updating algorithm that seeks a balance between capturing market reality and minimizing turnover. The following subsections detail the filtering steps employed.

3.1.1 Special Coins

Certain cryptocurrencies neither reflect genuine market exposure nor exhibit typical tradable behavior. Two prime examples are:

- **Stablecoins**, whose prices are theoretically pegged to a fiat currency, so they do not represent "true market dynamics".
- **Wrapped tokens** or liquid staking derivatives, which simply mirror an underlying coin. Including both the original and its wrapped version would double-count the same exposure.

We therefore exclude any coin identified as stable or wrapped (or similarly pegged). This step makes sure that we only include coins that truly reflect market consensus at any point of time.

3.1.2 Exponential Weighted Moving Average and Market Cap

To ensure our universe reflects actively traded coins, we adopt a market-cap filter guided by an exponentially weighted moving average (EWMA). The EWMA smooths the market cap time series to reduce the impact of short-term fluctuations while still capturing long-term trends. Mathematically, the EWMA for the market capitalization of coin i at time t is defined as:

$$\mathrm{EWMA}_{t,i} = \lambda \cdot M_{t,i} + (1 - \lambda) \cdot \mathrm{EWMA}_{t-1,i}$$

where:

- $M_{t,i}$ is the raw market capitalization of coin i at time t,
- $\lambda \in (0, 1)$ is the smoothing parameter (decay factor), with higher values giving more weight to recent observations,
- $\mathrm{EWMA}_{t-1,i}$ is the EWMA of coin i from the previous time step.

The filtering process proceeds as follows:

1. Collect daily historical market capitalizations for each coin.
2. Apply EWMA smoothing to each coin's raw market cap time series using a chosen decay factor λ.
3. On each date t, sum these EWMA'd market caps across all coins, then rank the coins in descending order of EWMA. A coin receives a "score" of 1 if it lies within, for example, the top 80% of total EWMA market cap; otherwise, it is scored 0.
4. The dynamic universe on date t includes any coin scored 1. Thus, we capture the top fraction of market cap without fixing a strict rank (e.g. top 50).

In addition, we impose the constraint to only include a coin in our trading universe if a coin has achieved a score of 1 for x consecutive days. We apply a similar logic for excluding a certain coin from our universe. This approach reduces turnover in the chosen set, while still adapting to major shifts in coin valuations over time.

3.1.3 Median Trading Volume

Although market cap often correlates with liquidity, we also rely on trading volume to ensure coins are sufficiently active. The simplest approach is to require that each coin's 30-day median volume surpass a specified threshold. Alternatively, one might replicate the above "EWMA and score" approach using volume data. In both cases, the goal is to exclude coins with negligible daily turnover that might generate liquidity risks or excessive slippage in actual trading.

3.1.4 Negative Shock

Occasionally, coins experience catastrophic events (e.g. FTX's collapse) or show deep, sudden crashes that may not be reflected solely by the market cap filters. We therefore incorporate an additional "negative shock" check by examining an EWMA of each coin's closing price:

1. For the 30 days preceding the current trading date, compute the EWMA of the coin's close.
2. Measure the percentage change in this EWMA across that 30-day window.
3. If the decline exceeds a chosen threshold (e.g. 70%), label it as having suffered a severe shock and exclude it.

This step helps us avoid maintaining positions in assets whose fundamental viability may have been irreparably damaged, thereby improving the reliability of the mean reversion signals.

Note that the choice of the threshold (70%) is arbitrary. One may even choose to adopt a *relative shock* filter, which would filter out coins whose relative performance within a certain time period stands out among peers. We have chosen to adopt the absolute performance threshold for simplicity.

On each day t, we would apply all the filters above to obtain our trading universe on day t, denoted as TU_t. The resulting daily universe adapts to market dynamics yet avoids coins with minimal liquidity, questionable data integrity, or fundamental collapses. This set of coins then forms the basis for PCA factor extraction and mean-reversion analysis, which we detail in the following sections.

3.2 Statistical Factors via PCA

Suppose we have p assets and n time periods. Let

$$\mathbf{R} \in \mathbb{R}^{n \times p}$$

denote the matrix of returns: each of its p columns contains the n historical returns of an asset. We write

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_1 & \mathbf{R}_2 & \cdots & \mathbf{R}_p \end{bmatrix},$$

where each $\mathbf{R}_i \in \mathbb{R}^n$ is the column-vector of returns for asset i (over n periods).

Then, denote by μ_i the sample mean return of asset i and by σ_i the sample standard deviation. Define the standardized return matrix $\tilde{\mathbf{R}}$ by

$$\tilde{\mathbf{R}}_i \;=\; \frac{\mathbf{R}_i \;-\; \mu_i}{\sigma_i}, \quad i = 1, \ldots, p$$

Hence, each column of $\tilde{\mathbf{R}}$ has mean 0 and standard deviation 1.

To extract statistical factors, we let

$$\mathbf{C} \;=\; \mathrm{corr}(\mathbf{R}) \;=\; \mathbf{V}\,\Sigma\,\mathbf{V}^T$$

be the (estimated) $p \times p$ correlation matrix of the returns.[1] Here, $\mathbf{V}$ is the matrix whose columns are eigenvectors of $\mathbf{C}$, and Σ is the diagonal matrix of eigenvalues (in descending order). We pick only the top $q \ll p$ eigenvectors:

$$\tilde{\mathbf{V}} \;\in\; \mathbb{R}^{p \times q},$$

where each column of $\tilde{\mathbf{V}}$ is one of the first q eigenvectors, corresponding to the largest q eigenvalues.

Mathematically, these q new eigenvectors are known as the *principal axes*, and they represent new directions in which the variance of our data is maximized. Moreover, each axis is *orthogonal* to each other, implying that the information retained along the direction of each axis is original.

From a returns perspective, each eigenvector can be viewed as a *portfolio weight* vector (an "eigenportfolio"), and projecting the returns onto these directions gives us the principal components (PC) or *factor returns.* However, before performing the matrix multiplication, we follow the convention of Avellaneda (2008) to divide each element of the eigenvector by the standard deviation of the asset's return. That is:

$$Q_i^{(j)} = \frac{v_i^{(j)}}{\sigma_i}$$

The left hand side refers to the dollars invested in asset i of eigenportfolio j.

Since we have used the correlation matrix to extract statistical factors, our eigenportfolio weights do not take into consideration the variance of each asset; thus, dividing by volatility ensures that our eigenportfolio is not too concentrated in a high-vol asset.

[1] The reason we eigendecompose the correlation matrix and not the covariance matrix is that we don't want few volatile coins to deterimine the principal axes. Intuitively, suppose a very volatile small-cap coin was included in our trading universe. Then, the generated principal component would indicate that the systematic factor component captured by the returns of this volatile coin is more important than that of other coins. Clearly, this is not what we want.

A subtle but non-trivial result of this normalization process is that true orthogonality is not guaranteed, since we are not multiplying the normalizing constant to the *column* vector but rather to the *row* vector of $\tilde{\mathbf{V}}$. In other words:

$$\left\langle \frac{v^{(i)}}{\sigma_i}, \frac{v^{(j)}}{\sigma_j} \right\rangle \;=\; \frac{1}{\sigma_i \sigma_j} \langle v^{(i)},\, v^{(j)} \rangle \;=\; 0 \quad \forall\, i \neq j$$

is *not* what is happening behind the scenes, but rather:

$$\left\langle \begin{bmatrix} \frac{v_1^{(i)}}{\sigma_1} \\ \frac{v_2^{(i)}}{\sigma_2} \\ \vdots \\ \frac{v_p^{(i)}}{\sigma_p} \end{bmatrix}, \begin{bmatrix} \frac{v_1^{(j)}}{\sigma_1} \\ \frac{v_2^{(j)}}{\sigma_2} \\ \vdots \\ \frac{v_p^{(j)}}{\sigma_p} \end{bmatrix} \right\rangle \;\neq\; 0$$

However, in practice, empirical correlation between these volatility-weighted eigenportfolios is close to zero, so we can safely assume that orthogonality is preserved.

To get the actual returns of the eigenportfolios, we define:

$$\mathbf{F} \;=\; \mathbf{R}\,\mathbf{Q} \;\in\; \mathbb{R}^{n \times q}$$

Row t of $\mathbf{F}$ (i.e. $\mathbf{F}_{t,\cdot}$) then gives the q-dimensional vector of factor returns on day t and the column of $\mathbf{F}$ can be interpreted as the return series of a specific eigenportfolio.

In the context of factor models, this is equivalent to saying that our q statistical factors explain a bulk of the variance of past returns, and each of the statistical factor in theory has zero correlation to each other.

3.3 Trading the Idiosyncratic Return Component

Before discussing how to model the auxiliary series as defined by equation (2.6), we first offer theoretical insights into why one would want to trade the "residuals" of such a factor model.

First, partition the right singular vector matrix $\mathbf{V}$ of the standardized return matrix $\tilde{\mathbf{R}}$ into two blocks:

$$\mathbf{V} \;=\; \begin{bmatrix} \mathbf{V}_1 & \mathbf{V}_2 \end{bmatrix}$$

with $\mathbf{V}_1 \in \mathbb{R}^{p \times q}$ representing the top q eigenvectors ($q \ll p$), and $\mathbf{V}_2 \in \mathbb{R}^{p \times (p-q)}$ the remaining eigenvectors.

Define

$$\mathbf{Y} \;=\; \tilde{\mathbf{R}}\,\mathbf{V}$$

so that

$$\mathbf{Y} \;=\; \begin{bmatrix}\mathbf{Y}_1 & \mathbf{Y}_2\end{bmatrix}$$

where $\mathbf{Y}_1 = \tilde{\mathbf{R}}\,\mathbf{V}_1 \in \mathbb{R}^{n\times q}$ and $\mathbf{Y}_2 = \tilde{\mathbf{R}}\,\mathbf{V}_2 \in \mathbb{R}^{n\times(p-q)}$. Notice that $\mathbf{Y}_1$ comprises projected factor returns, but instead of using the original return matrix $\mathbf{R}$ and the vol-weighted eigenportfolio matrix $\mathbf{Q}$, we substitute the standardized return matrix $\tilde{\mathbf{R}}$ and the truncated eigenvector matrix $\mathbf{V}_1$.

