Why Using GLS and PCA Creates an Edge in Cryptocurrency Futures Trading

Eber Sluis (571807), Hugo Groenewegen (542628), Efe Kaya (568899) and
Menno Smit (580595)

(Group D)

EI faculty supervisor:	dr. Rutger-Jan Lange, Associate Professor
Firm supervisor:	Arnout Tilgenkamp
Date final version:	25th April 2025

Abstract

This paper investigates whether forward-curve modeling of Bitcoin (BTC) and Ethereum (ETH) futures yields a tradable edge, using second-interval data from Binance, OKEX, and Deribit (June 2023–September 2024). We model log futures premia with liquidity, spread, average prices, and tenor-based features, estimated via ordinary least squares (OLS), weighted least squares (WLS), generalized least squares (GLS), and robust regression (RR), all evaluated by residual mean reversion. We construct two approaches of capital allocation; equal-weighted and directionally. The latter adjusts the positions to the relative sizes of the mispricings. Furthermore, we use a benchmark to select the two most under- and overpriced assets. In addition to the benchmark, we impose several adjustments for two other strategies. The GLS model with the first component of PCA shows the highest mean reversion across all the models. Further showing strong profitability with a profit and loss (PnL) to maximum capital employed of 38.8% per annum. Tenor-adjusted strategies underperform due to exaggerated mispricing resulting from short-term bias. Further, we find that most of the PnL is generated by instruments with longer maturities, due to more mispricing on this end of the forward curve. Finally, combining models reduces variance and model risk. This makes the directional allocation more profitable than equal-weighted since we can increase our risk. Yet, using the best model equal-weighted remains more profitable. Our findings suggest that feature selection and trade-entry conditions enhance profitability, linking traditional modeling to the cryptocurrency futures market and offering traders actionable signals.

Contents

1 Introduction

Over the past decade, the cryptocurrency market has evolved from a niche speculative asset class into a significant component of global financial markets. Bitcoin (BTC) and Ethereum (ETH), the two largest digital assets by market capitalization, have become key instruments for both retail and institutional investors. However, their high volatility, fragmented liquidity, and complex market dynamics present unique challenges to accurate pricing and risk management (Cheng et al., 2025). In traditional finance, futures curves are widely used to extract information on expected returns, risk premia, and arbitrage opportunities (De Roon et al., 1998; Bianchi et al., 2023). Yet, in cryptocurrency markets, the pricing structure of futures contracts remains less explored, with limited academic research addressing whether these markets exhibit systematic and predictable patterns in their term structures (Almeida and Gonçalves, 2024).

From a theoretical perspective, this research is relevant for the ongoing discussion of market efficiency in cryptocurrency derivatives, addressing whether these markets adhere to traditional financial theories or exhibit unique characteristics that create persistent arbitrage opportunities (Mokni et al., 2024; Alexander et al., 2024). From a practical point of view, our findings have direct trading implications for market makers, quantitative hedge funds, and institutional investors looking to optimize execution strategies and risk management in cryptocurrency futures markets. This is why the following question is answered in this research:

Can curve fitting Bitcoin and Ethereum futures prices create an edge in trading?

To address this question, we model the log-forward premium using our base model. The base model consists of a relation between the log-forward premia and a tenor. This relation is derived from the transformed no-arbitrage pricing condition for futures and should hold under the perfect market hypothesis (PMH). Since this does not always hold for cryptocurrencies, we add features to adjust for deviations from the PMH, both directly and via principal component analysis (PCA). In doing so, we answer the following sub-questions: 'To what extent does the cryptocurrency market satisfy the theoretical no-arbitrage condition?', and; 'Which additional features and models can approximate the deviations from the PMH?'. These questions are answered based on the average mean reversion of the residuals of the models, where more mean reversion implies that we can better leverage our fitted curves. After answering these questions, the best models are used in trading strategies, where profit and loss (PnL), among other metrics, will support answering the main research question.

The base model and its extensions are estimated using four types of regressions; ordinary least squares (OLS), weighted least squares (WLS), generalized least squares (GLS), and robust regression (RR). OLS serves as the baseline but is too strict on the error assumptions considering volatile cryptocurrency markets. WLS weights observations by the inverse of their respective residual variances. GLS further generalizes this by additionally capturing heteroscedasticity. Finally, RR provides a framework where we better handle the presence of outliers, which is in our interest as outliers are our mispricings. The data consists of second interval prices from Binance, OKEX, and Deribit, from June 13, 2023, up to September 1, 2024. These are the three most liquid cryptocurrency exchanges (CoinMarketCap, 2025). It includes a variety of order

book information such as bid-ask spreads and mid-prices for spot assets, perpetuals, and futures (for a variety of maturities).

After jointly fitting these models for BTC and ETH, we select the optimal model combinations in terms of features and least squares (LS) methods. We find that the GLS models perform best in terms of mean reversion of residuals, due to their ability to capture the deviations from the assumptions of OLS. We take this one step further, by applying PCA to the features used by the GLS method. This leads on average to a higher mean reversion, therefore effectively reducing variance, without creating too much bias.

We consider three different sets of trading rules. First, a benchmark strategy, always trading with the two most under- and overpriced instruments (BM). Secondly, a modified strategy, that scales for the maturity of the instrument and imposes a minimum value for the residual for the instrument to be traded on (MOD), and finally the best instrument strategy, that only trades on the instruments with below average half-life times (BI).

We also make a distinction between types of allocation, namely an equal-weighted (EW) and directional (DI) one. In the EW we take short and long positions with an equal amount, whereas using DI we take the relative mispricings into account, which allows us to have a net long or net short position.

Finally, we combine these 3×2 options into six different trading strategies. The trading strategies lead to a significant profit throughout our dataset. During our trading window, we obtain PnLs ranging up to \$1.35 million based on a maximum capital employed of \$2.7 million, achieving a return on maximum exposure of 38.8% per annum.

The remainder of the paper is structured as follows. Section 2 covers an in-depth review of this subject in the literature. Sections 3, 4, and 5 will provide the set-up for our research, with an explanation of the data, methodology, and trading strategies used respectively. Section 6 states the obtained results in our framework. Finally, we discuss the outcomes in a broader context and suggest improvements for further research.

2   Literature review

The rapidly evolving cryptocurrency market is almost indispensable in investors' portfolios (Kajtazi and Moro, 2019). This is the reason why many researchers try to find profitable trading strategies. This started with Makarov and Schoar (2020), focusing on finding arbitrage within the different market dynamics of the Bitcoin, and it has evolved into pairs-trading strategies, that focus on mean reversion of forecast residuals (Fil and Kristoufek, 2020). However, as Corbet et al. (2018) and Fassas et al. (2020) found, the exploration of the cryptocurrency futures market and their pricing structures remains limited. This is primarily due to the lack of comprehensive data and the high volatility that complicates model accuracy and consistency.

Our research addresses this gap by investigating the degree to which different curve-fitting models create alpha when applied to BTC and ETH futures prices. We add features based on liquidity and spread to the curve-fitting model. This builds on Akyildirim et al. (2023), who show that there is a unique opportunity to find in this trading field, due to the inherent characteristics and continuous trading of cryptocurrencies. The reason why this might lead to profitable trading is that cryptocurrencies often jump together, as demonstrated by Bouri

et al. (2020). This means that trading on the mean reversion decreases the uncertainty of the unpredictable move of the individual instruments and leaves only the predictable part of the difference behind.

This research contributes to the literature in two ways. First, it tries to diminish the concerns raised by Liu and Tsyvinski (2020), who raised important questions about whether these markets adhere to traditional yield curve theories. It does so by successfully using conventional techniques on the unconventional BTC and ETH. Second, on a practical level, similar to Zhang et al. (2023), we highlight the usefulness for market participants, such as traders and market makers, in optimizing their execution strategies and profits in these evolving markets.

Even when having a curve fit on the BTC and ETH futures prices with relatively high mean-reversion of the residuals, it remains a difficult task to create a profit in the out-of-sample period, due to the speculative nature of the coins (Cheng et al., 2025). Aharon et al. (2021) achieve mixed results with their method in the out-of-sample forecasting period. And Malladi and Dheeriya (2021) only generate profit when disregarding the transaction costs. Alexander and Dakos (2020) try to overcome this, and eventually conclude that it requires highly sophisticated models to predict BTC returns.

To tackle these challenges, we use evidence from Sathyanarayana and Gargesa (2019) that the theoretical no-arbitrage pricing condition can be utilized. They do this by effectively selecting the models with the best forecasting performance. In this paper, this will be taken one step forward by integrating it into an effective trading strategy. We find more support in the selection of our features based on liquidity and spread in the work of Orte et al. (2023). They show that technical indicators, like ours, are significant in their machine-learning models. After feature selection, both liquidity and spread are still present in the models.

3 Data

The dataset is obtained from Maverick Derivatives and consists of exchange-API collected spot, futures, and perpetual prices for BTC and ETH, for a variety of different maturities. Our analysis focuses solely on current spot and futures instruments, while perpetuals are excluded. The futures, which have maturities that are weekly, monthly and quarterly, update with the release (expiry) of new (old) instruments. These futures give the set of current instruments I_t. This set is updated at each second from June 13th, 2023, to September 1st, 2024, covering a total of 446 days.

The data is sourced from three major cryptocurrency exchanges: Binance, OKEX, and Deribit. These exchanges are among the most liquid and dominant in the market, making them representative of the broader cryptocurrency ecosystem (CoinMarketCap, 2025). The structure of the dataset and how it is split into separate time series is visualized in the Appendix, Figure 10.

The dataset is sampled at a one-second frequency, with each observation containing a timestamp rounded up to the nearest second. Additionally, it includes order book information, such as bid and ask-prices, order sizes, trading volumes, and mid-point prices.

The dataset consists of approximately one billion rows and is stored in 446 Parquet files which we concatenated to 35 Parquet files to make it easier to work with. Due to its large size

5

(exceeding 75 GB), handling this dataset requires high-performance computational resources with excessive RAM. We utilized a 128GB RAM virtual machine from Google's Vertex AI for this. To optimize memory usage, the dataset is processed in batches of four Parquet files, covering roughly 40 days each, during aggregation.

To be able to make forecasts once per hour, we aggregate the data into hourly intervals by selecting the last observed price at the end of each hour. This approach is chosen because using average or median prices could result in non-tradeable values, which would lead to inconveniences when converting the model into a real-world trading strategy. The descriptive statistics of the aggregated data are shown in Table 1, and the time series of hourly BTC and ETH spot return series can be found in the Appendix, Figure 11.

Table 1: Descriptive statistics of hourly BTC and ETH spot return series in percentages

	Count	Mean	Std	Min	25%	50%	75%	Max	Skewness	Kurtosis
BTC	10511	0.01	0.51	-6.12	-0.18	0.01	0.20	5.55	-0.15	14.74
ETH	10511	0.01	0.60	-8.32	-0.22	0.01	0.24	9.29	-0.17	23.50

Notes: Count = number of observations, Std = standard deviation, Min = minimum value, 25%/50%/75% = quantile values, Max = maximum values.

The dataset is complete, meaning there are no missing observations in terms of time intervals. However, during periods of low market activity, prices may remain unchanged for extended periods. Because this is tick data, no new entries are recorded when there are no price changes, so no interpolation is necessary, and these cases are handled naturally during data cleaning by choosing the last available observations as the current observation.

Regarding outliers, no specific checks or filtering are performed. We do not remove or winsorize extreme price movements, as they are considered an inherent part of market behavior. Since data formats differ between exchanges, pre-processing is necessary. We standardize timestamps from different exchanges to a uniform time zone (CET) to ensure consistency across all observations.

4 Forward Curve Fitting

In this section, we discuss the implementation of fitting a curve through the futures premia. Firstly, we use the no-arbitrage price of the futures to derive the base model. Then, we discuss the theoretical conditions for this model and propose several additional features to correct for deviations from these conditions. Subsequently, we evaluate the use of PCA in extracting orthogonal features. Finally, we cover the model selection procedure, which is based on the mean reversion of the residuals.

4.1 Forward curve base model

We begin with the theoretical no-arbitrage price of a future at time t (Baxter and Rennie, 1996)

$$F_{t,i(\tau)} = S_t \cdot e^{r_t \cdot \tau} \quad \Leftrightarrow \quad p_{t,i(\tau)} := \log(F_{t,\tau}/S_t) = r_t \cdot \tau. \tag{1}$$

6

Here, $F_{t,i(\tau)}$ is the futures mid-price of instrument $i \in I_t$ with time-to-maturity or tenor $\tau > \frac{1}{365.25}$ years (one day). Furthermore, $S_t > 0$ is the spot price, and r_t is the theoretical risk-free rate per annum. The restriction of τ greater than one day is set to avoid instruments with extreme volatility and outliers, also known as the 'Samuelson effect' (Chevallier, 2012). To simplify modeling, we assume that r_t is constant across time-to-maturity, which is an often made assumption in financial analysis, as used in, for example, Black and Scholes (1973). One can see that the futures price is a function of the spot price of the underlying, the interest rate, and the tenor. Here, we omitted the cost of carry and income yield, or dividend, which are theoretically added and subtracted, respectively, to and from the interest rate (Chen, 2025). This can be done as there is no standardized carrying cost or dividend for cryptocurrencies. By expressing it as a log-linear relation and incorporating an intercept and error term, we obtain the base model for estimating the forward premium $p_{t,i(\tau)}$ as

$$p_{t,i(\tau)} = \beta_{0,t} + \beta_{1,t} \cdot \tau + \varepsilon_{t,i(\tau)}, \quad \varepsilon_{t,i(\tau)} \overset{\text{i.i.d.}}{\sim} (0, \sigma_\varepsilon^2). \tag{2}$$

Here, we expect the estimated coefficients to be $\hat{\beta}_{0,t}^{OLS} = 0$ and $\hat{\beta}_{1,t}^{OLS} = r_t$, as per the no-arbitrage pricing theory. The risk-free rate r_t is usually stable over short intervals but could change significantly at each second t. Furthermore, $\varepsilon_{t,i(\tau)}$ is the error of the equation at time t for instrument $i \in I_t$, which is ideally normally distributed. This is, however, not a necessity when estimating it with OLS. In our case, $\varepsilon_{t,i(\tau)}$ is the mispricing compared to our theoretical price.

