MICROSTRUCTURE AND TRADING SYSTEMS

ITESO, Universidad Jesuita de Guadalajara

Project:

004 Pairs Trading

Professor:

Luis Felipe Gómez Estrada

Presented by:

David Campos Ambriz – 744435

Contents

Summary

In this report, a “pairs trading” statistical arbitrage strategy is developed, designed to capitalize on temporary deviations of a pair of cointegrated assets that share a certain equilibrium and economic relationship, maintaining a neutral approach to market changes and trends. This strategy, founded on the theory of cointegration, ensures the existence of a long-term equilibrium between two time series, indicating that their spread is stationary and tends to revert to its mean in the face of any market anomaly that moves both assets away from each other. This self-correcting property presented by the theory will be the main focus that the model will work to exploit for any potential profits.

To leverage this idea, a selection methodology is applied to filter candidate pairs through three statistical tests on the training set. A Pearson correlation filter, the Engle-Granger test, and the Johansen test to validate cointegration.

Furthermore, to overcome the limitations of a static model, a dual approach of Kalman filters modeled in the sequential decision analysis framework will be applied in the strategy. The first of the filters estimates a dynamic hedge ratio so that it can adapt to changes in the linear regression of the prices. And the second filter, which stabilizes a trading signal by iteratively estimating the cointegration vector of the error correction model. Finally, the strategy was validated by optimizing the Z-score entry threshold θ through iterated tests to maximize the Calmar ratio as a performance metric.

Strategy Description and Rationale

Overview of pairs trading approach

Within the world of “trading”, the pairs trading strategy is a strategy based on statistical arbitrage whose objective is to detect temporary anomalies within an equilibrium relationship between two assets and, based on that, use these signals to generate a profit. The interesting thing about this premise is that, unlike other trading strategies, this one has an approach that is, to a certain extent, neutral to market movements, as it is based solely on the convergence of the spread between the two selected assets, meaning it does not depend on the general trend or direction of the market.

The execution of this strategy stems from the exercise of identifying a pair of assets that, after passing different statistical analysis tests, could be said to “move together” or appear to behave similarly in the long term; therefore, one of the proposals of this strategy is that, should both assets separate, they would show signals of reverting to the mean (coming back together). Stated this way, when the spread between both assets widens beyond a certain level, the strategy would come into action, taking a long position in the “undervalued” asset and, simultaneously, opening a short position in the asset that is “overvalued” according to the theory. In this way, the profit from this strategy would be expected to be obtained when this difference reverts to the mean, which is the moment when both positions would be closed.

Why cointegration indicates arbitrage opportunity

For a strategy like pairs trading, using a simple correlation as the base argument to operate and justify the use of both assets would be insufficient, as this only measures their joint movements in the short term, and nothing ensures that this will continue for a long time. Therefore, in this strategy, cointegration is used as the statistical basis to identify a relationship that holds in the long term.

As such, the concept of cointegration, developed by Granger (2001), points out that a “special” relationship can exist between two time series that, although *a priori* each series individually may be non-stationary, there actually exists a linear combination between both series that is stationary.

That is, if x_t and y_t are non-stationary, they can be cointegrated by a parameter A, which, in combination, could now be considered stationary, such that:

$$z_t = x_t - Ay_t$$

Written this way, it could be said that a long-term equilibrium relationship exists, where z_t would represent the error of said equilibrium, which measures how much both assets have deviated from the equilibrium that “unites” them.

That said, the arbitrage opportunity arises because, if both assets were not cointegrated, the term z_t would also be a series integrated of order 1 that would not return to its mean, and therefore both series would separate indefinitely. However, since z_t is stationary, this series would tend to return to its mean, and it is precisely this self-correcting tendency that the pairs trading strategy seeks to exploit. Additionally, the theory of this test indicates that when two variables are cointegrated, their behavior must be described by an error correction model, which states that a future change in prices Δx_t and Δy_t may be defined in a certain way by the disequilibrium of that moment, which would indicate that today's spread may provide certain information about tomorrow's price direction; information that can be used to open new positions in both assets.

Justification for Kalman filter use in dynamic hedging

The Kalman filter is a dynamic algorithm that serves as an adaptive tool for modeling time series by estimating the “true state” of a system from noise in the data. The filter operates in a two-step cycle where it first “predicts” by estimating where the state should be, and then “updates” by correcting the given estimate with the new observation made (The AI Quant, 2024).