Since $\mathbf{V}$ is an orthogonal matrix, we can right-multiply by $\mathbf{V}^T$ to $\mathbf{Y}$ and represent the standardized return matrix as:

$$\tilde{\mathbf{R}} \;=\; \mathbf{Y}\,\mathbf{V}^T \;=\; \begin{bmatrix}\mathbf{Y}_1 & \mathbf{Y}_2\end{bmatrix}\begin{bmatrix}\mathbf{V}_1^T \\ \mathbf{V}_2^T\end{bmatrix} \;=\; \mathbf{Y}_1\,\mathbf{V}_1^T \;+\; \mathbf{Y}_2\,\mathbf{V}_2^T$$

Recall the original return matrix:

$$\mathbf{R} \;=\; \tilde{\mathbf{R}}\,\mathrm{diag}(\boldsymbol{\sigma}) \;+\; \mu$$

where μ is an $n \times p$ matrix whose rows are the mean return vectors, and $\mathrm{diag}(\boldsymbol{\sigma})$ is the $p \times p$ diagonal matrix of return volatilities $(\sigma_1, \ldots, \sigma_p)$.

Substituting $\tilde{\mathbf{R}} = \mathbf{Y}\,\mathbf{V}^T$ yields

$$\mathbf{R} \;=\; \mu \;+\; \big(\mathbf{Y}_1\,\mathbf{V}_1^T \;+\; \mathbf{Y}_2\,\mathbf{V}_2^T\big)\,\mathrm{diag}(\boldsymbol{\sigma})$$

We interpret

$$\mathbf{R}_{\text{sys}} \;=\; \mathbf{Y}_1\,\mathbf{V}_1^T\,\mathrm{diag}(\boldsymbol{\sigma}) = \tilde{\mathbf{F}}\boldsymbol{\beta} \quad \big(\text{where } \tilde{\mathbf{F}} = \mathbf{Y}_1,\ \boldsymbol{\beta} = \mathbf{V}_1^T\,\mathrm{diag}(\boldsymbol{\sigma})\big)$$

as the systematic return component, and

$$\boldsymbol{\epsilon} \;=\; \mathbf{Y}_2\,\mathbf{V}_2^T\,\mathrm{diag}(\boldsymbol{\sigma})$$

as the idiosyncratic component. Hence,

$$\mathbf{R} \;=\; \mu + \tilde{\mathbf{F}}\boldsymbol{\beta} \;+\; \boldsymbol{\epsilon} \tag{3.1}$$

In practice, the first q principal components in $\mathbf{Y}_1$ typically represent the systematic risk factors, while $\mathbf{Y}_2$ corresponds to smaller-variance directions in the data (residuals or noise). Assuming that each idiosyncratic return component $\boldsymbol{\epsilon}_i$ is well-behaved and mean-reverting, one can conjecture that large deviations arising from noise should revert to equilibrium.

This framework also illustrates why retaining too many eigenvectors is unwise: if we do so, the idiosyncratic component's variance becomes too small, so even if mean reversion occurs, it may not overcome trading costs.

Because drift in the return space is often negligible, we can represent the residual return of asset i by

$$\boldsymbol{\epsilon}_i \;=\; \mathbf{R}_i \;-\; \mathbf{F}\boldsymbol{\beta}_i \tag{3.2}$$

where factor exposures $\boldsymbol{\beta}_i$ are estimated by regressing the asset's returns on the eigen-portfolio returns.

Finally, actual trading signals are generated on the auxiliary series described in equation (2.6).

3.4 Modeling with the Ornstein-Uhlenbeck Process

This section provides an overview of how we model each auxiliary series $X_i(t)$ with an Ornstein-Uhlenbeck (OU) process. We begin by introducing the standard continuous-time formulation, then discuss two methods for discretely modeling and estimating parameters. One is a naive (Euler-based) approach that approximates the SDE when Δt is small, and the other is an exact method derived from the closed-form OU solution, allowing parameter recovery even if Δt is not negligible. We follow convention of Avellaneda (2008) and use the latter approach.

3.4.1 Brief Exposition

We assume each auxiliary series $X_i(t)$ follows

$$dX_i(t) \;=\; \kappa_i\big(m_i - X_i(t)\big)\,dt \;+\; \sigma_i\,dW_i(t) \tag{3.3}$$

where

$$\begin{aligned}
X_i(t) \;&=\; \text{the mean-reverting series for asset } i \text{ at time } t,\\
\kappa_i \;>\; 0 \;&=\; \text{rate of mean reversion},\\
m_i \;&=\; \text{long-run mean level},\\
\sigma_i \;&=\; \text{volatility of stochastic shocks},\\
W_i(t) \;&=\; \text{standard Wiener process for asset } i \text{ at time } t
\end{aligned}$$

The parameter κ_i controls how quickly $X_i(t)$ pulls back toward m_i. A larger κ_i implies a faster return to equilibrium, while σ_i governs the amplitude of random fluctuations around the mean.

In practice, we often observe X_i at discrete times $t_0, t_1, \ldots$, spaced by Δt. Let us take a look at baisc methodologies for obtaining parameter estimates of the OU process.

3.4.2 Naive Approach: Euler Discretization

A simple approximation uses the Euler-Maruyama method. We write

$$X_{h+1} - X_h = \kappa_i\big(m_i - X_h\big)\,\Delta t \;+\; \sigma_i\,\sqrt{\Delta t}\,\eta_{h+1}$$

where $\eta_{h+1} \sim \mathcal{N}(0,1)$. Hence,

$$X_{h+1} = \kappa_i \; m_i \Delta t + (1 - \kappa_i \Delta t) X_h \;+\; \sigma_i \sqrt{\Delta t}\; \eta_{h+1}$$

For Δt small (so that $\kappa_i \Delta t \ll 1$), this discrete model looks like an AR(1) process. Hence, this provides motivation to fit a linear regression of X_{h+1} on X_h directly and reverse engineer the OU parameter estimates from the regression estimates. Note the quality of approximation may degrade as Δt grows (Gillespie, 1996). In practice, the next method is preferable for precise parameter estimation.

3.4.3 Exact Approach: Closed-Form Discretization and Parameter Estimation

When Δt is not negligible, we should use an exact update formula by solving the differential equation.

Introducing the integrating factor $u(t) = X_i(t)\, e^{\kappa_i t}$, we can solve for the exact solution:

$$X_i\big(t_0 + \Delta t\big) = e^{-\kappa_i \,\Delta t}\, X_i(t_0) \;+\; \big(1 - e^{-\kappa_i \,\Delta t}\big)\, m_i \;+\; \sigma_i \int_{t_0}^{t_0+\Delta t} e^{-\kappa_i \big((t_0+\Delta t)-s\big)}\, dW_i(s)$$

In discrete form, for any integer h,

$$X_{h+1} = e^{-\kappa\,\Delta t}\, X_h \;+\; \big(1 - e^{-\kappa\,\Delta t}\big)\, m \;+\; \sigma \sqrt{\frac{1 - e^{-2\,\kappa\,\Delta t}}{2\,\kappa}}\; \eta_{h+1}$$

with $\eta_{h+1} \sim \mathcal{N}(0,1)$. This leads naturally to an AR(1)-like regression:

$$X_{h+1} = \underbrace{\big(1 - e^{-\kappa\,\Delta t}\big)\, m}_{a} \;+\; \underbrace{e^{-\kappa\,\Delta t}}_{b}\, X_h \;+\; \tilde{\zeta}_{h+1}$$

where $\mathrm{Var}\big(\tilde{\zeta}_{h+1}\big) = \sigma^2\, \frac{1-e^{-2\,\kappa\,\Delta t}}{2\,\kappa}$. By fitting

$$X_{h+1} = a + b\, X_h + \tilde{\zeta}_{h+1}$$

over $h = 1, \ldots, n-1$, one obtains $\hat{a}$, $\hat{b}$, and $\widehat{\mathrm{Var}}(\tilde{\zeta})$. Substituting

$$b = e^{-\kappa\,\Delta t}, \quad a = m\big(1 - b\big), \quad \mathrm{Var}\big(\tilde{\zeta}\big) = \sigma^2\, \frac{1 - e^{-2\,\kappa\,\Delta t}}{2\,\kappa}$$

yields

$$\hat{\kappa} = -\frac{1}{\Delta t}\, \ln(\hat{b}), \quad \hat{m} = \frac{\hat{a}}{1 - \hat{b}}, \quad \hat{\sigma} = \sqrt{\frac{\widehat{\mathrm{Var}}\big(\tilde{\zeta}\big)\, \big(2\,\hat{\kappa}\big)}{1 - e^{-2\,\hat{\kappa}\,\Delta t}}}$$

One may also define $\hat{\sigma}_{\mathrm{eq}} = \sqrt{\frac{\widehat{\mathrm{Var}}(\tilde{\zeta})}{1-\hat{b}^2}}$ as the "equilibrium" standard deviation of $X_i(t)$. This approach does not rely on $\kappa\,\Delta t \ll 1$ and is more accurate for moderate or large Δt.

3.4.4 Mean reversion Time

A useful scalar is $\tau_i = 1/\kappa_i$, denoted as the *mean reversion time*, which can be thought of as a proxy for how long it takes $X_i(t)$ to revert to its long-run mean m_i. One may discard series whose estimated τ_i greatly exceeds a chosen threshold, as the convergence may be too slow to be profitable. Avellaneda advocates setting the threshold to be one-half of the estimation period for the OU process (e.g., if sixty days of data are used to estimate the OU parameters, then set $\tau_{\text{threshold}} = \frac{30}{252}$).[2]

3.4.5 Signal Generation

After obtaining $\hat{\kappa}_i$, $\hat{m}_i$, and $\hat{\sigma}_i$ via the exact method, we can determine when the series is "far" from $\hat{m}_i$. Defining

$$\hat{\sigma}_{\text{eq},\,i} = \sqrt{\frac{\widehat{\text{Var}}(\tilde{\zeta}_i)}{1 - \hat{b}_i^2}} \quad \text{and} \quad s_i(t) = \frac{X_i(t) - \hat{m}_i}{\hat{\sigma}_{\text{eq},\,i}},$$

we can adopt threshold rules (e.g., ± 2 standard deviations) for opening and closing positions. These rules parallel the s-score approach originally proposed by Avellaneda (2008).

3.5 Execution and Backtesting

3.5.1 Execution Nuance

The original paper by Avellaneda (2008) does not fully clarify how to implement the backtest in practice, and most subsequent work applying Avellaneda's framework focuses on ETF-based strategies, which are simpler than those involving eigenportfolios.