To give an intuition, when we estimate the base model at the first timestamp in the data, we obtain Figure 1. Looking at the regression line of the OLS curve through the log premia, there seems to be a reasonable fitting of the general trend of the data. The R^2 of 0.93 confirms this. The value of the $\beta_{1,1}$ coefficient (slope) here is 0.028. This can be interpreted as a 2.8% increase of the log(premium) when the maturity increases by one year. The slope coefficient is not constant over time but increases slowly and steadily to 0.045 in the first month of data, as can be seen in Figure 12 in the Appendix. In this same figure, it becomes clear that the intercept coefficient is indeed close to zero for all timestamps, leading us to impose the intercept to be zero. We do, however, leave the intercept in the regression analysis, as we are trading futures only and do not force our fitted line through the origin. In the next 14 months of the data, the slope coefficient increases even further towards a high point of 0.18 in April 2024, after which it declines again (see Figure 15). Given that we expect the slope to be the risk-free rate, it is no surprise that the US 10-year treasury rate also has a peak in April 2024. This is a confirmation of the validity of our base model setup.
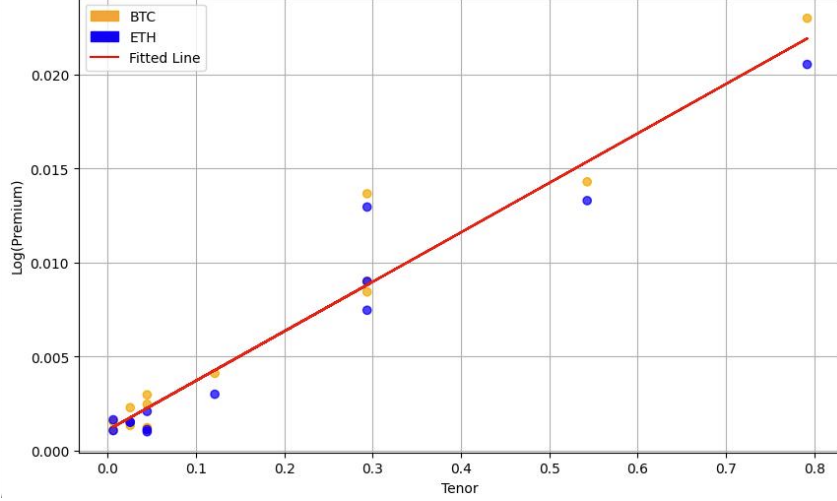
Figure 1: BTC and ETH log premia plotted over maturities (years), including the OLS fitted curve. The regression formula for this timestamp is: $p = 0.0012 + 0.028\tau$, with $R^2 = 0.93$

When estimating β_t, least squares methods minimize the objective function

$$\min_{\beta_t} \sum_{i,j \in I} w_{t,(i,j)(\tau)} \left(p_{t,i(\tau)} - x_{t,i(\tau)}\beta_t \right) \left(p_{t,j(\tau)} - x_{t,j(\tau)}\beta_t \right), \tag{3}$$

in which $x_{t,i(\tau)}^{(1 \times k)}$ is the regressor vector, and $w_{t,(i,j)(\tau)}$ the weight given to the multiplication of the two observation errors, all for time t and instruments $i, j \in I_t$.

When utilizing OLS, this comes down to the restriction

$$w_{t,(i,j)(\tau)} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases} \tag{4}$$

in Equation 3, which simplifies the expression. This is useful under the assumptions described by Heij et al. (2004). These seven assumptions are evaluated in their mentioned order by several tests, all performed on a 5% significance. We test

- $\langle A1 \rangle$ exogeneity by means of the test from Hausman (1978).
- $\langle A2 \rangle$ a zero error mean by means of a two-sided t-test.
- $\langle A3 \rangle$ homoscedasticity of the errorsby means of the test from Breusch and Pagan (1979).
- $\langle A4 \rangle$ no autocorrelation by means of Durbin and Watson (1950).
- $\langle A5 \rangle$ constant OLS parameters by means of the test from Quandt (1960).
- $\langle A6 \rangle$ linearity by means of the Ramsey (1969) Reset test.
- $\langle A7 \rangle$ normal distributed errors by means of Jarque and Bera (1980).

When assumptions $\langle A1 \rangle, \langle A2 \rangle, \langle A5 \rangle$, and $\langle A6 \rangle$ are fulfilled, the OLS estimator is unbiased (Heij et al., 2004). If additional assumptions $\langle A3 \rangle$ and $\langle A4 \rangle$ hold, by means of Gauss-Markov, the estimator is the best linear unbiased estimator (BLUE). Finally, if all assumptions hold, OLS gives the uniformly minimum variance unbiased estimator (UMVUE).

When $\langle A3 \rangle$ does not hold, this is equivalent to some observations containing more variance than others. WLS can reinstate this assumption by inversely weighing observations with their

error variance, which is, for $i, j \in I_t$,

$$w_{t,(i,j)(\tau)} = \begin{cases} (\Omega_{ii})^{-1}, & \text{if } i = j, \\ 0, & \text{if } i \neq j, \end{cases} \tag{5}$$

in Equation 3. Here, Ω is the covariance matrix of the errors. By inverting Ω, observations with high variance are reduced in weight, this way there is more emphasis on lower variance observations (Heij et al., 2004).

When both $\langle A3 \rangle$ and $\langle A4 \rangle$ are not matched simultaneously, meaning that Ω does not equal an identity matrix $\Omega = I$, GLS can correct for this directly. It does so by multiplying the OLS regression on the left-hand side with the upper diagonal L', where $\Omega = LL'$ according to Cholesky's decomposition (Cholesky, 2005). In Equation 3 this comes down to

$$w_{t,(i,j)(\tau)} = (\Omega_{ij})^{-1}, \forall i, j \in I. \tag{6}$$

Finally, while traditional OLS, WLS, and GLS still work without $\langle A7 \rangle$, they are not specifically designed for heavy-tailed errors. RR is designed to work efficiently for non-Gaussian errors and to more efficiently deal with outliers (Rousseeuw and Leroy, 2003), of which an illustration is given in Figure 14. The latter indirectly also decreases the severity of $\langle A3 \rangle, \langle A4 \rangle$ on the estimation by shrinking the weights of extreme residuals. One popular variant is the Huber Regression described by Huber (1992). The Huber regression iteratively obtains a new k^{th} estimate $\hat{\beta}_t^k$, after which it calculates $\hat{\varepsilon}_{t,i(\tau)}^k = p_{t,i(\tau)} - x_{t,i(\tau)}\hat{\beta}_t^k$. Subsequently, this is used to define the weight in Equation 3, for $i, j \in I_t$, as

$$w_{t,(i,j)(\tau)} = \begin{cases} 0, & \text{if } i \neq j, \\ \frac{1}{2}, & \text{if } i = j \text{ and } |\hat{\varepsilon}_{t,i(\tau)}^k| \leq \delta, \\ \frac{\delta\left(|\hat{\varepsilon}_{t,i(\tau)}^k| - \frac{\delta}{2}\right)}{(\hat{\varepsilon}_{t,i(\tau)}^k)^2}, & \text{if } i = j \text{ and } |\hat{\varepsilon}_{t,i(\tau)}^k| > \delta. \end{cases} \tag{7}$$

in which δ is the threshold value. We set this to $\delta = 1.345$ as this balances out efficiency of estimation and robustness (Huber, 1992). Clearly one can see that for small residuals the loss is quadratic, while for larger residuals linear. For the first estimate $\hat{\beta}_t^1$, the OLS estimator is used. We utilize a maximum iteration of 50. Furthermore, we use a convergence rate, meaning the difference in log-likelihood values between iterations until considered converged, of 10^{-4}.

Testing the assumptions for the base model for the first month of the data set gives Table 2. Here, one can see that the percentages of violations of $\langle A1 \rangle, \langle A2 \rangle, \langle A5 \rangle, \langle A6 \rangle$ are low, indicating there is not enough proof to reject these over time at each second. On the other hand, $\langle A3 \rangle, \langle A4 \rangle$ are significantly violated at least half of the time which signals that GLS might be effective. Finally, $\langle A7 \rangle$ is violated about 37% of the time, indicating that the residuals are often not normally distributed.

Table 2: OLS Assumptions Violations for Base Model

Assumption	$\langle \mathbf{A1} \rangle$	$\langle \mathbf{A2} \rangle$	$\langle \mathbf{A3} \rangle$	$\langle \mathbf{A4} \rangle$	$\langle \mathbf{A5} \rangle$	$\langle \mathbf{A6} \rangle$	$\langle \mathbf{A7} \rangle$
Base Model Violation (%)	0.00	8.21	55.85	65.61	0.00	27.69	36.59

Notes: The results stem from the first month of the data set, calculating statistics every hour on the regression, and then averaging the number of violations over each point in the data. In our regressions, the number of observations is small, resulting in a low power of the corresponding tests.

Importantly, the no-arbitrage pricing theory utilized in Equation 2 requires the PMH conditions to hold as described by Ritchken and Boenawan (1990). There should be (1) no transaction costs, (2) equal debit and credit interest rates, (3) no taxes, (4) allowance for short-selling, and (5) unlimited divisibility of assets. The first and second points are violated, while the third point is violated based on the country. Short-selling is allowed; divisibility is not unlimited, but many brokers offer partial coins, which proxies this condition. As the conditions do not fully hold, this could lead to distortions from the no-arbitrage condition and therefore the base model, which would require correction. This can be assessed based on the mean-reversion properties of the base model residuals, which is explained in Section 4.4.

4.2 Additional features

In addition to the base model features in Equation 2, we add features that explain a deviation from the usual no-arbitrage condition and can enhance the mean reversion of residuals. These features are contained in the feature set $F = \left\{ \tau \cdot \mathbb{I}(\tau < \tau_c), \overline{\text{midp}}_{t,i(\tau)}, \overline{\text{liq}}_{t,i(\tau)}, \overline{\text{Spr}}_{t,i(\tau)} \right\}$.

1. $\tau \cdot \mathbb{I}(\tau < \tau_c)$: For most investors, the debit interest rate exceeds the credit interest rate, meaning that their spread $r_d - r_c$ exceeds zero, leaving condition (2) from the PMH unfulfilled. As a result, $F^d_{t,i(\tau)} := S_t \cdot e^{r_{d,t} \cdot \tau} \geq S_t \cdot e^{r_{c,t} \cdot \tau} := F^c_{t,i(\tau)}$, which results in a higher than theoretical futures price than under the credit interest rate. In principle, this is covered by a higher parameter $\beta_{1,t}$ in Equation 2. What is not covered, however, is a change in spread over maturity, which occurs when both r_c and r_d follow distinct yield curve structures based on different risk profiles. This can especially give significant changes over a longer horizon when expectations start to differ significantly (Edelberg and Marshall, 1996). To keep our models focused on estimating the underlying relationship under the PMH without distortions from this changing spread, we separately estimate the slope of the short to middle end and the long end of the curve. This leads to the feature $\tau \cdot \mathbb{I}(\tau < \tau_c)$, in which τ_c determines the point where the slope changes. We set $\tau_c = \frac{2}{3}$ years as this was in our data set typically in between the middle and long end of the forward curve as can be seen in Figure 1.

2. $\overline{\text{midp}}_{t,i(\tau)}$: The average mid-price per hour is used, $\overline{\text{midp}}_{t,i(\tau)}$ with $i \in I_t$, which could proxy a structural deviation from the no-arbitrage price. For example, some countries have legal trading restrictions towards the exchanges that offer the lowest transaction costs, which causes a different price per exchange and therefore a structural deviation across exchanges (Ba and Ömer Faruk Şen, 2024). The mid-price at time t for instrument $i \in I_t$ is calculated

as

$$\text{midp}_{t,i(\tau)} = \frac{A_{t,i(\tau)} + B_{t,i(\tau)}}{2}, \tag{8}$$

in which $A_{t,i(\tau)}, B_{t,i(\tau)}$ are the best bid and ask-prices respectively. Subsequently, the average mid-price is used over the last hour.

3. $\overline{\text{liq}}_{t,i(\tau)}$: The average liquidity per hour, $\overline{\text{liq}}_{t,i(\tau)}$ with $i \in I_t$, explains why the futures do not equal the typical no-arbitrage price. It is useful as it explains that a less liquid future reverts slower to the no-arbitrage price and therefore explains a temporal deviation. Liquidity at time t for instrument $i \in I_t$ is here calculated as

$$\text{liq}_{t,i(\tau)} = A_{t,i(\tau)} \cdot V^a{}_{t,i(\tau)} + B_{t,i(\tau)} \cdot V^b{}_{t,i(\tau)}, \tag{9}$$

in which $V^a{}_{t,i(\tau)}$ and $V^b{}_{t,i(\tau)}$ represent the highest bid and ask-volumes respectively. Subsequently, the average liquidity is used over the last hour.