Considering this, within the application of cointegration models, there is the limitation of the cointegration parameter, which in many cases is assumed to be static, ignoring that in real markets the market situation is changing by nature and presents regime changes that mark new trends, ultimately invalidating the idea of a fixed hedge ratio. Therefore, the application of Kalman filters in this strategy would allow for modeling the spread and “predicting” the hedge ratio in a certain way, because instead of assuming a fixed relationship, the Kalman filter would be able to continuously adjust this parameter by learning from the new information it receives and thereby adapt to changes in the market.

This project implements two Kalman Filters to model the problem as a sequential decision process:

1. Filter 1 (Dynamic Hedge Ratio): This filter estimates the hedge ratio (β_1) in real-time. It models the regression $Price\_Y = \beta_0 + \beta_1 * Price\_X$, treating $[\beta_0, \beta_1]$ as the hidden state to be estimated. Its objective is to determine how much of one asset to hedge against the other.
2. Filter 2 (Dynamic VECM Signal): This filter dynamically estimates the cointegration vector $[e_1, e_2]$. It uses the result of a rolling-window Johansen test as the "noisy" observation and generates a stabilized spread (VECM) signal. Its objective is to determine when to trade.

Expected market conditions for strategy success

As mentioned previously, the pairs trading strategy is designed to be theoretically “risk-neutral,” meaning its outcome is independent of whether the market is bullish or bearish and solely depends on the persistence of favorable conditions and the non-manifestation of adverse conditions:

- Favorable:

  Generally, the strategy tends to perform well in markets that exhibit a certain amount of short-term noise and volatility, as these are the very concepts that generate deviations in the equilibrium of both assets, which the model monitors in search of the moment when the spread reverts to its mean.

- Adverse:

  The worst-case scenario for a model like this is the generation of a profound change in the market regime that indefinitely breaks the cointegration relationship between both assets.

Pair Selection Methodology

Continuing with the previous idea, for the pairs trading strategy, the selection of a robust pair is determinant for the success or failure of the model, as choosing an inadequate pair, even if it has a strong correlation, can generate large losses in the long term by not demonstrating a constant equilibrium between them. Therefore, this project employs a methodology of a three-stage sequential process applied to the training set (60% of the 15 years of data) with the objective of finding a pair that passes all three tests to be considered candidates.

Correlation screening criteria and results

The first test or filter serves to identify pairs with a strong co-movement in the short term by calculating the Pearson correlation on the prices of all possible pair combinations to find candidates that pass the minimum correlation threshold of 0.6.

ticker1	ticker2	correlation	passed_corr	eg_p_value	passed_eg	jo_trace_stat	passed_jo	passed_all
CVX	OKE	0.9396	TRUE	0.0017	TRUE	24.71	TRUE	TRUE
ET	WMB	0.9087	TRUE	0.0037	TRUE	17.62	TRUE	TRUE