To understand why eigenportfolios add complexity, consider how one would execute a single trade on day t. Suppose we have:

1. Extracted new PCA factors from the correlation matrix,
2. Performed factor regressions to obtain residual returns and their corresponding auxiliary series,
3. Estimated Ornstein-Uhlenbeck (OU) parameters,

[2] The denominator is the number of business days within a year. Since cryptocurrencies trade for all 365 days, the denominator in this case would be 365.

4. Filtered out coins whose mean reversion time τ_i exceeds a threshold $\tau_{\text{threshold}}$,

5. Formed a *mean reversion list* mrl_t.

Next, for each coin $i \in mrl_t$, compute the daily s-score. Suppose coin k triggers a long signal ($s_t <= s_{bo}$). Then, to remain factor neutral, we go long \$1 in coin k and short $\$\beta_{k,j}$ in each eigenportfolio $j = 1, \ldots, q$.[3] By construction, this "hedged portfolio" involves leverage of $1 + \sum_{j=1}^{q} |\beta_{k,j}|$.

In practice, each eigenportfolio is itself a synthetic basket of all coins in TU_t, including coin k. Also, there are q different eigenportfolios, each formed from the same coin set. Denoting

$$p_t = \#mrl_t = \text{the number of coins in the mean reversion list at time } t,$$

we can have up to p_t hedged positions. This raises questions about whether to treat each hedged position separately (so each coin forms its own "mini-portfolio") or whether to aggregate weights across all coins and manage leverage at the aggregate level. For this study, we chose to adopt the latter approach, which is operationally much simpler.

3.5.2 Drift in Dollar Weights

A second issue involves the daily rebalancing of weights to maintain hedge ratios derived from factor regressions. Unlike a price-spread model, here the hedge ratios are in dollar terms: if the portfolio return drifts relative to the security return, the initial weights also drift. Maintaining factor neutrality requires frequent rebalancing to restore the desired dollar weights, but such frequent trades can be costly.

To address this, we run several backtest variations:

- Maintaining a beta-neutral position daily via aggregated coin weights
- Rebalancing beta-neutrality weekly with aggregated coin weights
- Rebalancing daily but aggregating returns through residual series of hedged portfolio

Because actual trades require specifying the number of coin units to buy or sell, it is crucial to have a clear method for updating the weight vector each day. Appendix A provides further details on the rebalancing mechanism.

[3] We generate signals based on the closing price and assume to trade at the close as well. While this may not reflect reality, we can take advantage of the fact that cryptocurrencies trade 24/7.

3.5.3 Dynamic Rebalancing of Statistical Factors

A key assumption in using statistical factors for hedging is that those factors remain relevant throughout the trading horizon. In reality, however, financial relationships evolve over time. Without rebalancing, our PCA-based factors could become outdated.

Hence, we implement dynamic rebalancing each month:

1. Recompute the trading universe TU_t,
2. Extract $q = 5$ new PCA factors from the updated correlation matrix,
3. Use these updated factors for the subsequent trades.

Our empirical results (see Section 4) support the view that rebalancing improves stability, as it helps ensure that the eigenportfolios track current market structure.

3.5.4 Walk-Forward Validation

Some replication reports focus solely on in-sample results for the PCA-based strategy, but in our view, this can be misleading. Because PCA, by construction, explains the variance of historical data, the in-sample auxiliary series may appear overly mean-reverting.

Instead, we split our data into:

- A sample period (2022-05-02 to 2022-07-01),
- A validation period (2022-07-01 to 2023-07-01),
- A test period (2023-07-01 to 2024-08-05).

On the validation set, we conduct a grid search (with constraints) to find the best s-score thresholds. We also test multiple backtest setups, as noted above. To simulate a realistic trading environment, we use a walk-forward approach with a rolling window of fixed length and a daily step. That is, each day we:

1. Re-estimate factor regression betas,
2. Recompute the auxiliary series and OU parameters,
3. Generate signals.

We then apply the parameter set (optimized from the validation phase) to out-of-sample (OOS) data, ultimately reporting the performance on the test set.

3.5.5 Parameter Choice

For this study, we use the following hyperparameters, primarily guided by both intuition and preliminary empirical analysis:

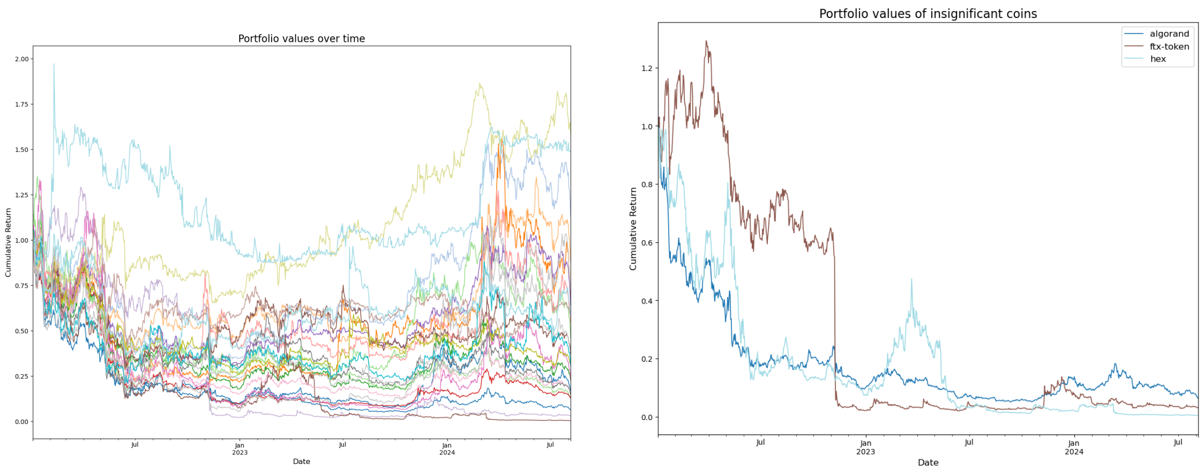
- PCA Estimation: 180 days
- Factor Regression: 60 days
- OU Estimation: 60 days
- q : number of top eigenvectors retained = 5
- $s_{thresholds}$: [-2.5 , -0.5, 2.5, 0.5] (from optimization)
- $\tau_{threshold}$: $\frac{60}{365}$ (from optimization)

4 Results

In this section, we present the results of our experiment. In the first half, we show the EDA results using constituents from $TU_{20220701}$. In the second half, we present the results from strategy backtesting.

4.1 EDA

4.1.1 Performance of Coins Over the Trading Period



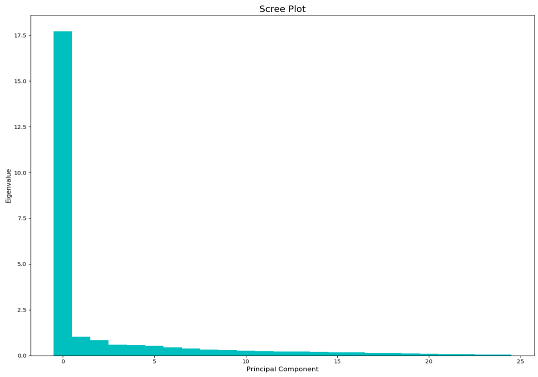
(a) Performance of all coins

(b) Performance of insignificant coins

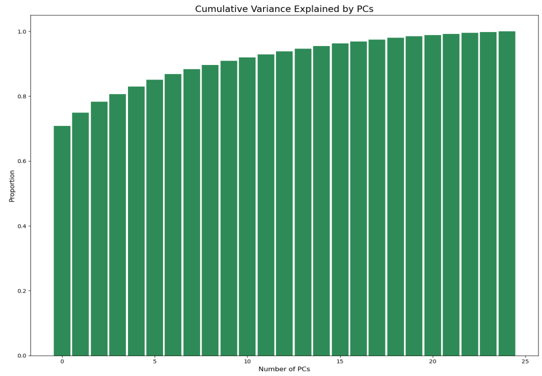
Figure 1: Performance Over the Entire Period

Figure 1a shows how a hypothetical dollar invested in one of the coins from $TU_{20220701}$ would have changed throughout the entire data collection period. As observed, the year 2022 was unfavorable for most of the coins in our universe, with only a few coins managing to reclaim their initial value by the end of the period (2024-08-05). Moreover, Figure 1b highlights coins whose final performance was below -80% by the end of the period, supporting our hypothesis that the trading universe should be resampled periodically to ensure the continued relevancy of our statistical factors.

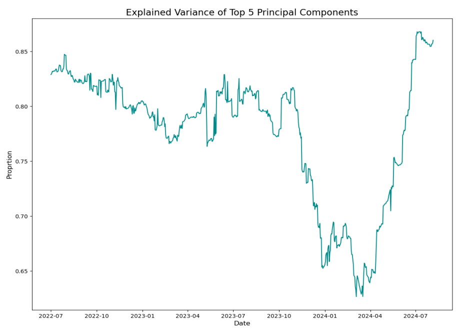
4.1.2 PCA Results



(a) Scree Plot



(b) Cumulative Variance Explained by PCs



(c) Total Variance Explained by 5 PCs

Figure 2: PCA Results

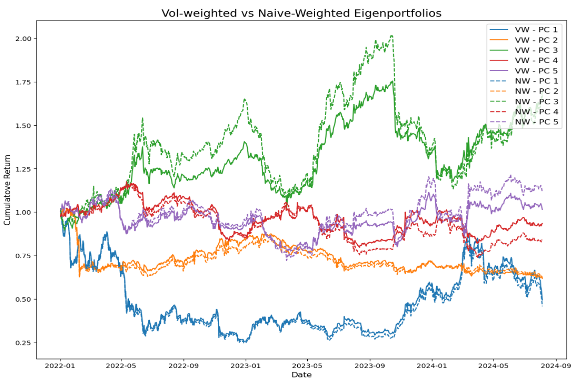
An important question we must ask ourselves is: how many eigenportfolios should we choose? In other words, what is the optimal number of eigenvectors to retain? Should this number be static or dynamic? The answer to this question is nuanced, with many studies attempting to address it using various methodologies. However, an intuitive way to approach this is to include enough eigenvectors to capture all systematic factors while avoiding the inclusion of "noise" eigenportfolios. As we observed in Section 3, including too many eigenvectors can reduce the variance of the idiosyncratic (or residual) return, making it less profitable for mean reversion strategies. Fortunately, the characteristics of the cryptocurrency market allow us to select only a few eigenportfolios while still capturing most of the variance within our trading universe.