4. $\overline{\text{Spr}}_{t,i(\tau)}$: Finally, the average spread per hour, $\overline{\text{Spr}}_{t,i(\tau)}$ with $i \in I_t$, is used to distinguish between exchanges and instruments where there is less or more agreement about the underlying interest rate and therefore the no-arbitrage price. Our relative spread variable is calculated at each time t for instrument $i \in I_t$ as

$$\text{Spr}_{t,i(\tau)} = 10^4 \cdot \frac{A_{t,i(\tau)} - B_{t,i(\tau)}}{midp_{t,i(\tau)}}, \tag{10}$$

Here, 10^4 is used to denote the spread in basis points. Subsequently, the average spread is used over the last hour.

4.3 Orthogonal Features

We will also make use of PCA, which is a widely used technique for dimensionality reduction in Least Squares (LS) problems. By transforming the original correlated variables into a set of orthogonal principal components, PCA helps retain the most significant information while reducing noise and redundancy. This is particularly beneficial in LS regression when dealing with high-dimensional data, as it mitigates multicollinearity and improves model interpretability. By selecting only the most relevant principal components, PCA ensures a more efficient and stable solution, leading to better generalization and computational efficiency. In general practices, an elbow plot can be used to determine the number of components, as it visualizes the amount of variance explained per component. In our case, however, we choose the number of PCs based on the best mean reversion. Our regression does not make use of a large number of variables, yet we still are interested in seeing if PCA improves mean reversion and to that extent, PnL. Additionally, the number of optimal PCs will indicate how much unique orthogonal information is contained in the features. In our case, PCA is applied to the features in Equation 2 and the additional features F. We run the regression for a different number of components and compute the mean reversion metrics and PnL.

4.4 Assessing power of models by their mean reversion

Due to the number of possible features and combinations used in the models mentioned in Subsection 4.2, we need to determine which models are the most suitable to use for trading strategies. We do this because we only want to focus on the most robust and effective models rather than to try all possibilities directly. Our main approach to assess this, will be via mean reversion. This is a much-used principle in financial markets and trading such as mentioned in Bouri et al. (2020).

Specifically, we are looking at the mean reversion of instruments' residuals with respect to our fitted line. The line is a theoretical value, and the instruments' premia fluctuate around this line reverting to it over time. We assume that the higher the mean reversion of a model, the more suitable it is to be used for a trading strategy. When a model's residuals' have mean reversion, it better captures the over- and underpricing of certain instruments, which one can capitalize on. On the other hand, when mean reversion is not present, this could lead to very negative scenarios. For example, continuously taking short positions of estimated overpriced instruments, without solid evidence that the price will go down, resulting in a possible infinite loss.

In general, prices of financial instruments are continuous processes, just like volatility. However, since we use hourly data we make the process a discrete one. To check the mean reversion of the residuals, we can fit an AR(1) model on the time series and compare the coefficients,

$$\epsilon_{t,i(\tau)} = \rho\epsilon_{t-1,i(\tau)} + v_t, \quad v_t \overset{\text{i.i.d.}}{\sim} (0, \sigma_v^2). \tag{11}$$

Since we want $\epsilon_{t,i(\tau)}$ to revert to zero, we want a low value for ρ. Ideally, it is going towards -1, however, values below 0.7 would already be showing signs of solid mean reversion. This means that the autocorrelation between consecutive residuals is lower, i.e. higher mean reversion. Also, it would mean more under- and overpricing on which we can trade. If ρ would be negative it would mean it will jump from positive (overpriced) to negative (underpriced). Furthermore, v_t are the error terms of the AR(1) regression.

Supportive to the AR(1) coefficient, we use the half-life time of mean reversion, which as the name suggests is the time it takes to achieve half the reversion to the mean, and is calculated as

$$t_{\frac{1}{2},AR(1)} = -\frac{\ln(2)}{\ln(\rho)}. \tag{12}$$

This equation uses the aforementioned ρ of the AR(1) model. The half-life, $t_{\frac{1}{2}}$. is positive in ρ which makes sense as a higher ρ means lower mean reversion, hence a longer half-life time of mean reversion. One might notice that in this formula ρ cannot be less or equal to zero as this is not in the domain of the (natural) logarithm. When ρ would be negative it would mean the process immediately reverts, hence no time can be calculated.

Since every model contains numerous series we gather averages of our results, that is, the values for $\bar{\rho}$ and $\bar{t}_{\frac{1}{2},AR(1)}$. Using the value of $\bar{\rho}$ in Equation 12, does not lead to the average of their respective $t_{\frac{1}{2}}$, as it is non-linear in ρ. We rank them per metric and sum the scores. We expect a high correlation between the scores as they are estimating the same principle of mean reversion. In case one might rank high based on $\bar{t}_{\frac{1}{2},AR(1)}$ but have the same $\bar{\rho}$, it indicates that

there is a wider dispersion in instruments for the higher half-life time model. This supports the idea to not only look for trading strategies per model but also per instrument.

For our model selection, we use the first 73 days of our total 446 days dataset as an in-sample set. Since we are not fitting parameters but only select the model settings we do not risk underfitting our model. It could be possible that the most mean reverting model changes over time, however, due to the amount of computing power this research requires, we decided to use a single model based on the first set of 73 days (the first five batched parquet files see Section 3).

From here we select the best model and further examine the mean reversion per instrument. This is because we are not bound by trading every instrument or only one model. Rather we select the instrument-model combinations that are best. Nevertheless, we prefer first looking at models as a whole rather than starting with specific instruments per model, as this is a more robust approach to model feature selection.

5   Trading strategies

Now, the theory is connected to empirics by a corresponding trading strategy.

5.1   Setup of the Trading Strategy

After assessing our model specifications and the obtained residuals, we construct trading strategies. Our starting point is our model framework where we select the LS method and the features we use. Per hour we have a set dependent variables, $p_{t,i(\tau)}$, which are our log premia. We then analyze the actual values of this $p_{t,i(\tau)}$ with respect to our fitted line. A point can be either on, above, or below the line. For the first instance, which in theory is possible but in practice will be very difficult, this instrument is perfectly priced according to our model. Points above the theoretical line, that is, $\epsilon_{t,i(\tau)} > 0$, imply that the asset is overpriced. Hence, these could be considered for a short position, as due to the mean reversion (see Section 4.4). Namely, it is expected that the price reverts to the line (i.e. the price drops). This applies the other way around for negative residuals using the same principle. This idea will be the basis for our trading strategies; however, many decisions and rules must be considered.

Before going over the trading strategies, we clarify the rules we implemented in the logic of our profit and loss (PnL) calculations. First, in our model, we calculate our $p_{t,i(\tau)}$ based on the mid-price of the instruments. Yet, in our trading simulation, we decide to only buy for the ask-price and sell for the bid-price. In our regressions, we exclude instruments with a maturity of less than one day to avoid the 'Samuelson effect' (Chevallier, 2012). We follow the same principle in our trading strategy where we force closing a position in an instrument that matures in less than one day.

Secondly, to prevent very large long or short positions, we impose a limit on our maximum exposure per instrument. This mainly comes from a risk management point of view. It could be the case that our model continuously over- or underprices a certain instrument that does not mean revert. Consequently, we would consistently engage in increasing our position which gives much exposure to one instrument (the 'Do not put all your eggs in one basket.' story). Hence, we impose that the exposure cannot be larger than our start capital. This limit can vary and is

dependent on the risk preference of the trader.

Finally, to actually realize PnL, we force closing our position in an instrument when the sign of a residual flips. For example, instrument X had the highest residual in iteration t, therefore we shorted it. In $t+j$, $j > 0$, X's residual is negative. Regardless of the amount of underpricing with respect to the other underpriced instruments, we buy to close this position. This results in our way to realize PnL which will be our lead indicator of how our models perform. The benefit of this approach is that we can better track the mean reversion our model is trading on, as it does not incorporate price changes on our books before closing positions.

5.2 Trading Strategies

To make our trading strategies symmetrical, we always consider the four most mispriced assets, the two highest and lowest residuals in absolute value. From here we consider six different trading strategies, where we split the strategies in two different ways. First, we make a distinction between relative mispricing. Since our residuals are not symmetrical, we put more emphasis on higher absolute values. This brings a trade-off; our model suggests higher mispricing, which we can capitalize on by shorting higher residuals but also gives us an unhedged exposure. An example can be seen in Table 3.

Table 3: Instrument residuals and weightings

Instrument	$\epsilon_{t,i(\tau)}$	Equal-weighted	Directional
A	0.004	25%	40%
B	0.002	25%	20%
C	-0.001	25%	10%
D	-0.003	25%	30%

Notes: Example of trading weights in equally weighted strategies versus directionally weighted strategies. In the equal weighted approach our exposure to instrument i will always be $\frac{1}{2 \times 2}$. In the directional approach , the weight in instrument $i = \frac{|\epsilon_i|}{\sum |\epsilon_j|}$, allowing for directional bets.

Table 3 represents an example of one iteration in our model, where assets A and B are the most overpriced instruments according to our model, and C and D are the most underpriced instruments. In this iteration on the directional approach (DI), we would have a 60% long position and 40% short, allowing us to profit more from our predicted mispricing, but also leaving us exposed to losing (more) money when prices drop. When we do equal weighted (EW), our allocation is always $\frac{1}{2 \times 2}$ regardless of the magnitude among the used residuals.

These two approaches of using the residuals will serve as the benchmark (BM), as no further extensive decision rules are imposed, and purely looks at (relative) mispricing. One of the most notable aspects of this approach is that it almost always trades on the longer end of the curve. Instruments on the longer end of the curve are more mispriced due to less certainty about the future, and thus deviations simply are higher. Furthermore, a negative aspect of our benchmark is that in every iteration it is forced to take a position by construction. This does not necessarily result in a higher PnL as sometimes it might be best to do nothing.

Therefore, to solve these issues, two new conditions are implemented to create the second trading strategy. Firstly, we scale for the maturity of the instrument τ (in years). We use the following transformation: $\tilde{\epsilon}_\tau = (1 + \epsilon_\tau)^{\frac{1}{\tau}} - 1$ where $\tau \in (\frac{1}{365}, 1)$. In this way we can annualize the mispricing and diminish the higher deviations on the long end of the curve. An example is shown in Appendix 12.

Secondly, this second strategy imposes a minimum value of 100 bps that the transformed residual must exceed. This way, in the case that mispricing is not attractive enough, it will skip this trade. This not only ensures that we choose our trades more selectively, but can also adapt for risk appetite, selecting a higher threshold for less risk appetite. This strategy with a minimum deviation and an annualization of residuals will be referred to as 'modified'.

A third and final approach used in this research is only selecting the instruments whose half-life time is lower than the average of the instruments used in the whole set. This way we try to make more use of the mean reversion of certain instruments which we assume leads to higher PnL. This strategy will be referred to as 'best instruments'.

The notation of the trading strategies is now as follows. We consider the set $D = \{EW, DI\}$ to include the direction of the strategy, either equally weighted or directional. Subsequently, we define the set $S = \{BM, MOD, BI\}$, which contains the benchmark, modified, and best instrument-focused strategies. This gives six trading strategies $T = S \times D$, in which '$\times$' refers to the Cartesian product. For example (BM,DI) means the benchmark, directional strategy.

5.3   Performance Metrics

Finally, we have our performance metrics to assess and compare different strategies. We use the following definitions:

Table 4: Summary of performance metrics

Metric	Definition
PnL	$\sum_{t=1}^{|I|} \sum_{i=1}^{N} (P_{\text{close},i,t} - P_{\text{open},i,t}) Q_{i,t}$
Trades	N
PnL per trade	$\frac{\text{PnL}}{N}$
Max drawdown to max employed	$-\max\left(\frac{\text{Peak}_t - \text{Through}_t}{\text{ME}_t}\right)$
Hit ratio	$\frac{PnL_+}{PnL_-}$
Avg. time to realize PnL	$\frac{1}{k} \sum_{i=1}^{k} l_i$, l_i is duration of unchanged PnL
Trading costs	$\gamma \times \sum_{t=1}^{T} \sum_{i=1}^{|I_t|} |q_{i,t}|$

Notes: PnL = profit and loss, trades = amount of trades. Performance metrics used to assess our trading strategies and their corresponding calculations. With $|I_t|$ being the cardinality of the set of instruments I_t, Q_i the size of instruments traded that iteration and $q_{i,t}$ is the absolute value in dollars traded in instrument i at time t.

With PnL being the sum of actual taken profit and losses of all traded instruments and

hit ratio as the total times the PnL increases divided by the total amount PnL decreases. In our metrics, we define maximum drawdown as the biggest drop from a peak throughout the entirety of the PnL. We scale this by the maximum capital employed we have (both long and short), to make fairer comparisons across strategies. This is different than the traditional way to compute maximum drawdown. The difference lies in that PnL is already a result of trading underlying assets, whereas for example, a stock is an underlying itself. We do not make a distinction between a drop from 100 to 50 and 150 to 100 (both decreasing our PnL with 50 units), whereas, for a stock price, this difference is decreasing our portfolio value with 50% or 33.3% respectively.

Furthermore, the average time to realize PnL is the average time it takes to trade on mean reversion, as explained in the final part of Subsection 5.1. Finally, we show our PnL without subtracting trading costs to make fairer comparisons of the causality of mean reversion to PnL; however, we still calculate these trading costs, which is the cost per transaction times the turnover (total $ amount traded). We assume the costs to be one basis point for γ, this assumption is valid for institutional investors.

5.4 Caveats and assumption of our trading strategies

In this subsection, we tackle the assumptions and caveats of our strategy and approach. Starting with gathering the variables, we assume that we can immediately trade at the beginning of hour t using the information of hour $t - 1$. Also, we assume that we can immediately trade on the prices as used in the regressions. As mentioned in Subsection 5.3, we display the PnL without fee, thus omitting every type of cost in our PnL graphs. We do calculate transaction costs, assuming one basis point. Furthermore, we assume that there is no slippage in the price.