The scree plot (Figure 2a) displays the eigenvalues in descending order, where the largest eigenvalue corresponds to the first eigenvector, or Principal Component 1 (PC 1). Figure 2b illustrates the contribution of the top i eigenvectors to the total variance of the data, $\forall i = 1, 2, \ldots, p$. Notably, over 60% of the total variance can be explained by the first principal component. This contrasts with the equities market, where the first principal component typically explains less than 40% of the variance.[4] This again highlights the higher concentration in the cryptocurrency market compared to the public equity market.

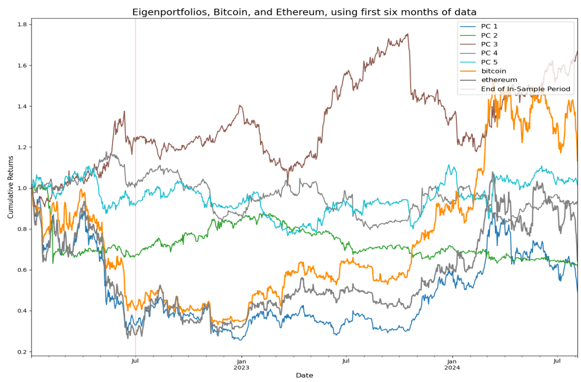
Figure 2c shows the total variance explained by the top five principal components over the entire data collection period. Figure 2b corresponds to the first data point in Figure 2c. The decline in explained variance coincides with the bear market period in the cryptocurrency market, while the subsequent increase aligns with the Bitcoin-led bull market that began in October of 2023. This pattern suggests that systematic factors play a more dominant role during periods of high volatility (typically driven by a few major coins) and become less significant during sideways market movements.

Based on this analysis, we can conclude that selecting five eigenportfolios effectively represents a large portion of the variance in our data. Moreover, Figure 2c is generated using $TU_{20220701}$; since our strategy involves periodically updating the trading universe, we can expect the explained variance to remain higher than that seen in Figure 2c over time.

4.1.3 Performance of Eigenportfolios



(a) Vol-weighted vs Naive-weighted Eigenportfolios



(b) Eigenportfolios, Bitcoin, and Ethereum

Figure 3: Eigenportfolio Performance

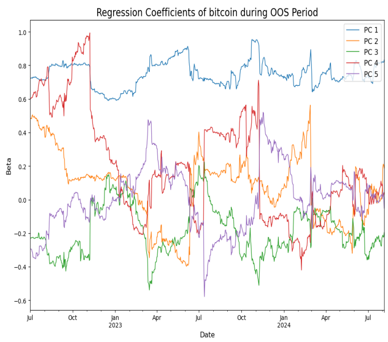
Avellaneda (2008) employs a vol-weighting scheme to compute the eigenportfolio weights.

[4]One may argue that this is a natural consequence of our investment universe consisting of fewer than 30 coins, compared to over 100 stocks in the public equity market. However, this is simply a consequence of the highly concentrated nature of the cryptocurrency market.

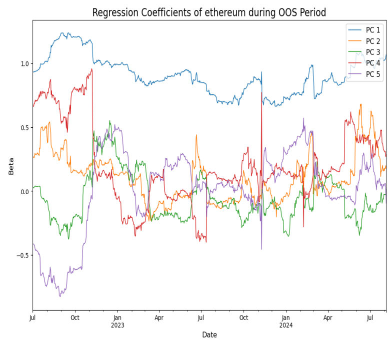
As shown in Figure 3a, both the vol-weighted and naive-weighted eigenportfolios (i.e., using the raw eigenvector as weights) exhibit similar performance, though the vol-weighted eigenportfolio is relatively less volatile overall. Figure 3b compares the performance of eigenportfolios with Bitcoin and Ethereum, with the initial value normalized to one. As discussed in the original paper, the principal eigenportfolio (PC 1) behaves similarly to a "market" portfolio. Interestingly, PC 1 "tracks" Ethereum more closely than Bitcoin, suggesting that coins in our trading universe generally have a higher correlation with Ethereum than Bitcoin.

An important observation is that our principal eigenportfolio, estimated using data from six months prior to 2022-07-01, begins to diverge from Ethereum and Bitcoin starting in late 2022. This indicates the need for periodic updates to the trading universe. However, the principal eigenportfolio tracks Ethereum and Bitcoin well for almost four months during the out-of-sample period, suggesting that an update frequency of less than three months may be sufficient.

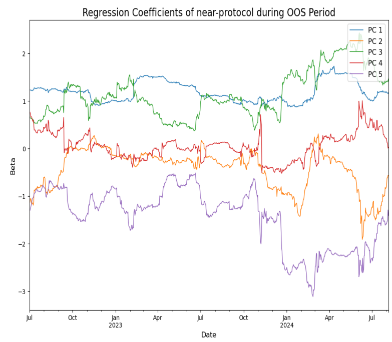
4.1.4 Stability of Hedge Ratios



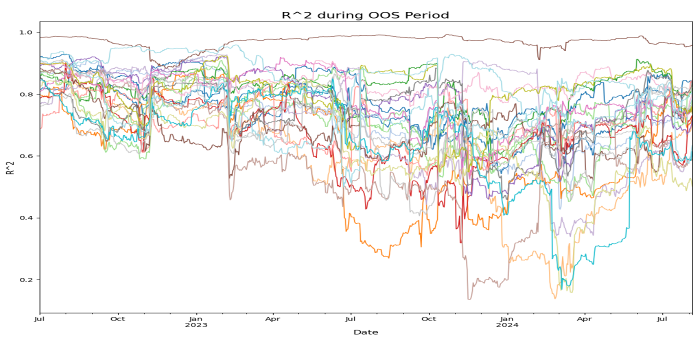
(a) Hedge Ratios for BTC



(b) Hedge Ratios for ETH



(c) Hedge Ratios for NEAR



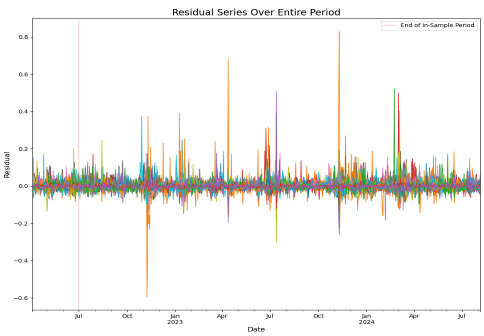
(d) R^2 Over Entire Period

Figure 4: Factor Regression Results

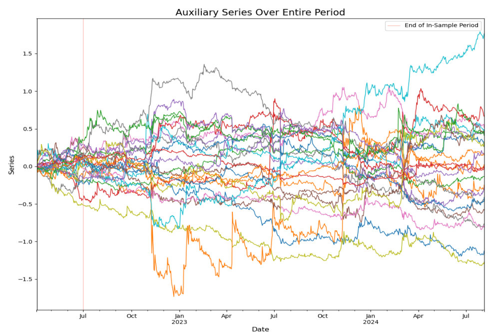
One of the main properties of a robust factor is that its influence on securities remains fairly consistent over a sufficiently long period of time. In other words, if we conduct a rolling regression of the coin return on the factor return, the resulting regression beta (hedge ratio) should not vary significantly over time. In particular, a change in sign indicates an opposite direction of effect, so constant fluctuations in sign suggest that the factor in question is unstable.

Figures 4a, 4b, and 4c display the rolling beta coefficients from regressing the coin returns on the eigenportfolio returns. Except for the beta coefficient corresponding to the principal eigenportfolio, the remaining betas exhibit non-trivial volatility, indicating that the assumption of static statistical factors is overly simplistic. The implication of Figure 4d is similar to that of Figure 2c; the explanatory power of statistical factors varies depending on the market regime.

4.1.5 Residual & Auxiliary Series



(a) Residual Series Over Entire Period



(b) Auxiliary Series Over Entire Period

Figure 5: Constructing the Idiosyncratic Component

Even after subtracting the systematic component of return, which accounts for almost 80% of the variance, most of the auxiliary series do not appear to exhibit mean-reverting behavior (Figure 5b). Another noteworthy observation is that the auxiliary series, which is simply the cumulative sum of the residual series, does not converge to zero at the end of the in-sample period. Recall that the construction of the regression equation enforces the sum of residuals to be zero; however, the residual series displayed in Figure 5a is computed without subtracting the drift (intercept) component.[5] Therefore, we can conjecture that drift in returns may be non-trivial, unlike what is typically observed in the equity space. In conclusion, the analyses presented in Sections 4.1.5 and 4.1.4 suggest that the statistical arbitrage strategy proposed in Avellaneda's paper (2008) may not be effective in the crypto space.

[5]Note that when we construct the hedged portfolio, we can only trade the coin and eigenportfolios. Hence, drift is an uncontrollable component, so we do not account for it when computing the residuals. The drift should be considered when generating the signal.

4.2 Backtesting Results

For the analysis in this section, assume an initial equity amount of 10,000. For out-of-sample performance, a specific level of leverage is utilized such that the realized volatility of the pairs trading strategy matches that of buy-and-hold Bitcoin. Trading cost and the risk-free rate is assumed to be zero unless specified otherwise. $\tau_{threshold} = \frac{30}{365}$ is used to test the in-sample performance.

As mentioned in Section 3.5.3, we test several variations of the strategy proposed in Avellaneda (2008). The different "versions" of the strategy are detailed below:

- **Version 1** is essentially the in-sample backtest, where statistical factors based on $TU_{20220701}$ are used to construct auxiliary series and generate trading signals. Regression betas, OU parameter estimates, and s-scores were computed daily, as in Avellaneda's paper (2008).
- **Version 2** mirrors the strategy discussed in the original paper. However, unlike the paper, statistical factors are extracted on a monthly basis to maintain relevancy. Betas are re-estimated daily so that each trading day uses a fresh 60-day rolling window of regressions to generate residuals, OU parameters, and signals. Positions are closed at each monthly rebalance date.
- **Version 3** performs monthly PCA recalculation as in Version 2, but betas are re-estimated weekly. That is, the same hedge ratio is used for a week to compute OU parameter estimates and s-scores. The objective of this version is to reduce turnover resulting from restoring the dollar hedge ratios on a daily basis. Instead, the hedged portfolio is allowed to "drift" for a week to see if performance can be maintained while reducing trading costs.
- **Version 4** adopts the same schedule as Version 2 for PCA and betas but adds an "inclusion-exclusion" counter: a coin must appear in the mean reversion list for a few consecutive days before opening a position, and positions are not immediately closed if that coin briefly disappears from the mean reversion list. Instead, there is a multi-day grace period before the position is forcibly exited. Again, the aim of this version is to reduce trading costs from frequent rebalancing.
- **Version 5** similarly recalculates PCA each month and re-estimates betas daily, but it computes daily P&L directly from the residual returns rather than based on aggregated coin weights.