We do make a currency assumption. That is we make no distinction between USD and USDT. This is not a viable assumption in a real-life trading strategy as the instruments are not the same. They do however trade approximately par throughout our trading window (see Appendix 13), as a result, our PnL will not show significant differences if two account for the USD/USDT price. Furthermore, for our goal, making a hard distinction and accounting for the difference between the two, is beyond the purpose of this research.

In terms of trading, we calculate our residuals based on the mid-price, but trade on the bid and ask-prices to create realistic PnL calculations. We do not consider the available size at the market at that point. If, for example, we identify mispricing for future X and want to buy $250,000$, we assume we can always execute this trade. For institutional investors this assumption will almost always hold at the sizes we use. Regarding shorting, we assume we can take as many short positions as we want until we hit the predefined cap of our starting capital ($1,000,000) for a specific instrument. Finally, we keep track of our duration risk but do not actively hedge this. We approximate the duration by the weighted dollar exposure of our maturities.

These caveats show the discrepancies between our approach and a real-life scenario, it is therefore something to be aware of when applying these strategies in real life.

6 Results

This section starts with a discussion of the mean reversion of the model extensions, and different feature configurations. Then, the report of the PnLs obtained by our trading strategies is presented. This is for our best model (GLS), a combination of two models (GLS and RR), and the application of PCA on GLS. Finally, we dive deeper into our method that generates the highest PnL.

6.1 Mean reversion

In order to find the model most suitable for trading, we assess the mean reversion metrics of the residuals of the individual instruments per model. We do this using the criteria introduced and explained in Subsection 4.4. Given that we consider four types of least squares and four features that can be selected, we end up with $4 \times 2 \times 2 \times 2 \times 2 = 64$ different possibilities for our model. We go over all the possible combinations and keep track of the mean reversion properties. We then rank each model by its mean reversion metrics and sort them by their cumulative ranking. The best nine models are shown in Table 5.

Table 5: Ranking of models by mean reversion of residuals

Mthd	Ind	Spr	Liq	MidPx	$\bar{\rho}$	$\bar{t}_{\frac{1}{2}}$	Rank $\bar{\rho}$	Rank $\bar{t}_{\frac{1}{2}}$	Rank Total
GLS	+	+	+	−	0.7045	2.8677	1	1	2
GLS	−	+	+	−	0.7218	3.1575	3	2	5
GLS	+	+	−	−	0.7116	3.3771	2	4	6
GLS	−	+	−	−	0.7309	3.3242	4	3	7
RR	+	+	+	−	0.7856	4.3359	6	5	11
WLS	+	+	+	−	0.745	4.6111	5	6	11
OLS	+	+	+	−	0.7865	4.465	7	7	14
OLS	+	+	−	−	0.7891	4.7004	9	8	17
WLS	+	+	−	−	0.7885	4.7209	8	9	17

Notes: BM= benchmark, EW= equal weighted, MOD=modified, DI=directional, BI=best instruments, GLS = generalized least squares, RR = robust regression, WLS = weighted least squares, RR = robust regression. In the left part, the model settings are shown, next are the values of the mean reversion criteria. On the right, the models are ranked by the sum of the scores, the lower the better. $\bar{t}_{\frac{1}{2}}$=half-life, $\bar{\rho}$ = average of autoregressive coefficient. (+) means features included and (−) features not included. This has been done for the first two and a half months of the data: 13 June 2023 - 26 August 2023.

From Table 5, we draw several conclusions. First, there is a correlation between the different mean reversion rankings. This is in line with our expectations, as a lower $\bar{\rho}$ generally should mean a lower average $\bar{t}_{\frac{1}{2}}$. Then, to examine the performance of the models and the features, we identify two main things. The GLS method generally performs better than the other least squares methods, with the four best mean reverting combinations all being GLS. Looking at the features, the spread feature is always selected for the best-performing models, whereas the

17

liquidity feature is selected in half of the best models. Furthermore, adding the average price of the previous hour of the instrument as a variable did not give any improvement. This variable was not selected in any of the top nine models. RR is the second best method in terms of mean reversion, showing that better handling of the outliers is beneficiary in this respect.

Examining the best model in terms of mean reversions (GLS including variables except the average price), we see that the average AR(1) values on the residuals for the best model are; $\bar{\rho} = 0.7045$, with $(\bar{t}_{\frac{1}{2}})$ 2.8677. This would suggest that the residuals move halfway towards the mean in approximately 2.87 hours. Values towards zero of the mean AR(1) coefficient emphasize that our conventional methods of curve fitting are well applicable in the field of trading.

A more advanced, next step, is orthogonalizing the features by applying principal component analysis (PCA). This way we reduce the amount of features used in our regression. Since we are not using that many features, using PCA might not be necessary, yet it is interesting to see what the effects are on the mean reversion and PnL. We run a 4×4 loop where we differ over the used regression method and the number of components used. The features used in the PCA are tenor, indicator ($\tau_c = \frac{2}{3}$), liquidity, spread, and the average price of the last hour.

Table 6: Comparison of Regression Methods Across Components

Method	1 $\bar{\rho}$	1 $\bar{t}_{\frac{1}{2}}$	2 $\bar{\rho}$	2 $\bar{t}_{\frac{1}{2}}$	3 $\bar{\rho}$	3 $\bar{t}_{\frac{1}{2}}$	4 ρ	4 $\bar{t}_{\frac{1}{2}}$
WLS	0.7189	2.7163	0.7630	3.5220	0.7750	3.8253	0.7716	3.7675
OLS	0.7230	2.8001	0.7635	3.6666	0.7661	3.6382	0.7651	3.6834
GLS	**0.6776**	**2.3286**	0.7084	2.7300	0.7377	2.8518	0.7114	2.8125
RR	0.7155	2.6886	0.7664	3.5207	0.7802	3.8451	0.7825	4.024

Notes: WLS = weighted least squares, OLS = ordinary least squares, GLS = generalized least squares, RR = robust regression. A comparison of our four least squares methods using the mean of the AR(1) regression coefficient ($\bar{\rho}$) of the residuals and mean half-life time ($\bar{t}_{1/2}$) across a different number of principal components. This has been computed for the first two and a half months of the data: 13 June 2023 - 26 August 2023.

Table 6 shows the average value of the AR(1) coefficient ($\bar{\rho}$) and the average half-life time ($\bar{t}_{\frac{1}{2}}$). We stop at four components, as five components would lead to the same results as regular LS, but then with orthogonalized factors. We see that increasing the number of components leads to worse mean reversion values, seeing an increase in the values of $\bar{\rho}$ and $\bar{t}_{\frac{1}{2}}$ and. This suggests that the variance explained by five features is largely captured by the first component, that they contain similar non-orthogonal information, and are therefore collinear. GLS has lower (i.e. better) metrics than the other methods, this is in line with the results of Table 5. In a mid-frequency environment this makes sense as the error terms, in this case mispricing, contain autocorrelation. Emphasizing the relevancy of exploring different regression methods and showing the power of GLS. Since the values overall are even slightly better than without using PCA (Table 5), there is relevance in using PCA with only one component as this reduces the dimensions while still explaining a big part of the variance and shows better mean reversion statistics.

For the final approach, we select only instruments whose half-life time is below (and thus better) than the average of the entire set of instruments.

Table 7: Comparison of GLS, GLS & RR, and PCA methods.

	GLS	GLS & RR	PCA
$\bar{\rho}$	0.7838	0.8139	0.7268
$\bar{t}_{1/2}$	5.54	6.97	3.49
$t_{1/2} < \bar{t}_{1/2}$ (%)	66.07	67.86	69.20

Notes: GLS = generalized least squares method, GLS & RR = model combination between generalized least squares and robust regression model, PCA = principal component analysis on the generalized least squares method. Mean reversion analysis of the three models using performance metrics $\bar{t}_{\frac{1}{2}}$=half-life, $\bar{\rho}$ = average of autoregressive coefficient. This has been computed for the entire dataset.

First, we see that the values of $\bar{\rho}$ and $\bar{t}_{\frac{1}{2}}$ in Table 7 deviate from those in Table 5. This is because in Table 5 we only use the first two and a half months to select our models as an in-sample period. However, new instruments are introduced frequently, which affects the metrics. We also see that the percentage of instruments that have values lower than the mean is above 50%. This means that the majority of the instruments have a half-life time lower than the mean, but the average is being increased by instruments that exhibit far higher half-life times. Finally, we see that PCA shows the best mean reversion metrics across the entire sample, followed by GLS, and the combination of GLS and RR with the worst scores. This is in line with the in-sample set of the first two and a half months.

6.2 Profit & Loss

After assessing our models and their mean reversion, we start trading. Mainly, we see how the mean reversion and rules affect our PnLs. As mentioned in Section 5.2, we have six different strategies that can be defined by two settings. The first allows for directional bets or not, and the second is the restrictions within the model (no restrictions, adjusted residuals with minimum bps, and the restrictions on only trading the most mean-reverting instruments). As our best model, we have a model using GLS and all our variables except the mean price of the instrument. This yields us the following performance metrics.

Table 8: PnL performance overview best GLS model

Metric	BM		MOD		BI	
	DI	EW	DI	EW	DI	EW
Final PnL ($)	1,022,123	**1,307,678**	125,950	741,286	818,187	869,872
PnL/Trade ($)	1,338	**1,777**	30	586	295	404
Transaction Costs ($)	**17,339**	18,531	37,240	31,674	46,127	53,862
PnL/Max Employed (%)	38.0	**43.6**	2.8	27.0	23.7	26.8
Max Drawdown (%)	-17.1	-14.7	**-3.9**	-11.1	-6.7	-8.4
Average Time (H)	31.9	33.2	**9.5**	20.2	11.6	12.7
Hit Ratio	1.035	1.089	0.965	**1.169**	1.100	1.089
Trades	764	736	4,142	1,264	2,770	2,151

Notes: DI = directional, EW = equal weighted, BM = benchmark, MOD = modified, BI = best instruments. Performance metrics for our six trading strategies over the entire dataset for the generalized least squares method.

Starting with the final PnL, in the benchmark equal weighted (BM,EW) approach (see Table 8), we achieve a PnL of approximately $1.3 million, which is the highest among all our strategies. For this PnL, the maximum capital employed at a certain point was approximately $3.0 million. This is followed by the benchmark directional approach (BM,DI). The second-highest method is trading only on instruments whose half-life time is below the average of all the available instruments. Again, the equal-weighted (BI,EW) is higher than the directional (BI,DI) approach. Surprisingly, this approach is significantly lower than taking all instruments, which is not in line with the assumption that better mean reversion statistics lead to higher PnL. However, looking at the maximum drawdown, the downside of more mean reverting instruments is much lower.

When we annualize the residuals, we see a significant decrease in PnL compared to their benchmark approaches. The directional approach (MOD,DI) shows a decrease in PnL of 88% and 43% for equal weighted (MOD,EW). Instruments close to maturity will blow up as the exponent of the annualization approaches 365 (excluding instruments with maturities of less than one day). This will make certain instruments appeal much more under or overpriced than they actually are, and make us trade on instruments that are not that profitable. Another shortcoming is that annualizing overshadows the residuals on the longer end of the curve. This end of the curve has more mispricing, due to more uncertainty, which leads to higher PnL. However, higher mean reversion is mostly exhibited by instruments closer to maturity. Yet, these instruments generate lower PnL as on the shorter end of the curve there is less mispricing. This is the reason why using the 'best instrument' generates lower PnL.

This effect can also be seen in the number of trades, average time to trade, and transaction costs. In our benchmark approach (BM,DI), (BM,EW), we tend to stick to our positions much longer, waiting for a mean reversion that we can trade on. Due to annualizing, maturity overshadows the effect of the actual residual. This makes more different instruments tradeable without actually being mispriced, leading to higher amounts of trades and transaction costs and lower average time between PnL realization. Note that the low number of trades for our

benchmark strategy is the result of wanting to trade on instruments that reached our cap. [1]

Examining PnL per trade and PnL per maximum capital employed. Both are rather in line with the ranking of final PnL, with a clear win for the benchmark approach, followed by selecting the most mean reverting instrument, and at last the annualized residuals. For the benchmark equal-weighted approach, we obtain a PnL per maximum capital employed (sum of long and short exposure) of 43.6%. Annualizing for our 446-day trading window we obtain a value of 34.5%. Important to consider, is that this is not a return, as we do not have our own cash at risk. Therefore, it is to be interpreted as a return on exposure.

Noticeable is that the PnLs and drawdowns have a negative relation, meaning that higher PnL also means a more negative maximum drawdown. This can be seen as a risk/reward trade-off which is one of the main principles of financial engineering and investing (Back, 2010). Nevertheless, in the long run, we still obtain a much higher PnL for the strategy with a higher drawdown. This parallel can somewhat be drawn to the relation between investing in stocks as opposed to fixed-income markets. Stocks are generally more volatile, yet over a longer horizon, they still beat fixed income in terms of returns (Bessembinder, 2018).