To understand how subtle changes in execution and rebalancing can affect trading outcomes, we implemented four backtest versions (for out-of-sample data) that share a common framework of PCA-based factor extraction, factor-regression hedge ratios, and OU mean reversion checks, yet differ in how frequently these steps are updated and how positions are managed. Contrary to our belief that reducing trading costs would be beneficial, the results indicated that Version 2 was the most stable variation. That is, the original strategy advocated by Avellaneda, with the added component of monthly rebalancing, outperformed the more "complicated" versions designed to reduce trading costs.

Hence, we present out-of-sample (OOS) results using Version 2 only.

4.2.1 In-Sample

Our in-sample period corresponds to the estimation period of the residuals (a 60-day period ending on 2022-07-01).

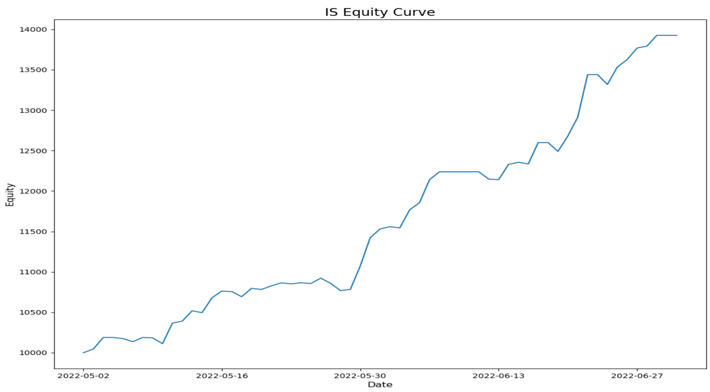


Figure 6: In-Sample Equity Curve of Strategy

When tested on the sample dataset, the strategy performs well, achieving a realized Sharpe ratio of approximately 18. However, this is an extremely biased result for a variety of reasons, with two main contributing factors:

1. We used statistical factors extracted from $TU_{20220701}$, but in practice, we cannot use a trading universe based on a *future* date. This introduces a typical example of look-forward bias.

2. Note that we have used statistical factors extracted from the *future* date; by definition, PCA selects new axes that maximize the variance of past data. Hence, it is natural for the sample dataset, when subtracting the systematic return component, to exhibit mean-reverting behavior. However, this does not imply that the statistical factors will effectively summarize the future systematic return component. This represents a critical error due to look-forward bias.

Therefore, the above results should be disregarded as they do not provide meaningful insights into how the strategy would perform in live trading. Instead, they simply indicate that the strategy *in theory* would perform well if the statistical factors extracted from past data indeed explain a significant portion of the future return data.

The next section provides a more "realistic" description of the strategy's performance.

4.3 Walk-Forward Validation & Optimization

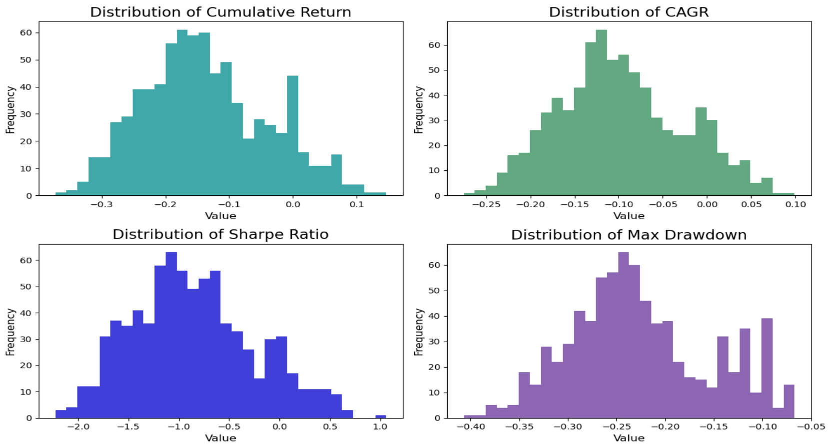


Figure 7: Distribution of Performance Metrics

Figure 7 shows the distribution of simulated performance metrics when we iterated over a constrained set of s-score thresholds. Overall, the strategy's out-of-sample performance is quite poor compared to the in-sample performance; the histogram peaks around -1 for the Sharpe ratio. Therefore, even if we adopt the most optimal parameter set retrieved from our validation dataset, we remain skeptical about whether that set of thresholds is truly robust.

In fact, when we evaluated the performance of the strategy using the optimal parameter set, we found that there were a lot less trades generated during the second half of the validation period (Figure 8). This is another indicator that the strategy is not robust.

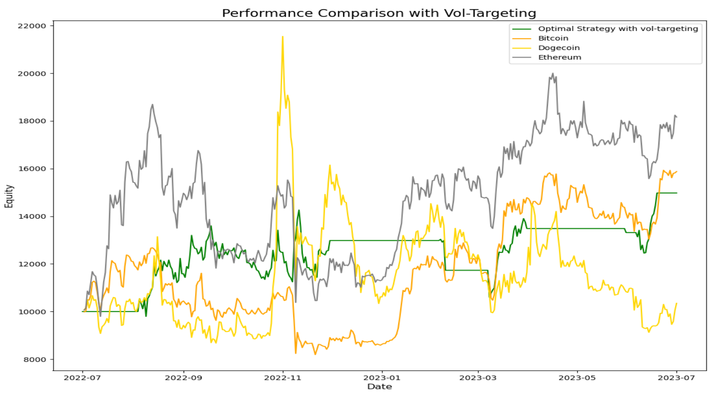
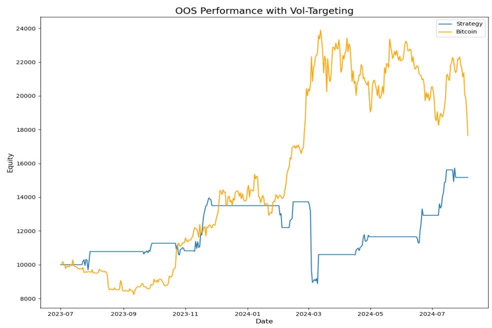


Figure 8: Optimal Strategy Performance Over Validation Period

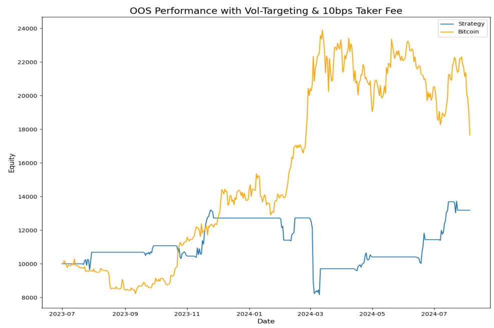
4.4 Testing Period

From optimizing on the validation dataset, the following parameter set is used.

- $s_{thresholds}$: [-2.5 , -0.5 , 2.5 , 0.5]
- $\tau_{threshold}$: $\frac{60}{365}$



(a) Vol-Targeting



(b) Vol-Targeting & 10bps Taker Fee

Figure 9: Test Set Performance

Metric	Vol-Targeting	Vol-Targeting & 10bps Taker Fee
Cumulative Return	51.77%	31.87%
CAGR	29.93%	18.97%
Annualized Volatility	35.43%	35.33%
Sharpe Ratio	0.84	0.54
MDD	-36.37%	-38.17%
MDD Period	107 days	107 days

Table 1: Performance Metrics for Test Set

From observing trading performance in Figure 9, it is difficult to verify the statistical significance of the above results, since there are not enough trades generated, with the total number of independent signals being 14 over the entire test period.

5 Discussion

5.1 Insufficient Number of Trades

The statistical arbitrage strategy explored relies on trading mean-reverting hedged portfolios, which are found via a filtering based on the $\tau_{threshold}$. However, setting a too small $\tau_{threshold}$ filters out *most* of the coins in the trading universe; this leaves with only a few coins per day as possible mean-reverting candidates. Hence, while a small trading universe allows estimating the population covariance matrix without suffering from the typical issues regarding curse of dimensionality (Ledoit, 2003), there comes with the cost of not enough signals being generated.

5.2 Persistence of Mean Reversion and Non-Mean Reversion

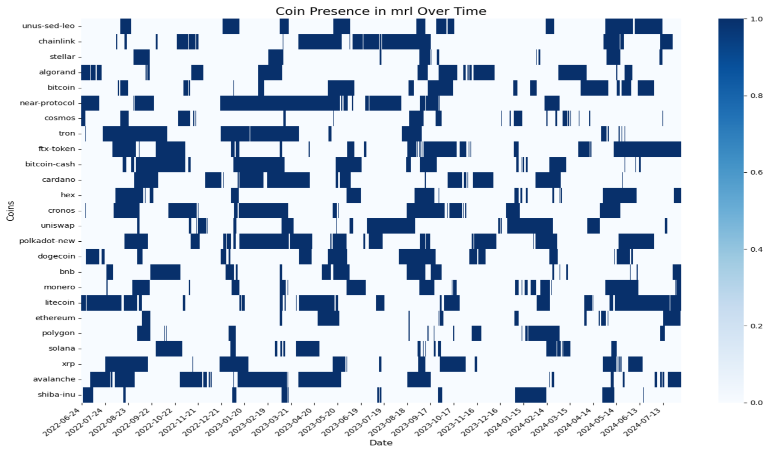


Figure 10: Coin Presence in mrl_t Over Entire Period

An interesting observation from the daily *mean reversion lists* mrl_t is that membership tends to be *sticky*: once a coin consistently exhibits mean-reverting behavior, it often remains classified as such for multiple successive days, and coins not deemed mean-reverting tend to stay out of the list. Figure 10 (the heatmap of coin presence in mrl_t) visually highlights how individual names repeatedly reappear or remain absent, suggesting that the process of "losing" or "gaining" mean reversion status is governed by more than short-term noise. In practice, this means a coin that is excluded from mrl_t today is likely to be excluded for some stretch of time, and similarly, one that is included may continue to appear for several days.

Such stickiness has direct implications for trading, especially for deciding how quickly to close positions if a coin *ceases* to be mean-reverting. Although a conservative approach might be to exit immediately whenever a coin disappears from mrl_t, this can induce excessive churn and transaction costs if a coin's classification fluctuates briefly but remains generally mean-reverting. Conversely, waiting too long risks holding a position that truly lost its stationarity. Hence, a more nuanced approach—where we allow up to ν consecutive days of exclusion before closing—can mitigate excessive turnover while cutting losses on genuinely non-mean-reverting names. Determining the right threshold ν largely depends on how strongly one values avoiding "false positives" (maintaining trades in coins that no longer revert) versus "false negatives" (forgoing profitable trades in coins that still revert). As the heatmap suggest, modeling these regime shifts with a carefully chosen grace period may better align real execution with the observed persistence of mean reversion status.

5.3 Mean Reversion Parameter: A Proxy for Stop Loss

As seen in Section 4.2, different variations of the strategy were explored, mostly aimed to reduce trading costs from attempting to restore hedge ratios on a daily basis and switching in and out of positions based on inclusion/exclusion from mrl_t. However, all of these variations failed to improve performance mainly due to their inability to get out of losing trades before suffering a large drawdown. This was a critical performance drag; however, using $\tau_{threshold}$ as a daily filter enabled us to get out of positions that experienced sudden spikes in price. In other words, if re-estimation of the series to fit the Ornstein Uhlenbeck process results in mean reversion parameter κ_i to become small enough, this acts as a stop loss, which in some cases act as a more reliable and logical stop loss than using for instance, a typical hard stop loss.

5.4 Limitations and Improvements

5.4.1 Data

The daily price series of cryptocurrencies and related market data were scraped from the CoinMarketCap API. Although the daily OHLCV data was mostly error-free, it does differ from data obtained directly from an exchange. Moreover, the test set consisted of data spanning just over a year. During this period, the cryptocurrency market experienced significant rallies, and mean reversion strategies generally underperform in such bullish environments. Testing the strategy on datasets that encompass a wider range of market regimes could yield more robust performance results.

Furthermore, for a more comprehensive study, one could test the strategy on intraday datasets and expand the trading universe to include a larger set of cryptocurrencies. This would help determine whether including smaller market cap coins improves the strategy's performance.

5.4.2 Too Many Hyperparameters

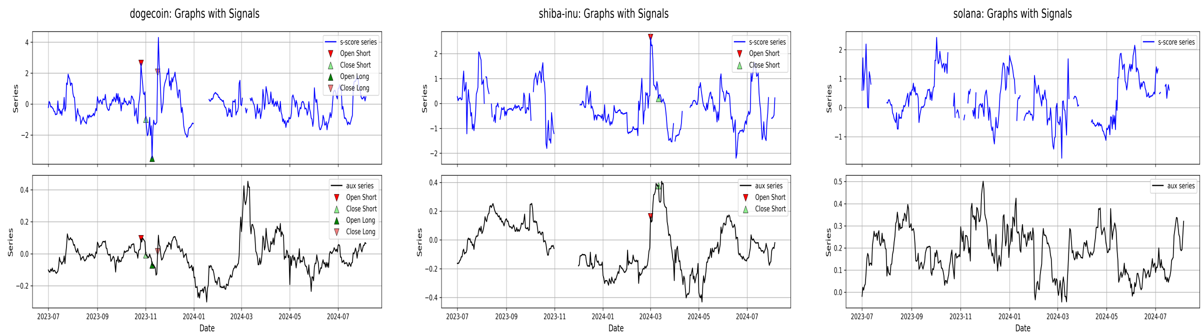
Admittedly, this strategy relies on many hyperparameters. For example:

- **Selection of Trading Universe:** EWMA parameter λ, negative shock threshold, median volume threshold, and market cap threshold.
- **Estimation Periods:** PCA estimation period (180 days), factor regression estimation period (60 days), OU process estimation period (60 days).
- **Eigenvectors to Retain:** While we have used the top five principal components for our eigenportfolios, this heuristic approach is not optimal in the strict sense.
- **Trading Thresholds:** $\tau_{threshold}$, s_{bo}, s_{sc}, s_{so}, and s_{bc}.

As more hyperparameters are introduced, the risk of overfitting the strategy to historical data increases, reducing the strategy's robustness and adding complexity that hampers controllability.

5.4.3 Trading the Residual Using Other Approaches

The s-scores are used as "signal lines" for the trading strategy. However, were the signals generated by these s-scores accurate in identifying entry points that led to mean reversion, and were the exit signals appropriately timed?



(a) Dogecoin (b) Shiba-Inu (c) Solana

Figure 11: S-scores (on top) and Auxiliary Series (on bottom) with Signals

Figure 11 presents a set of selected graphs displaying how the signals performed during the out-of-sample period. For instance, Figure 11a illustrates a case where the signals correctly captured moments of overvaluation and undervaluation, with trades executed effectively. Note that the blue series represents "signal lines", and the black series represents the performance of trades following execution.

Figure 11b demonstrates an example where trading based on the signal resulted in a loss. Despite entering a short position in Shiba-Inu based on the s-score series, this trade would have incurred an approximate 20% loss on the capital allocated to this coin.

Figure 11c showcases a case where no signal was generated. However, reviewing the auxiliary series with the benefit of hindsight, it appears that mean reversion strategies could have been effectively applied. This suggests that the methodology of generating signals based solely on s-scores may not be optimal.

For instance, alternative approaches, such as using Bollinger Bands or z-scores, could yield more frequent executions and potentially improve performance.

5.4.4 Idiosyncratic Component Does Not Mean-Revert

A more fundamental limitation of this strategy is the possibility that the idiosyncratic component of cryptocurrency returns does not exhibit mean reversion. This could be due to either: 1) The statistical factors estimated by PCA fail to eliminate all systematic return components, or 2) Cryptocurrencies, as an asset class, have higher idiosyncratic risk compared to publicly traded equities.

A potential solution to issue (1) is to incorporate fundamental factors that may better capture the systematic return components of cryptocurrencies. However, unlike equities, there is not a lot of evidence on robust factors for cryptocurrencies.

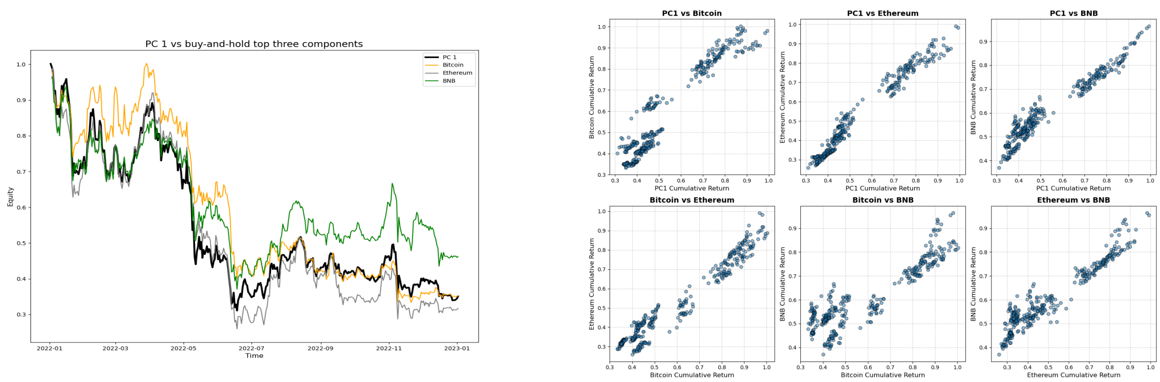
For issue (2), it is plausible that many cryptocurrencies experience sharp inflows and outflows due to coin-specific factors, leading to events commonly referred to as "pump-and-dumps". This phenomenon is particularly prevalent among small-cap coins.

5.5 Bonus Strategy: Integrating Cointegration with PCA

While the original approach proposed by Avellaneda (2008) did not yield outstanding results, we believe that PCA can still be effectively utilized to develop profitable pairs trading strategies. This section presents a brief discussion on how one can use the first principal component, or the principal eigenportfolio, to construct a simple pairs trading strategy based on cointegration.

Unlike equities, there is no widely accepted "market index" for cryptocurrencies. Bitcoin and Ethereum often serve as market proxies, but their status as reliable benchmarks is questionable. At times, these assets experience idiosyncratic inflows and outflows. For instance, the late-2024 crypto bull market was primarily driven by Bitcoin, while most altcoins failed to deliver strong performance.

There is substantial evidence in the literature suggesting that the principal eigenportfolio can serve as a proxy for the market factor. This is reflected in Figure 12a, which shows the normalized price movements of the principal eigenportfolio and major cryptocurrencies over time. Given that the principal eigenportfolio can be actively traded using its vol-weighted eigenvector components, we explore whether it can be used to construct a pairs trading strategy via cointegration analysis.



(a) Principal Eigenportfolio and mainstream coins over a year

(b) Scatterplot of cumulative returns

Figure 12: Exploratory Data Analysis for Bonus Strategy

5.5.1 Methodology

1. **Finding the Market Factor**: Define an estimation period and an out-of-sample period. PCA is conducted using 365 days of historical data to estimate the "market factor." The trading universe at the end of the formation period $TU_{20230101}$ is used.

2. **Cointegration Analysis**: Conduct Engle-Granger cointegration regressions (in both directions) between the constituents of $TU_{20230101}$ and the first eigenportfolio.

3. **Selection Criteria**: Only select coins for trading if the p-value from ADF test for cointegrating residuals is below 0.1 for both directions.

4. **Trading Simulation**: Allocate $1 of initial equity to each cointegrated pair in an equally-weighted manner, and trade until the next rebalance date (2 months). The rolling z-score strategy is applied for trading the spread.

5. **Rebalancing**: At the next rebalance date, extract a new trading universe TU_t, perform PCA on the latest data, and conduct cointegration analysis. Select coins that meet the cointegration criteria.

6. **Equity Allocation**: Set the new initial equity to the final equity value from previous trading period and evenly distribute it among the new set of cointegrated pairs. Continue trading for 2 months using betas from the regression analysis.

7. **Repeat Until End of Data**: The above process is repeated until the end of the dataset.