We visualize the generation of PnL for the different strategies over the whole data set. Here, we allocate a maximum of $1,000,000$ at each hour, and impose a maximum absolute position of $1,000,000$ per instrument.
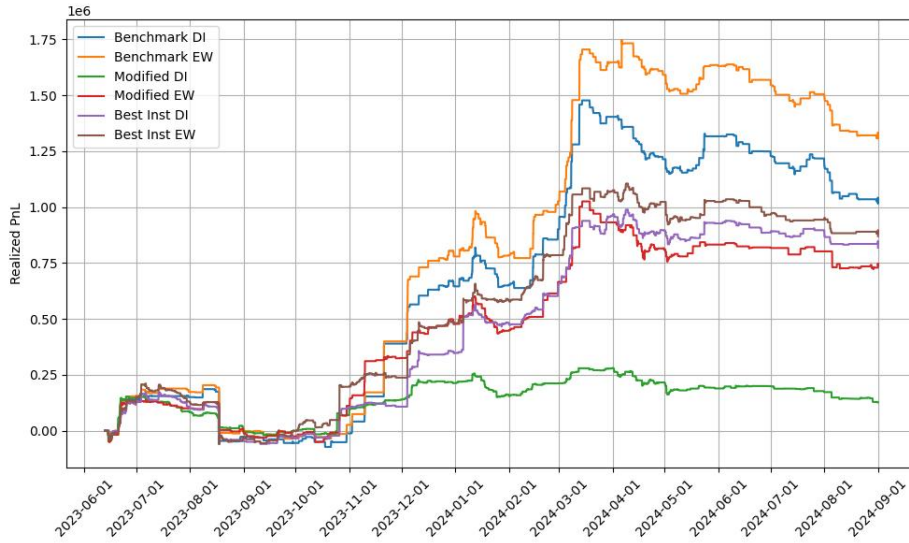

Figure 2: All six strategies; PnL (in millions) using the optimal GLS model

In Figure 2 we see that all strategies develop quite equally up to December 1st 2023. Subsequently, (MOD, DI) remains quite flat, indicating that it does not select trades that lead to much profit as it overinvests in short-tail trades. The other strategies develop similarly to each other with the (BM,DI) and (BM,EW) increasing faster due to their ability to capture the best trades.

---

[1]Trades and transaction costs do not have a linear relation as the magnitude of the trades (and therefore the transaction costs) are different. Hence, (MOD,DI) has almost six times the number of trades of (BM,EW) but only twice the costs.

6.3 Model Combination and PCA

An extra layer to advance our current approach is a combination of models. Model combinations are a widely used method in econometrics to reduce variance in final forecasts. We take the average of the residuals of the best GLS and RR model, this yields the following performance metrics.

Table 9: PnL performance overview model combination best GLS and RR

Metric	BM		MOD		BI	
	DI	EW	DI	EW	DI	EW
Final PnL ($)	**1,256,987**	943,014	582,734	341,973	894,628	825,235
PnL/Trade ($)	**2,064**	1,577	685	101	378	489
Transaction Costs ($)	**13,235**	15,044	21,333	31,969	38,987	42,308
PnL/Max Employed (%)	**44.3**	31.2	21.4	8.1	28.9	28.7
Max Drawdown (%)	-14.9	-15.9	-8.4	**-4.7**	-8.8	-9.8
Average Time (H)	41.3	43.3	29.8	**10.2**	13.0	15.9
Hit Ratio	0.871	0.871	1.065	**1.172**	1.078	0.992
Trades	609	598	851	3,380	2,368	1,689

Notes: DI = directional, EW = equal weighted, BM = benchmark, MOD = modified, BI = best instruments. Performance metrics for our six trading strategies over the entire dataset for the generalized least squares and robust regression combination method.

Again, starting with the PnL, we observe (BM,DI) achieve the highest PnL of approximately $1.26 million. Followed by (BM,EW) approach with a PnL of $0.94 million. The second-highest method is trading only on instruments whose half-life time is below the average of all the available instruments. Again, this is the result of excluding the instruments with higher tenors. Again, this is the reason why using the 'best' instruments generates lower PnL. Finally, the modified strategy generates the lowest PnL. If we compare this with Table 8, we notice differences in the directional or equal-weighted approach. In Table 8, equal-weighted was the most profitable method for all the strategies, yet when combining, the directional approach shows to be generating a higher PnL. An explanation for this is since combining models is a form of diversification, we reduce our model risk, allowing us to make directional bets. This is again an interpretation of the often recurring risk/reward trade-off in investing.

Further looking at the amount of trades, average time to trade, and transaction costs. The amount of trades is lower in this approach, which is an immediate result of combining the models. When averaging out forecasts or predictions you decrease the influence of outliers, however, outliers (i.e. mispricing in our case) is what we trade on. The lower trading activity then leads to a higher trading time and lower transaction costs.

Regarding the PnL per trade and PnL per maximum capital employed, we see mixed results. More specifically, improvements for the directional approach and worse metrics for equal weighted. Our best strategy had a PnL per trade of $1,777 which decreased to $1,577, and the PnL to maximum capital employed decreased with 12.4% to 31.2%. The directional approach increased those metrics to $2,064 and 44.3% respectively. The modified and best instrument

approach shows similar behaviors, where the equal-weighted strategy worsens in PnL and the directional performance increases.

Furthermore, both in Table 8 and 9 the best instrument models show the least differences in their performance metrics among the three strategies. The reason for this is that the instruments selected, are close to each other in terms of absolute value and thus show no big deviations in the allocated weights as opposed to equal-weighted ones.

Finally, we take a look at the maximum drawdown and the hit ratio. Although slightly worse, the maximum drawdowns are somewhat similar to the single-model approach. Only the modified approach shows the aforementioned switch between the equal-weighted and directional approaches. This is in line with expectations given the PnL generated. The hit ratios are significantly worse when using the model combination of GLS and RR. Especially for our benchmark approach the hit ratios are well below one, meaning that we have more trades that decrease or PnL than increase. Still, we remain profitable, meaning that we have more density on the negative side, but a positive skewness in our PnL distribution (Appendix 16).
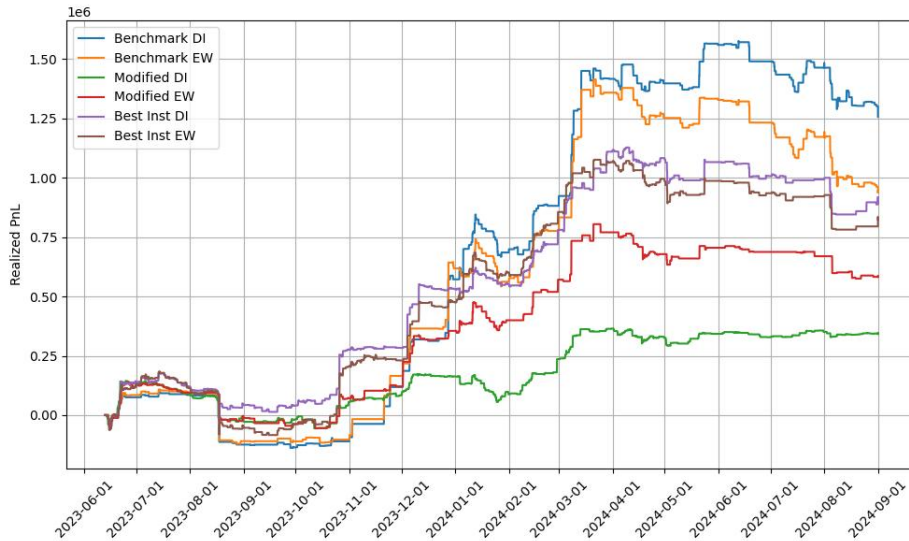
Figure 3: All six strategies; PnL (in millions) using the average residuals of the GLS and RR models

We see similar patterns similar to those in Figure 2. Firstly, there is again a quite flat increase in (MOD, DI), a slightly larger increase in (MOD, EW), and the rest of the trading strategies develop quite similarly. An important difference here is that (BM,DI) ends up with a higher PnL compared to (BM,EW), which could be attributed to better diversification of the model combinations for trading and therefore less need to keep it equally weighted.

Table 10: PnL performance overview best PCA on GLS model

Metric	BM		MOD		BI	
	DI	EW	DI	EW	DI	EW
Final PnL ($)	**1,354,924**	1,182,804	577,933	234,035	905,052	1,240,983
PnL/Trade ($)	549	**558**	273	45	205	314
Transaction Costs ($)	50,909	53,068	53,008	**35,181**	62,234	108,924
PnL/Max Employed (%)	**49.3**	39.1	23.1	5.0	26.0	38.2
Max Drawdown (%)	-16.0	-15.8	-10.2	**-3.6**	-6.9	-8.4
Average Time (H)	12.1	13.2	12.0	**7.2**	8.1	8.0
Hit Ratio	1.089	1.094	0.981	0.985	1.038	**1.159**
Trades	2,466	2,118	2,118	5,158	4,425	3,958

Notes: DI = directional, EW = equal weighted, BM = benchmark, MOD = modified, BI = best instruments. Performance metrics for our six trading strategies over the entire dataset for the principal component analysis on generalized least squares method using one component.

Looking at Table 10, we see that for the optimal model combination selected from Table in Table 6 (GLS with 1 PCA-component), rather similar results hold compared to Table 8. The highest PnL is still achieved by one of the benchmark approaches, however, the (BI,EW) now comes in second, showing that the PCA model better makes use of the mean reversion of the models. Also, a difference, is that the number of executed trades significantly increased for the benchmark approach. This can be interpreted as the models detecting more variability across the instruments that are most mispriced and also capture more mean reversion[2]. Although (BM,DI) generated the highest PnL, (BI,EW) has a higher hit ratio and a significantly lower max drawdown (8.4% compared to 16.0%). Therefore, on a longer horizon, using BI with PCA can be better, as it is less susceptible to shocks and shows a more robust selection of trades. This again shows that PCA is a better fit to leverage mean reversion on specific instruments.

Further, to compare Table 10 to Table 8 and 9. We examine no consistent improvements across the different residuals, meaning we can not conclude that one method is superior to others. Yet, PCA strategies generally achieve higher PnL. Also, PCA achieved the best results when only selecting above-average mean reverting instruments.

PCA, especially with only using the first component, reduces variance. Since cryptocurrency returns exhibit high volatility, PCA can reduce variance, however, introducing a bias. Yet, this bias-variance pays off, as for BI we have the best metrics using PCA comparing Tables 8, 9 and 10.

Finally, graphically looking at the PnL series using GLS with PCA (one component).

---

[2]To clarify the capturing of more mean reversion; we have a limit on exposure per instrument. Allowing more trades on an instrument that reached its limit, can only be the case if the position has been closed in the meantime, and we only close positions if they exhibit a sign switch (i.e. mean reversion).
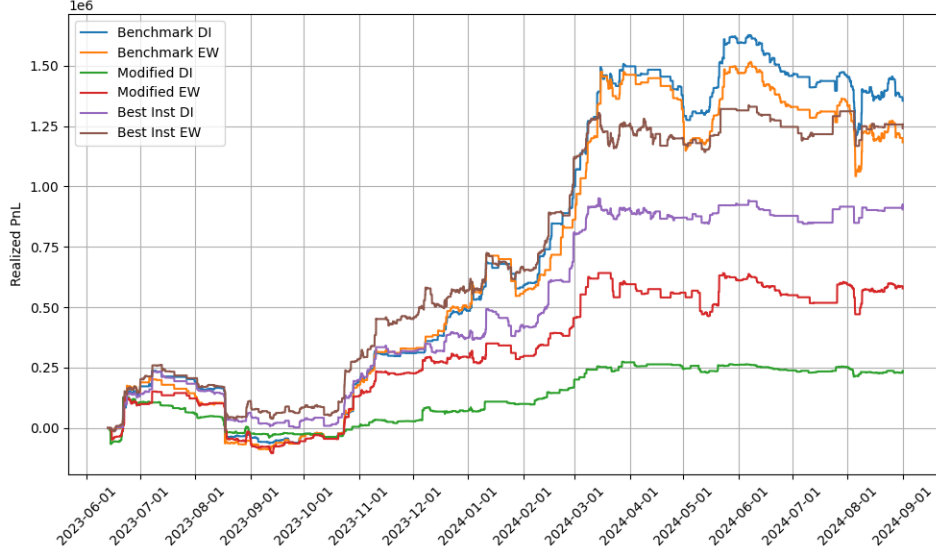
Figure 4: All six strategies; PnL (in millions) using GLS with PCA (one component)

Figure 4 displays the realized PnL for the six different strategies over the period June 2023 to September 2024. Again, compared to Figures 2 and 3, no significant differences in the PnL trajectories. Again (BM,DI) exhibits the highest overall performance, reaching a peak value exceeding 1.5 million before decreasing slightly below this level. (BM,EW) and (BI,EW) follow a similar upward trajectory but lag slightly behind (BM,DI), particularly from March 2024 onward. (MOD,DI) shows the weakest performance, remaining nearly flat with minimal growth, while (MOD,EW) achieves moderate gains but underperforms relative to the rest.
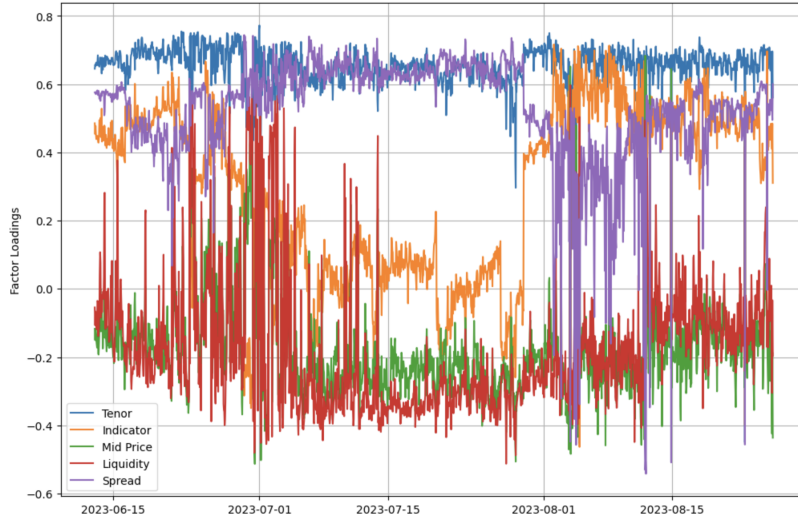
Figure 5: Factor loadings over time of the first principal component, using the in-sample dataset (73 days)

The time series plot of factor loadings illustrates distinct differences in stability and magnitude among the factors. Tenor (blue) consistently maintains high loadings, fluctuating around 0.6 to 0.8, indicating its strong and stable contribution to market dynamics. This is in line with our expectations that the tenor is the most important variable, supported by the theoretical

price of a futures instrument. Spread (purple) stays on the same trajectory as the tenor being high and stable, but experiences a sudden volatile period. During this period, the magnitude of the indicator alternates with the loading of the spread. In contrast, liquidity (red) exhibits high volatility, frequently oscillating between -0.4 and 0.4, suggesting an unstable and reactive role. This is somewhat in line with the results of Table 5, where it is not always included in the most mean-reverting models. The indicator (orange) starts at approximately 0.5 but experiences a steady decline to near 0.1 by mid-July, before sharply rebounding to previous levels in early August. This pattern suggests a diminishing influence of this factor over time, due to the distribution of maturities of instruments in the regression, which makes it more difficult to have a significant influence. The sudden recovery is hence the opposite effect, where due to the increase of a more diverse maturities set, its significance was restored. Meanwhile, the average mid-price of the instrument (green) remains relatively minor throughout, with low and stable loadings around 0 to 0.2, indicating a less dominant role. These dynamics suggest that while some factors maintain a persistent influence, others are more sensitive to evolving market conditions, highlighting the need for adaptive risk models and trading strategies. The less significant role of the mid-price is fully in line with Table 5 where it is not present in any of the top ten models.

6.4 Analysis of highest PnL generating approach

In this subsection, we further examine our most profitable strategy, which uses GLS in combination with PCA and its first principal component. The (BM,DI) generated a PnL of $1,354,924$ (excluding its $50,909$ transaction costs). In addition to the highest PnL, it had a return on maximum exposure of 49.3% which is annualized equal to 38.8%.
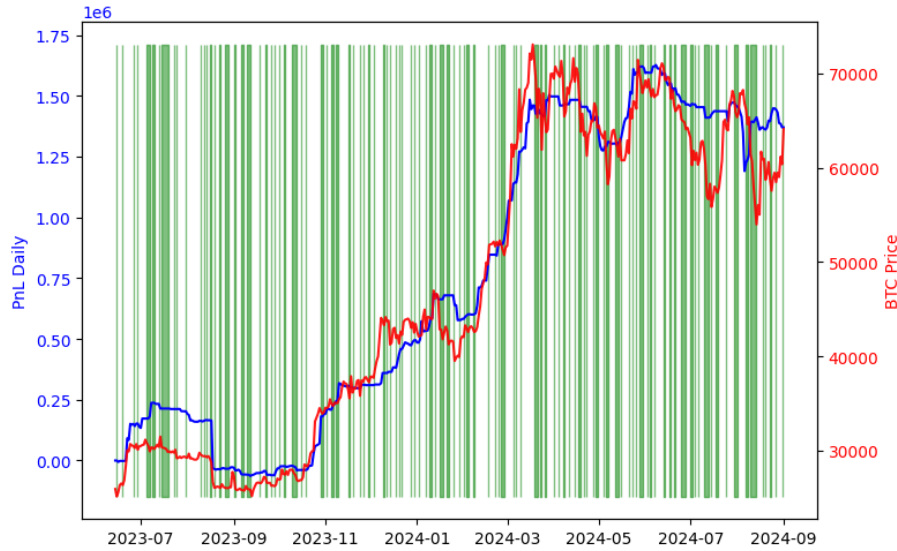

Figure 6: PnL (in millions) vs. BTC price

Notes: The data range from 13 June 2023 to 1 September 2024, profit and loss and Bitcoin price taken at end-of-day. The green bars indicate the days that given a drop in the Bitcoin price, we have a lower drop in PnL (84.2%).

In Figure 6 we see, over the course of our trading window, the development of our PnL (left

axis) and the spot price of BTC (right axis). The first observation is the high correlation of the two series (ρ=0.982), which indicates that our strategy gains from the rise of BTC prices and loses vice versa. The downside of this graph is that we are not as market-neutral as expected, meaning that further development should be to improve our hedges. Yet, our drops in PnL are much less steep than those of the BTC price itself. The green bars in Figure 6 indicate the one-day periods that, given Bitcoin dropped in price, the relative decrease of our PnL was lower (due to good hedging) than that of the spot price of Bitcoin. This was the case in 84.2% of the instances. That means, that despite the high correlation with Bitcoin, which encounters heavy drops, we still manage to be better hedged in more than four out of five times. An example of this can be seen in the period of June-August 2024, where Bitcoin drops heavily whereas our PnL drops much more moderately, also showing an increase where BTC drops. Furthermore, the PnL series is a much smoother line. This is the result of being hedged, but also investing in multiple instruments, which is the result of diversification. The period March 2024 - May 2024 is an example of this; BTC experiences many jumps, and our PnL stays stable showing no big gains or losses.
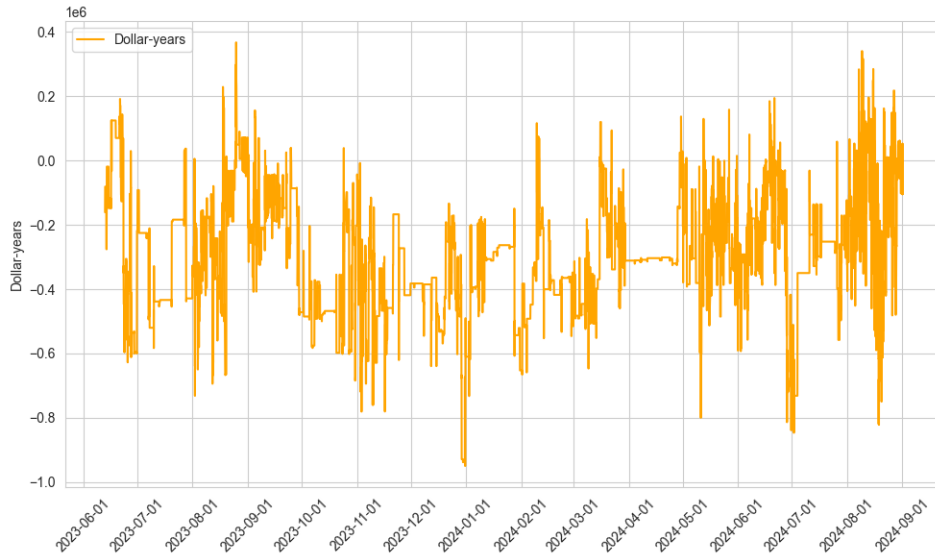

Figure 7: Dollar-years exposure (in millions) of our portfolio 13 June 2023 - 1 September 2024
Notes: Dollar-years is the weighted sum of our positions in instruments and their respective tenors.

Figure 7 illustrates the strategies' duration exposure over time, which is approximated by the weighted sum of tenor and dollar exposure. There are clear fluctuations in the weighted dollar amount allocated to different tenors. The economic consequence of a short-duration exposure, as depicted in many negative regions of the graph, is a sensitivity to rising interest rates. If rates increase, the portfolio benefits, whereas falling rates could lead to losses. Even though we do not actively hedge this risk, Figure 7 gives us insights into how we position ourselves with respect to the maturities of our instruments. Since we are mostly short-duration, this means that we are short more on the longer end of the curve (instruments with high maturities) and long on the shorter end of the curve (instruments with high maturities). This suggests that our model generally estimates that instruments with high maturities are overpriced, and vice versa, that instruments with low maturities are underpriced. The interpretation of going considerably more

27

short on longer maturing instruments is that our model generally estimates that the implied rate of return baked into the futures price, is higher than the cost of financing (mostly driven by the interest rate).

Next, we examine the difference between the out from BTC and ETH instruments in terms of PnL attribution. Figures 8 and 9 display bar charts in descending order of the PnL generated per instrument.
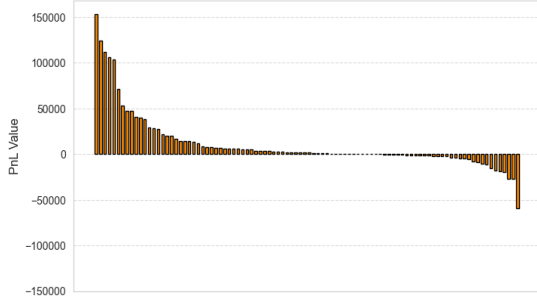

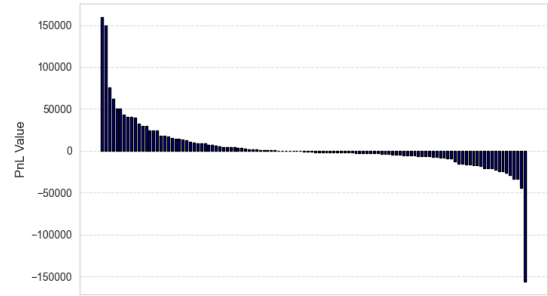Figure 8: PnL distribution of BTC instruments (Total PnL: $1,019,127)

Figure 9: PnL distribution of ETH instruments (Total PnL: $351,137)

Table 11: Summary statistics for ETH and BTC PnL

	Total PnL	Mean	Std	Min	5%	Median	95%	Max
BTC	1,019,127	10,616	31,158	-59,369	-17,962	1,586	79,619	153,321
ETH	351,137	3,027	31,088	-155,744	-24,901	-1,389	44,557	159,928

Notes: PnL = profit and loss, std = standard deviation, min = minimum value, 5%/Median/95% = quantile values, max = maximum value. Calculated over the entire dataset.

From the total PnL, 75.2% was driven by Bitcoin trades, dominating the gains. Also from a distributional aspect, BTC instruments have a mean PnL of $10,616$ as opposed to $3,027$ for ETH. This is even though they have similar standard deviations ($31,158$ and $31,088$ respectively). For ETH, the median is negative ($-$1,389$), meaning that the majority of the instruments of ETH are losing us money. The argument for not using ETH in our strategy is even stronger when we look at the tails. The 5th and 95th percentile scores are significantly worse for ETH, demonstrating skewness towards the losses. Finally, the most losing instrument for ETH decreased our PnL with $155,744$, whereas for BTC this was $59,369$. All of this amplifies that PnL analysis is a crucial part of trading strategies. By doing so, we can mitigate losses by excluding certain instruments or only trading on specific parts of the curve. This way we also have more to allocate to the more profitable elements of the strategy.

7 Conclusion

This paper finds profitable trading strategies, when modeling the futures premia of Bitcoin and Ethereum, using Principal Component Analysis (PCA) on a generalized least squares (GLS) estimated futures curve, among other approaches. We have done this using data from the three largest cryptocurrency exchanges from June 13, 2023, to September 1, 2024. We model the

forward curve first via the no-arbitrage futures condition using ordinary least squares (OLS). Later, we extend this using additional features that were proposed to correct for unfulfilled conditions of the perfect market hypothesis (PMH), and three different regression methods, namely weighted least squares (WLS), GLS, and robust regression (RR). We select the best-performing models based on the properties of the mean reversion of the residuals, after which we explore several trading strategies. A higher mean reversion of the residuals means that the futures prices move faster towards the curve, and therefore are more suited to be traded on.

The results show that the best-performing models are all using the GLS method and benefit from the additional liquidity and spread features. Importantly, GLS significantly improves the mean reversion compared to OLS, which agrees with the structural deviation from homoscedasticity and autocorrelation assumptions. What leads to even more improvement in mean reversion, is orthogonalising the features used in the optimal GLS setting using PCA. The best PCA configuration uses one component, again in combination with GLS. Given that the mean reversion improved, this means that the variance was successfully reduced compared to the GLS setup with all the features, without creating too much bias.

From here, we create six different trading strategies: a benchmark strategy, always trading with the two most under- and overpriced instruments; a modified strategy, that scales for the maturity of the instrument and imposes a minimum value for the residual for the instrument to be traded on; the best instrument strategy, only trades on the instruments with below average half-life times. All three of these strategies can be traded on both equally weighted (EW; equal long and short position) and directionally (DI; adjusted to relative mispricing), leading to our six strategies. We then evaluate the best-performing GLS setup, the model combination of GLS and RR as well as the PCA model with GLS on all six of the strategies and calculate trading performance metrics.

For our best model approach we achieve the best performance from our benchmark trading strategy. The EW approach is preferred here, giving us a net zero long-short exposure which mitigates our market risk. Also, trading on instruments with higher maturities is more profitable than lower ones, due to more mispricing.

For our approach, combining the two best different LS methods (GLS and RR), we continue to generate the highest PnL using our benchmark approach. However, now using DI leads to a higher PnL than EW. The result of combining models is that it reduces variance in our forecasts, which leads to more stable outcomes. This stability is then used to make more confident directional bets that pay off.

For the PCA approach, we again find the most profits in our benchmark model. Here, the strategy profits from using directional bets as well. This indicates that the relevant residual size is captured effectively.

We see that with the trading strategies for all three of our models, the benchmark trading strategy outperformed the modified and the best instrument strategies. The modified strategy that adjusts our residuals for the tenor by annualizing, does not improve and worsens the selection of instruments. This is explained by the fact that it blows up the residuals of instruments on the short end. Due to this instability, unnecessary amounts of trading are done. This not only increases trading costs but also results in more mispriced instruments on the longer end of

the curve to not be traded on. Hence, trading on the longer end of the curve generates a higher PnL. Selecting instruments whose mean reversion statistics are better than average did also not lead to an improvement in performance. This is counter-intuitive since the idea of trading here is identifying mispriced assets, and capitalizing on the idea that they revert back to our theoretical line. It can, however, be explained because of the fact that short-end instruments have higher mean reversion, but lower residuals. This would lead to this strategy to not being able to capture as much profit. Again emphasizing that the longer end generates more PnL.