5.5.2 Results

Table 2 summarizes the results of the cointegration analysis conducted using the trading universe $TU_{20230701}$. The selected coins for trading are Polkadot, Shiba-Inu, Solana, and BNB. The regression beta obtained from the cointegration test serves as the hedge ratio.

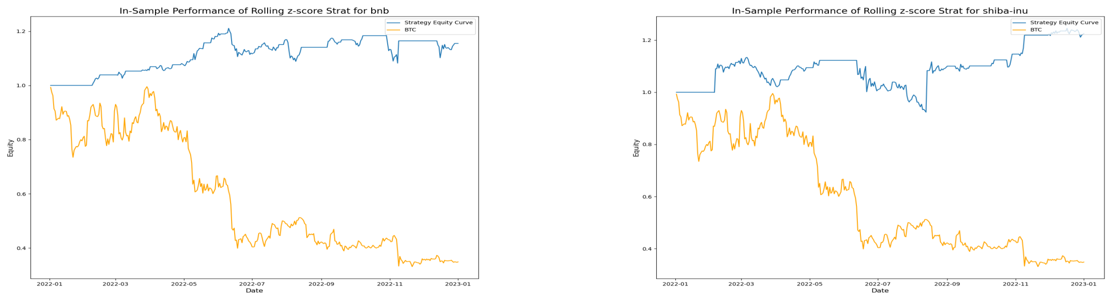
A key assumption of this strategy is that fractional trading of coins is possible. While this is often unrealistic in equity markets, it is feasible in the cryptocurrency market, where fractional units can be traded. In practice, traders should ensure they have sufficient equity to meet minimum order size requirements.

To trade the spread, a rolling z-score strategy is applied. A 20-day moving average is used to de-trend the spread, and entry/exit points are defined using upper and lower bands of ± 2 standard deviations. These parameters can be optimized, as discussed in Section 3.5.5.

Figure 13 shows two example equity curves from testing the strategy using sample data. As one can see, the in-sample performance all ended in a positive return during a period when most cryptocurrencies decreased in price. While the in-sample performance does

Coin	ADF Stat 1	P-Value 1	ADF Stat 2	P-Value 2	Best ADF Stat	Best P-Value	Direction
Polkadot-New	-3.9540	0.0313	-4.8290	0.0017	-4.8290	0.0017	Coin $\rightarrow$ PC1
Shiba-Inu	-4.2059	0.0148	-4.3786	0.0085	-4.3786	0.0085	Coin $\rightarrow$ PC1
Solana	-3.7425	0.0553	-4.2410	0.0132	-4.2410	0.0132	Coin $\rightarrow$ PC1
BNB	-3.6530	0.0692	-3.9452	0.0321	-3.9452	0.0321	Coin $\rightarrow$ PC1

Table 2: Engle-Granger Test Results

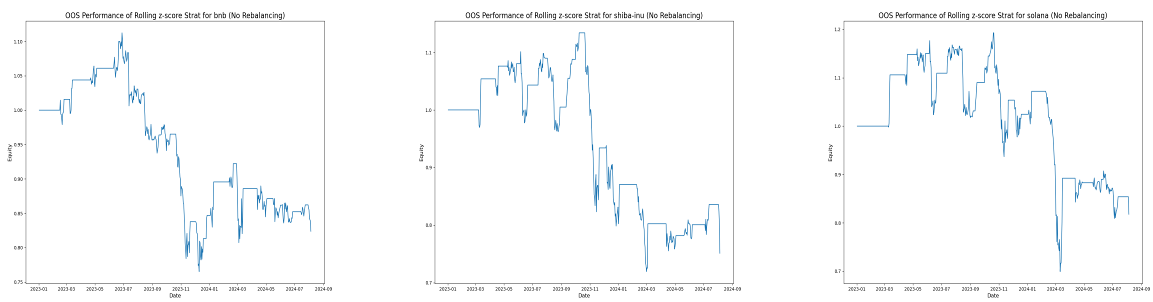


(a) BNB (b) Shiba-Inu

Figure 13: Examples of In-Sample Performance of Bonus Strategy

not indicate future performance in any way, it can serve as a rough measure of how the out-of-sample performance equity curve should look if the strategy works well.

As discussed in Section 3.5.4, it is important to select a new trading universe periodically and compute new principal eigenportfolio to maintain relevancy and avoid trading "insignificant" coins. To motivate the adoption of rebalancing, we first present a set of OOS performance results of the bonus strategy without using rebalancing. That is, the principal eigenportfolio estimated based on $TU_{20230101}$ will be used for the entire out-of-sample period (2023-01-01 $\sim$ 2024-08-05).



(a) BNB (b) Shiba-Inu (c) Solana

Figure 14: Out-of-Sample Performance of Bonus Strategy without Rebalancing

Figure 14 highlights the importance of rebalancing. As one can see, the strategy's performance holds up for the first 2$\sim$3 months but then starts to deteriorate rapidly as we move further into the future. Hence, this motivates the usage of a rebalancing period of 2$\sim$3 months. In our case, we have chosen to use 2 months as the rebalancing frequency.

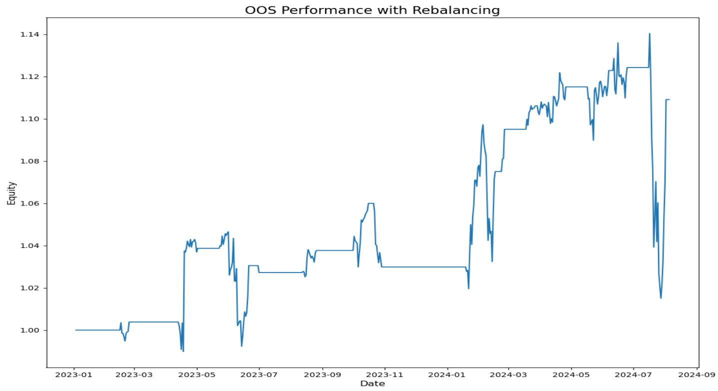


Figure 15: Out-of-Sample Performance of Bonus Strategy with Rebalancing

Metric	Bonus Strategy with Rebalancing
Cumulative Return	0.1091
CAGR	0.0459
Annualized Volatility	0.0898
Sharpe Ratio	0.5115
Maximum Drawdown (MDD)	-0.1099
MDD Period	11 days

Table 3: Performance Metrics for Bonus Strategy with Rebalancing

The rebalanced strategy in Figure 15 shows a significant improvement compared to a non-rebalanced approach. While the overall performance is not exceptional, it is important to note that this result was obtained from the first out-of-sample test without any parameter optimization.

Additionally, the performance drop near the end of the period was due to trading only a single cointegrated pair (Solana) from 2024-06-25 to 2024-08-05. For other trading periods, the strategy traded at least two pairs and at most seven pairs. Thus, to avoid excessive concentration risk, one potential improvement would be to under-allocate capital when fewer pairs are available for trading.

In conclusion, PCA can be leveraged to construct meaningful pairs trading strategies, and the principal eigenportfolio can serve as a reliable synthetic factor portfolio when assets are grouped into meaningful groups.

6 Conclusion

In this study, we applied the PCA-based statistical arbitrage strategy of Avellaneda (2008) to the cryptocurrency market, incorporating practical extensions such as adopting a dynamic trading universe and rebalancing periodically to extract new statistical factors. Our initial expectation was that a factor-neutral approach would isolate sufficient idiosyncratic mean reversion to yield profitable trades. In reality, the results were notably weaker than anticipated, highlighting that the idiosyncratic components in cryptocurrency returns may not consistently revert to a long-run mean.

Nonetheless, the findings offer several important insights that can inform practitioners and researchers. First, we demonstrated how PCA can serve as a tool for extracting underlying "market-like" factors in a domain where no canonical benchmark index exists. While this is a widely known fact, there is no research that directly integrates the principal eigenportfolio and cointegration approach to construct a pairs trading strategy. Second, we showed that repeated rebalancing—updating both the trading universe and principal factors—can help the model remain relevant in a rapidly changing environment.

Yet the study underscores key limitations. The need to specify numerous hyperparameters (e.g. thresholding for negative shocks, daily or monthly rebalancing windows, etc.) introduces complexity and potential overfitting. Our reliance on daily data and a relatively short out-of-sample horizon also raises concerns about statistical power, given that only a small number of trades were generated. Moreover, it remains an open question whether the assumed mean-reversion property of auxiliary series is valid in cryptocurrency markets, or whether alternative factor models—potentially integrating more fundamental or on-chain data—might do a better job of capturing systematic behavior.

In theory, there is no fundamental reason why the movements of cryptocurrency markets should be intrinsically tied to other asset classes. Unlike traditional financial instruments, there is no universally accepted model for determining the intrinsic value of a cryptocurrency. As a result, the key drivers of the cryptocurrency market are highly dynamic, shifting from one regime to another and evolving with changing narratives. Given this inherent fluidity, it is logical to rely on statistical tools to analyze market behavior—this is where techniques such as PCA can offer valuable insights into the ever-changing market dynamics of cryptocurrencies. While this study is neither exhaustive nor groundbreaking, we hope it serves as a source of curiosity, providing practitioners, academics, and students with both entertainment and a glimpse of insight into this complex and evolving landscape.

Appendix

Appendix A: Rebalancing and Weighting Mechanism

This appendix provides a detailed description of the daily rebalancing logic and weight-updating scheme used in our PCA-based statistical arbitrage strategy.

A.1 Setup and Notation

Price and Return Matrices. Let

$$\mathbf{P} \in \mathbb{R}^{(n+1)\times p} \quad \text{and} \quad \mathbf{R} \in \mathbb{R}^{n\times p}$$

respectively, denote the daily close-price matrix and the simple-return matrix of p coins observed over $n+1$ days. To be explicit, note that $p = p_{t_s} = \#mrl_{t_s}$ where $t_s \in$ {rebalance dates}.

In matrix form:

$$\mathbf{P} = \begin{bmatrix} P_{0,1} & \dots & P_{0,p} \\ P_{1,1} & \dots & P_{1,p} \\ \vdots & \ddots & \vdots \\ P_{n,1} & \dots & P_{n,p} \end{bmatrix}, \qquad \mathbf{R} = \begin{bmatrix} R_{1,1} & \dots & R_{1,p} \\ \vdots & \ddots & \vdots \\ R_{n,1} & \dots & R_{n,p} \end{bmatrix},$$

where

$$R_{t,i} = \frac{P_{t,i} - P_{t-1,i}}{P_{t-1,i}} \quad \text{for} \quad t = 1,\dots,n, \quad i = 1,\dots,p.$$

We index time $t = 0$ as the initial date, so $t = n$ is the final date.