If we compare the three models' best trading strategies to each other, we come to the following conclusion. The PCA on the GLS model, using the first principal component finds a PnL of approximately $1.35 million relative to a maximum capital employed of around $2.7 million. This is the highest one of the three considered models. We further see that of the total PnL, 75.2% is driven by Bitcoin instruments. Furthermore, our PnL shows a high correlation with the Bitcoin spot (ρ=0.982), but still manages to be well hedged against drops in the price of Bitcoin.

Regular GLS has the second-best PnL of around $1.31 million and the GLS and RR combination has the lowest PnL of approximately $1.26 million. PCA performs even better when we look at PnL per maximum capital employed since this figure is higher for both the other models. The fact that GLS and GLS using PCA have better performance metrics than the combination of GLS and RR model confirms our hypothesis of GLS being the best method based on the highest mean reversion, as well as the added benefit of PCA.

From this we come to answer the research question affirmatively. There is indeed an edge to be found in cryptocurrency futures trading when using a GLS curve fit with PCA orthogonalized features. This can be taken into the broader economic context than just our research. We show that the conventional curve fitting models and estimation techniques, which have been shown to be statistically relevant for decades, are applicable in the unconventional cryptocurrency futures market.

Some theoretical limitations include the assumption of a non-changing time-to-maturity which largely simplified modeling from a stationarity perspective. Furthermore, the aggregation of the data into hour intervals instead of seconds, loses information that can potentially increase predictive power. For future research, we propose the addition of more features to further correct for PMH conditions, such as funding rate, volatility, and sentiment analysis of the news. Furthermore, selecting the used model (combinations) via a rolling window could increase the performance as the mean reversion abilities of the models adapt over time. In addition, the use of machine learning models such as neural networks to examine the behavior of the residuals, also can increase predictability and therefore profitability. Finally, exploring methods such as recursive least squares to leverage the information within the hour.

References

Aharon, D. Y., Umar, Z., and Vo, X. V. (2021). Dynamic spillovers between the term structure of interest rates, bitcoin, and safe-haven currencies. *Financial Innovation*, 7(1):1–59.

Akyildirim, E., Cepni, O., Corbet, S., and Uddin, G. S. (2023). Forecasting mid-price movement of bitcoin futures using machine learning. *Annals of Operations Research*, 330(1):553–584.

Alexander, C., Chen, X., Deng, J., and Wang, T. (2024). Arbitrage opportunities and efficiency tests in crypto derivatives. *Journal of Financial Markets*, 71(1):1–20.

Alexander, C. and Dakos, M. (2020). A critical investigation of cryptocurrency data and analysis. *Quantitative Finance*, 20(2):173–188.

Almeida, J. and Gonçalves, T. C. (2024). Cryptocurrency market microstructure: a systematic literature review. *Annals of Operations Research*, 332(1):1035–1068.

Ba, H.-L. and Ömer Faruk Şen (2024). Explaining variation in national cryptocurrency regulation: Implications for the global political economy. *Review of International Political Economy*, 31(5):1472–1495.

Back, K. (2010). *Asset pricing and portfolio choice theory*. Oxford University Press.

Baxter, M. and Rennie, A. (1996). *Financial Calculus: An Introduction to Derivative Pricing*. Cambridge University Press.

Bessembinder, H. (2018). Do stocks outperform treasury bills? *Journal of Financial Economics*, 129(3):492–518.

Bianchi, R. J., Fan, J. H., Miffre, J., and Zhang, T. (2023). Exploiting the dynamics of commodity futures curves. *Journal of Banking & Finance*, 154(1):1–49.

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654.

Bouri, E., Roubaud, D., and Shahzad, S. J. H. (2020). Do bitcoin and other cryptocurrencies jump together? *The Quarterly Review of Economics and Finance*, 76(1):396–409.

Breusch, T. S. and Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47(5):1287–1294.

Chen, J. (2025). Cost of carry. *Investopedia*, 1(1):1.

Cheng, C.-H., Yang, J.-H., and Dai, J.-P. (2025). Verifying technical indicator effectiveness in cryptocurrency price forecasting: a deep-learning time series model based on sparrow search algorithm. *Cognitive Computation*, 17:21.

Chevallier, J. (2012). Advanced topics: Time-to-maturity and modeling the volatility of carbon prices. *Econometric Analysis of Carbon Markets: The European Union Emissions Trading Scheme and the Clean Development Mechanism*, pages 181–207.

Cholesky, A.-L. (2005). Sur la résolution numérique des systèmes d'équations linéaires. *Bulletin de la Sabix. Société des amis de la Bibliothèque et de l'Histoire de l'École polytechnique*, (39):81–95.

CoinMarketCap (2025). Top cryptocurrency derivatives exchanges. *CoinMarketCap*.

Corbet, S., Lucey, B., Peat, M., and Vigne, S. (2018). Bitcoin futures—what use are they? *Economics Letters*, 172(1):23–27.

De Roon, F. A., Nijman, T. E., and Veld, C. (1998). Pricing term structure risk in futures markets. *Journal of Financial and Quantitative Analysis*, 33(1):139–157.

Durbin, J. and Watson, G. S. (1950). Testing for serial correlation in least squares regression: I. *Biometrika*, 37(3):409–428.

Edelberg, W. and Marshall, D. (1996). Monetary policy shocks and long-term interest rates. *Economic Perspectives*, 20(2):1–18.

Fassas, A. P., Papadamou, S., and Koulis, A. (2020). Price discovery in bitcoin futures. *Research in International Business and Finance*, 52(9):1–30.

Fil, M. and Kristoufek, L. (2020). Pairs trading in cryptocurrency markets. *IEEE Access*, 8(1):172644–172651.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, 46(6):1251–1271.

Heij, C., de Boer, P., Franses, P. H., Kloek, T., and van Dijk, H. K. (2004). *Econometric methods with applications in business and economics*. Oxford University Press.

Huber, P. J. (1992). *Robust Estimation of a Location Parameter*, volume 1. Springer.

Jarque, C. M. and Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, 6(3):255–259.

Kajtazi, A. and Moro, A. (2019). The role of bitcoin in well diversified portfolios: A comparative global study. *International Review of Financial Analysis*, 61(1):1–62.

Liu, Y. and Tsyvinski, A. (2020). Risks and returns of cryptocurrency. *The Review of Financial Studies*, 34(6):2689–2727.

Makarov, I. and Schoar, A. (2020). Trading and arbitrage in cryptocurrency markets. *Journal of Financial Economics*, 135(2):293–319.

Malladi, R. K. and Dheeriya, P. L. (2021). Time series analysis of cryptocurrency returns and volatilities. *Journal of Economics and Finance*, 45(1):75–94.

Mokni, K., El Montasser, G., Ajmi, A. N., and Bouri, E. (2024). On the efficiency and its drivers in the cryptocurrency market: the case of bitcoin and ethereum. *Financial Innovation*, 10(1):39.

Orte, F., Mira, J., Sánchez, M. J., and Solana, P. (2023). A random forest-based model for crypto asset forecasts in futures markets with out-of-sample prediction. *Research in International Business and Finance*, 64(1):1–29.

Quandt, R. E. (1960). Tests of the hypothesis that a linear regression system obeys two separate regimes. *Journal of the American Statistical Association*, 55(290):324–330.

Ramsey, J. B. (1969). Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society*, 31(2):350–371.

Ritchken, P. and Boenawan, K. (1990). On arbitrage-free pricing of interest rate contingent claims. *The Journal of Finance*, 45(1):259–264.

Rousseeuw, P. J. and Leroy, A. M. (2003). *Robust regression and outlier detection.* John wiley & sons.

Sathyanarayana, S. and Gargesa, S. (2019). Modeling cryptocurrency (Bitcoin) using vector autoregressive (VAR) model. *SDMIMD Journal of Management*, 10(2):1–18.

Zhang, C., Ma, H., Arkorful, G. B., and Peng, Z. (2023). The impacts of futures trading on volatility and volatility asymmetry of bitcoin returns. *International Review of Financial Analysis*, 86(1):1–55.

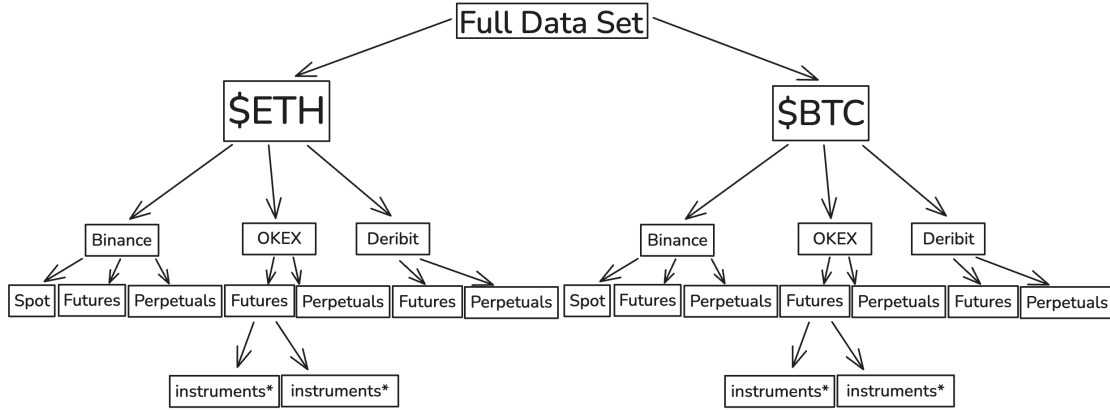# 8 Appendix

8.1 Figures and Tables

Figure 10: Preprocessing of the dataset visualized. **Note:** The '*' in the visualization indicates the different instruments/maturities used for this analysis, which might differ depending on which day the data stems from.
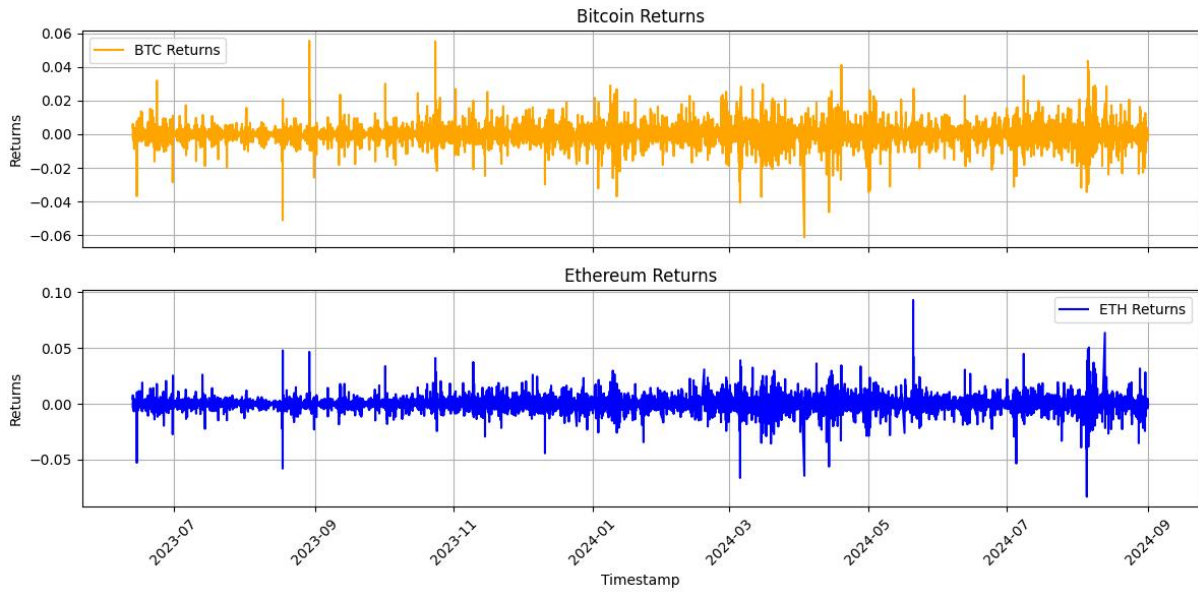