Factor Decomposition & Hedge Ratios. Each coin $i \in \{1,\dots,p\}$ has a factor-exposure vector $\boldsymbol{\beta}_i = \left(\beta_{i,1},\dots,\beta_{i,5}\right)^\top$ relative to the top 5 PCA factors (eigenportfolios). In a fully *factor-neutral* position, if we go \$1 *long* (or short) in coin i, we must go $\beta_{i,j}$ dollars *short* (or long) in factor j, for $j = 1,\dots,5$. The eigenportfolios themselves are linear combinations of the p coins, so the net result is a position in each of the p coins (some possibly long, some short).

A.2 Constructing the (Daily) Weight Vector

Signal Generation. On each trading day t, for coins in mrl_t, we decide whether a signal is generated for some coin k. The signal is discrete; $+1$ signifies undervaluation, -1 overvaluation, and 0 neutral. Denoting this integer by $\text{signal}_{k,t} \in \{-1, 0, +1\}$, we collect all coin signals into a vector $\mathbf{sig}_t = [\text{signal}_{1,t},\dots,\text{signal}_{p,t}]^T$.

Combining Coin + Factor Positions. If $\text{signal}_{k,t} = +1$, we take \$1 *long* in coin k plus $\$(-\beta_{k,j})$ in factor j. However, recall a factor itself is a basket of coins. Denote by

$$\mathbf{Q}_t = \begin{bmatrix} w_{1,t}^{(1)} & w_{1,t}^{(2)} & \dots & w_{1,t}^{(5)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{p,t}^{(1)} & w_{p,t}^{(2)} & \dots & w_{p,t}^{(5)} \end{bmatrix}$$

the volatility-weighted matrix for the 5 eigenportfolios at day t. Each column corresponds to an eigenportfolio; each row $(i, \cdot)$ gives the fraction of dollar exposure on coin i. Then shorting $\beta_{k,j}$ dollars of factor j translates to

$$-\beta_{k,j} \times \text{column } j = -\beta_{k,j} \begin{bmatrix} w_{1,t}^{(j)} \\ \vdots \\ w_{p,t}^{(j)} \end{bmatrix}$$

Summing over $j = 1, \dots, 5$ yields a position of:

$$\sum_{j=1}^{5} -\beta_{k,j} \begin{bmatrix} w_{1,t}^{(j)} \\ \vdots \\ w_{p,t}^{(j)} \end{bmatrix} = -\mathbf{Q}_t \boldsymbol{\beta}_k$$

which is a length-p vector specifying *how many dollars* are shorted (negative entry) or longed (positive entry) in each coin i. Adding the $\$+1$ long in coin k modifies the kth entry by $+1$. This logic is reversed if $\text{signal}_{k,t} = -1$ (short coin k, long factor). Since we iterate this process for all coins $i \in \{1, 2, \dots, p\}$, we will have at most p non-zero vectors denoting the dollar position in each coin. Summing over *all* traded coins forms the *aggregate* net-exposure vector

$$\mathbf{v}_t \in \mathbb{R}^p$$

with entries $\mathbf{v}_t(i)$ denoting net dollar exposure to coin i.

Normalizing and Risk-Free Asset. We scale $\mathbf{v}_t$ so that its *sum of absolute values* equals a chosen leverage $\mathcal{L}$. Specifically,

$$\mathbf{v}_t' = \frac{\mathbf{v}_t}{\sum_{i=1}^{p} \left|\mathbf{v}_t(i)\right|} \times \mathcal{L}$$

If $\sum_{i=1}^{p} |\mathbf{v}_t(i)| = 0$ (i.e. no signals), we default to $\mathbf{v}_t' = \mathbf{0}$. Next, we introduce a risk-free (cash) component $w_{\text{rf},t} = 1 - \sum_{i=1}^{p} \mathbf{v}_t'(i)$, ensuring that the *net* portfolio weights sum to 1 (if we prefer a fully invested approach). The final weight vector becomes

$$\mathbf{w}_t = \begin{bmatrix} w_{\text{rf},t} \\ \mathbf{v}_t'(1) \\ \vdots \\ \mathbf{v}_t'(p) \end{bmatrix} \in \mathbb{R}^{(1+p)}$$

A.3 Daily Portfolio Valuation, Transaction Cost, & Rebalancing

Portfolio Return. Let $\mathbf{w}_{t-1}$ be the *final* weight vector chosen at day $t-1$. Over the interval $(t-1, t]$, each coin i realizes a return $R_{t,i}$. The total P&L for day t is

$$\mathrm{PnL}_t \;=\; \Big(\sum_{i=1}^{p} w_{t-1}(i)\, R_{t,i}\Big) \times\; E_{t-1}$$

where E_{t-1} is the equity at the close of day $t-1$, and $w_{t-1}(0)$ is the (risk-free) remainder with zero return (assuming no interest). Then the equity at day t prior to transaction costs is

$$E_t^{(\mathrm{pre})} \;=\; E_{t-1} + \mathrm{PnL}_t$$

Transaction Cost. After computing PnL_t, we impose a cost based on *turnover* whenever we shift from $\mathbf{w}_{t-1}$ to $\mathbf{w}_t$. Specifically, if the sum of absolute differences

$$\mathrm{turnover}_t \;=\; \sum_{i=1}^{p} \big|\, w_t(i) - w_{t-1}(i)\big|$$

is nonzero, then we subtract

$$\mathrm{Cost}_t \;=\; \mathrm{turnover}_t \;\times\; E_t^{(\mathrm{pre})} \;\times\; (\mathrm{bps}/10{,}000)$$

where bps refers to taker fee in basis points. In this simplified model, market impact and bid-ask spread is not considered. The net equity becomes

$$E_t \;=\; E_t^{(\mathrm{pre})} \;-\; \mathrm{Cost}_t.$$

Forced Closure on Rebalancing Days. On specified rebalancing dates (e.g. monthly), we forcibly set the old position to zero weight, realize the P&L, and pay transaction costs on the full absolute exposure from $\mathbf{w}_{t-1}$. We then open new positions $\mathbf{w}_t$ if the strategy signals require it, incurring another cost. This ensures a "clean slate" each rebalance day and is especially important if the set of tradable coins changes (due to dynamic filtering).

In summary, this rebalancing approach aims to preserve the "factor-neutral" property of the strategy, even as individual coin returns drift. Whenever new signals emerge or rebalancing occurs, we recompute the final weight vector, aggregate all coin/factor exposures, and debit the transaction cost from the daily equity.

References

[1] Avellaneda, Marco, and Jeong-Hyun Lee. *Statistical Arbitrage in the U.S. Equities Market.* June 2008.

[2] Chan, Ernest P. *Algorithmic Trading: Winning Strategies and Their Rationale.* Hoboken, N.J., John Wiley & Sons, Inc, 2013.

[3] Do, Binh Huu, et al. "A New Approach to Modeling and Estimation for Pairs Trading." Research Gate, Jan. 2006. `www.researchgate.net/publication/267721035_A_New_Approach_to_Modeling_and_Estimation_for_Pairs_Trading`

[4] Gatev, Evan, et al. "Pairs Trading: Performance of a Relative-Value Arbitrage Rule." *Review of Financial Studies*, vol. 19, no. 3, 2006, pp. 797–827. `www-stat.wharton.upenn.edu/~steele/Courses/434/434Context/PairsTrading/PairsTradingGGR.pdf`, `https://doi.org/10.1093/rfs/hhj020`.

[5] Gillespie, Daniel T. "Exact Numerical Simulation of the Ornstein-Uhlenbeck Process and Its Integral." *Physical Review E*, vol. 54, no. 2, 1 Aug. 1996, pp. 2084–2091. `pdfs.semanticscholar.org/57a2/3a111b40528af0cb1ba86f90d75e8d11a21d.pdf`, `https://doi.org/10.1103/physreve.54.2084`.

[6] Ledoit, Olivier, and Michael Wolf. "Honey, I Shrunk the Sample Covariance Matrix." *The Journal of Portfolio Management*, vol. 30, no. 4, 31 July 2004, pp. 110–119. `https://doi.org/10.3905/jpm.2004.110`.

[7] Markwick, Dean. "Stat Arb - an Easy Walkthrough." Dean Markwick, 15 July 2023. `dm13450.github.io/2023/07/15/Stat-Arb-Walkthrough.html`.

[8] Shah, Agam, et al. "Principal Component Analysis Based Construction and Evaluation of Cryptocurrency Index." *Expert Systems with Applications*, vol. 163, Jan. 2021, p. 113796. `www.sciencedirect.com/science/article/pii/S0957417420306151`, `https://doi.org/10.1016/j.eswa.2020.113796`.

[9] Shlens, Jonathon. *A Tutorial on Principal Component Analysis.* Apr. 2014. `arxiv.org/abs/1404.1100`, `https://doi.org/10.48550/arXiv.1404.1100`.

[10] Soo, Carolyn, et al. *MS&E448 Project Statistical Arbitrage.* 11 June 2017. `stanford.edu/class/msande448/2017/Final/Reports/gr5.pdf`.

[11] Vidyamurthy, Ganapathy. *Pairs Trading.* John Wiley & Sons, 2 Feb. 2011.

Disclaimer

This report is for informational and academic purposes only. It is not investment advice and does not constitute any offer or solicitation to offer or recommendation of any investment product. Nothing contained in this report constitutes a solicitation, recommendation, endorsement, or offer by the author or any third-party service provider to buy or sell any securities or other financial instruments in this or in any other jurisdiction in which such solicitation or offer would be unlawful under the securities laws of such jurisdiction. All content in this report is information of a general nature and does not address the circumstances of any particular individual or entity. Nothing in the report constitutes professional and/or financial advice, nor does any information in this report constitute a comprehensive or complete statement of the matters discussed or the law relating thereto. The author is not a fiduciary by virtue of any person's use of or access to this report. You alone assume the sole responsibility of evaluating the merits and risks associated with the use of any information in this report before making any decisions based on this report. In exchange for accessing this report, you agree not to hold the author, its affiliates, or any third-party service provider liable for any possible claim for damages arising from any decision you make based on information made available to you through this report.