Figure 11: Timeseries of hourly BTC and ETH spot return series from June 13th 2023 to September 1st 2024.
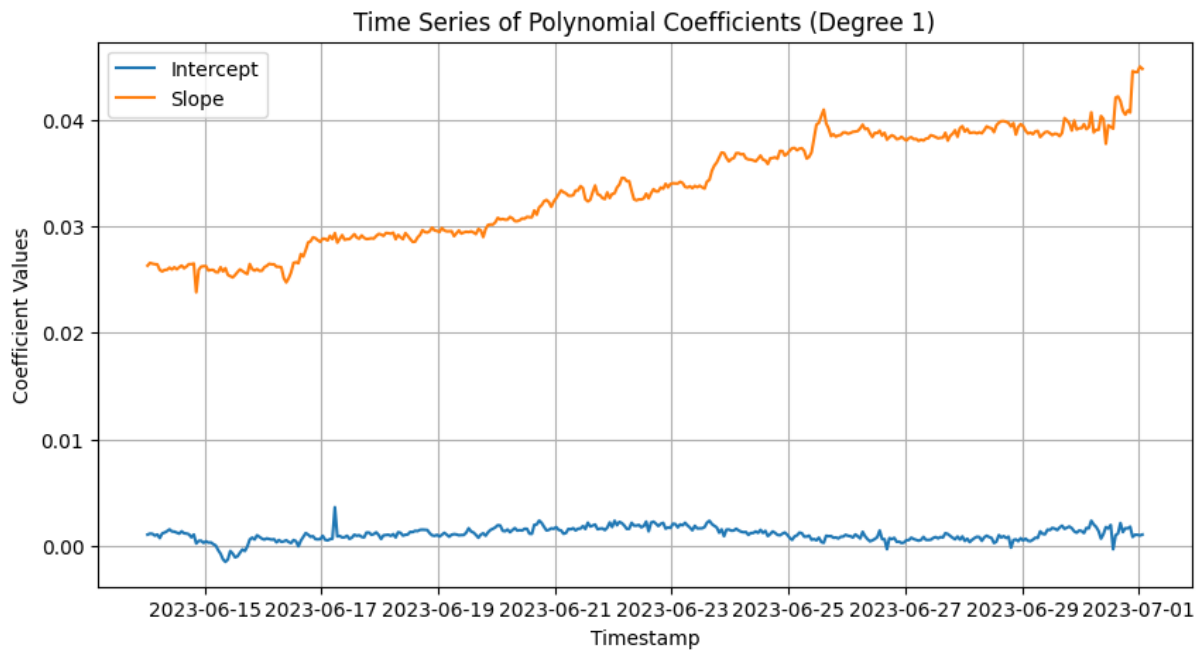
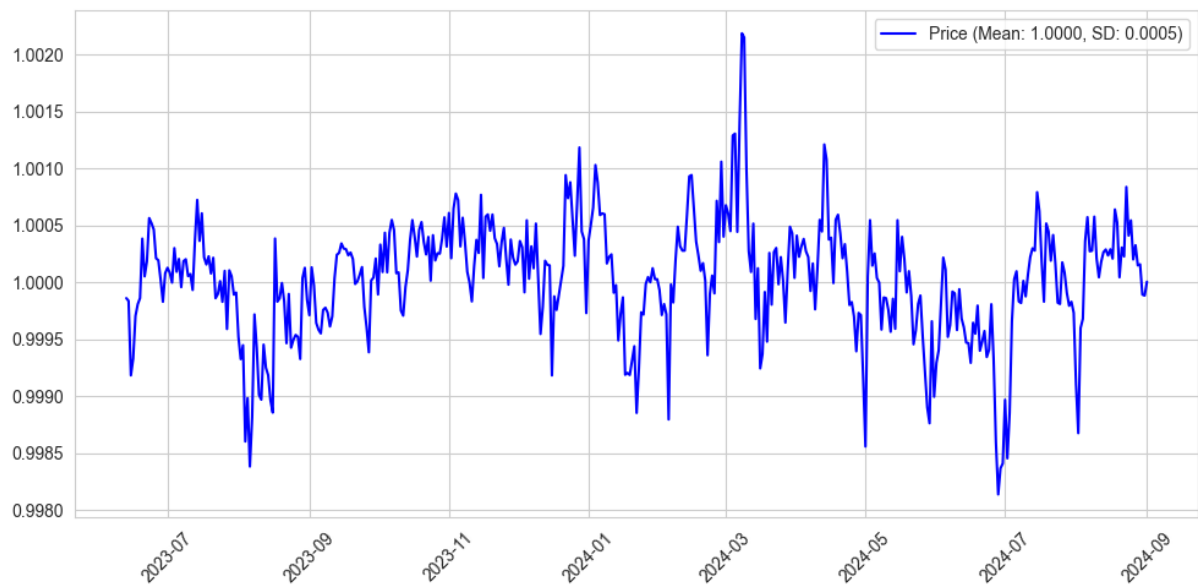Figure 12: Base model intercept and slope coefficients over the first month of the dataset.

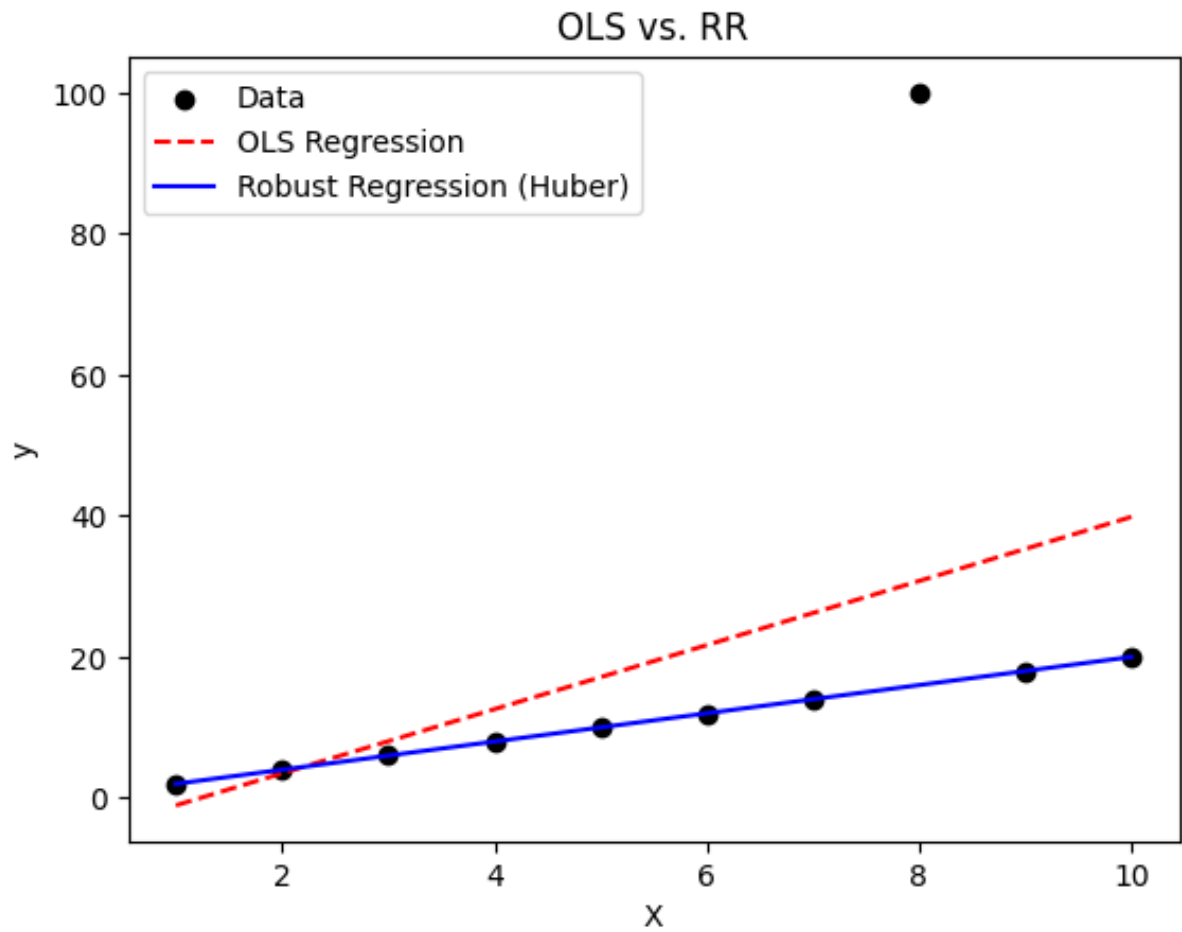Figure 13: USD/USDT price from June 13th 2023 to September 1st 2024.

Figure 14: Illustration of the difference between OLS and RR, RR handling the presence of outliers in a more robust manner.
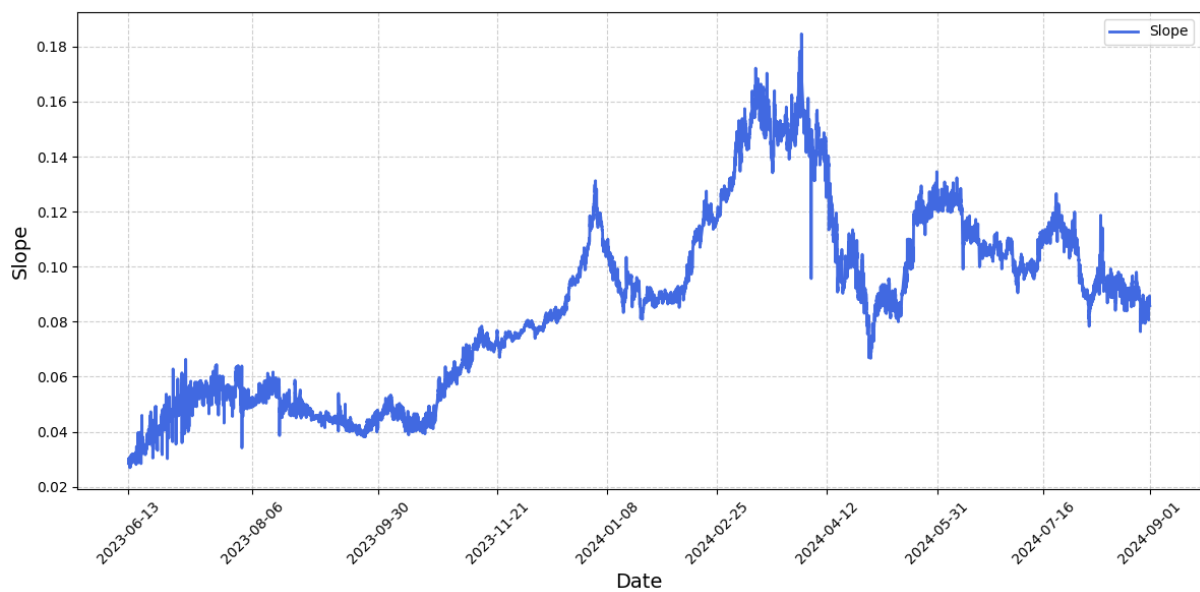

Figure 15: Base model slope coefficient over the entire dataset.

Table 12: Instrument data with residuals and tenor

Instrument	$\epsilon_{t,i(\tau)}$	τ	$\tilde{\epsilon}_{t,i(\tau)}$
A	0.020	0.8	0.025
B	0.015	0.4	0.038
C	0.010	0.1	0.105
D	-0.020	0.3	-0.065
E	-0.030	0.5	-0.059
F	-0.040	0.8	-0.050

Notes: Example of residual scaling for different maturities following $\tilde{\epsilon}_\tau = (1 + \epsilon_\tau)^{\frac{1}{\tau}} - 1$ where $\epsilon =$ residual and $\tau = \text{tenor} \in (\frac{1}{365}, 1)$. Note that with the benchmark approach instruments A, B, E, and F would be selected. After the correction for τ, instruments A and F are replaced with C and D respectively
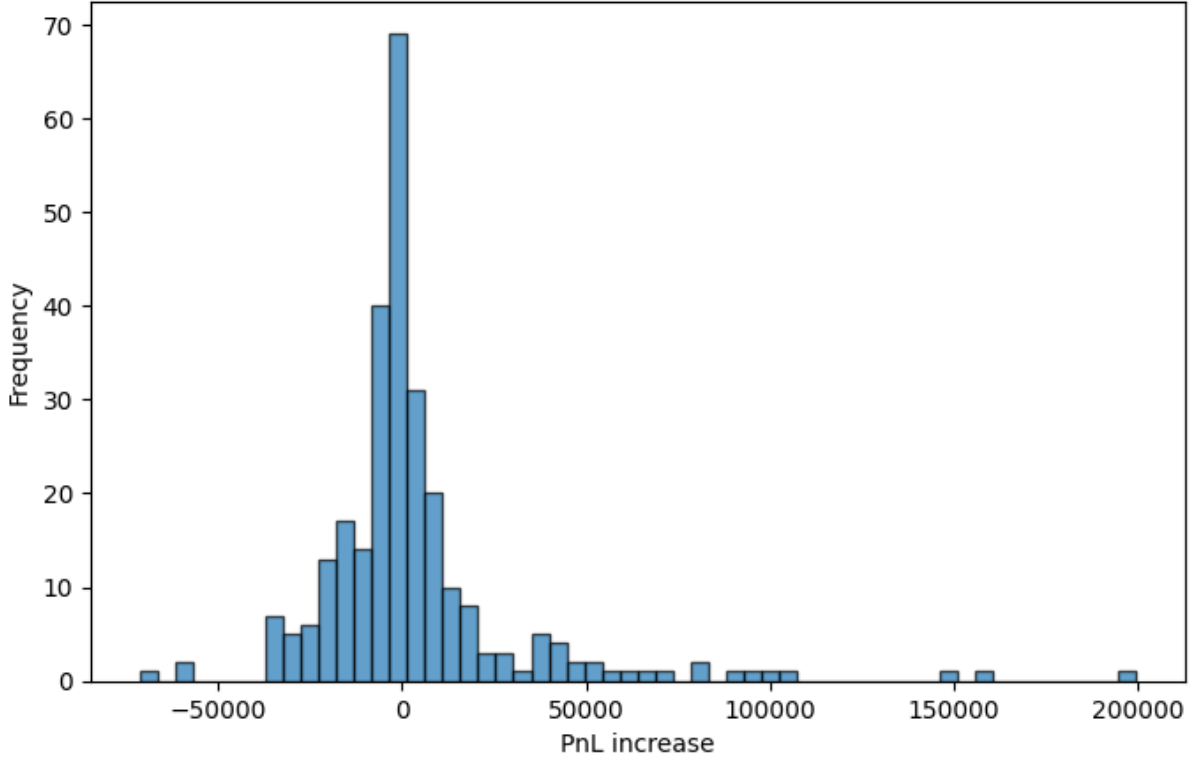

Figure 16: PnL distribution of the benchmark strategy, directional with a GLS/RR combination.

8.2   Generative AI Declaration

In certain parts of this report, we used generative AI to address some code errors, refine particular phrases, check grammar, correct spelling, or format tables. The tools used for this are ChatGPT and Grammarly. However, not all instances required AI assistance, and almost all tasks were completed without relying on these tools.