MPC	OKE	0.9714	TRUE	0.0503	FALSE	11.36	FALSE	FALSE
PSX	VLO	0.9666	TRUE	0.076	FALSE	8.59	FALSE	FALSE
MPC	PSX	0.9564	TRUE	0.1397	FALSE	8.61	FALSE	FALSE
FANG	PSX	0.9391	TRUE	0.8382	FALSE	3.36	FALSE	FALSE
FANG	VLO	0.9311	TRUE	0.0824	FALSE	4.3	FALSE	FALSE
EOG	FANG	0.9124	TRUE	0.0824	FALSE	12	FALSE	FALSE
HAL	MPC	0.8607	TRUE	0.9157	FALSE	6.26	FALSE	FALSE
EOG	PSX	0.8607	TRUE	0.9157	FALSE	6.26	FALSE	FALSE
HAL	MPC	0.8572	TRUE	0.3933	FALSE	9.39	FALSE	FALSE
OKE	WMB	0.8562	TRUE	0.261	FALSE	11.37	FALSE	FALSE
OKE	PSX	0.8506	TRUE	0.3973	FALSE	3.9	FALSE	FALSE
OKE	VLO	0.8395	TRUE	0.1141	FALSE	14.13	FALSE	FALSE
MPC	OKE	0.8263	TRUE	0.3432	FALSE	6.3	FALSE	FALSE
BKR	SLB	0.8213	TRUE	0.1076	FALSE	13.95	FALSE	FALSE
EOG	VLO	0.819	TRUE	0.8605	FALSE	5.52	FALSE	FALSE
FANG	OKE	0.8152	TRUE	0.909	FALSE	4.48	FALSE	FALSE
CVX	PSX	0.8015	TRUE	0.3628	FALSE	9.36	FALSE	FALSE
CVX	MPC	0.8015	TRUE	0.1247	FALSE	12.33	FALSE	FALSE
CVX	VLO	0.7985	TRUE	0.1201	FALSE	10.34	FALSE	FALSE
EOG	OKE	0.7854	TRUE	0.9467	FALSE	4.41	FALSE	FALSE
CVX	EOG	0.7632	TRUE	0.6132	FALSE	6.73	FALSE	FALSE
BKR	HAL	0.768	TRUE	0.2102	FALSE	10.73	FALSE	FALSE
CVX	FANG	0.765	TRUE	0.5201	FALSE	7.91	FALSE	FALSE
EOG	ET	0.7503	TRUE	0.261	FALSE	11.87	FALSE	FALSE
COP	CVX	0.7106	TRUE	0.5257	FALSE	10.05	FALSE	FALSE
COP	EOG	0.702	TRUE	0.6402	FALSE	8.94	FALSE	FALSE
COP	OKE	0.6892	TRUE	0.7848	FALSE	7.41	FALSE	FALSE
BKR	DVN	0.6525	TRUE	0.2698	FALSE	11.12	FALSE	FALSE
DVN	SLB	0.6325	TRUE	0.3838	FALSE	5.76	FALSE	FALSE
ET	FANG	0.6241	TRUE	0.5761	FALSE	6.26	FALSE	FALSE
COP	ET	0.6241	TRUE	0.4556	FALSE	11.39	FALSE	FALSE
HAL	OXY	0.6226	TRUE	0.0551	FALSE	11.07	FALSE	FALSE
COP	WMB	0.6178	TRUE	0.3626	FALSE	12.02	FALSE	FALSE
COP	MPC	0.6075	TRUE	0.6167	FALSE	6.49	FALSE	FALSE
ET	MPC	0.6063	TRUE	0.7873	FALSE	5.33	FALSE	FALSE
DVN	OXY	0.6028	TRUE	0.0558	FALSE	12.4	FALSE	FALSE
DVN	KMI	0.602	TRUE	0.4788	FALSE	4.63	FALSE	FALSE
OXY	WMB	0.1469	FALSE	0.2406	FALSE	17.04	TRUE	FALSE

Table 1. Top 40 Pairs results

As observed in Table 1, a large number of pairs in the energy sector surpassed the correlation threshold, which is expected given their high exposure to the same macroeconomic factors (e.g., oil and gas prices). However, when their results in the rest of the tests are analyzed, it is noted that in the end, only two pairs remained that managed to pass all three tests.

Simple cointegration test (Engle-Granger method)

The second filter, and the most critical, is the Engle-Granger cointegration test. This test evaluates whether a long-term equilibrium exists between the two assets and consists of:

- Perform an Ordinary Least Squares (OLS) regression between the prices of the pair:

$$P_y = \beta_0 + \beta_1 P_x + \epsilon$$

- Extract the residuals ϵ from the regression, which represent the spread.
- Apply the ADF test on the residuals

As a result of this, the null hypothesis of the ADF is that the residuals have a unit root (i.e., they are not stationary). If the residuals are stationary, it would be concluded that the pair in question is cointegrated, based on the criterion of rejecting the null hypothesis if the p-value of the test is less than 0.05.

As can be seen in Table 1, this test is restrictive enough to cause most of the tested pairs to fail. This could indicate that the relationship between many assets is actually something momentary and not something that represents a long-term equilibrium.

Johansen's Cointegration method

As a final confirmation, the pairs that pass the Engle-Granger filter are subjected to the Johansen test. Unlike Engle-Granger, which is a two-step method, the Johansen test is a more robust multivariate approach that determines the number of cointegrating relationships (or cointegration "rank") in a time series system (QuantStart, 2017)

In this case, the trace statistic is used to validate the null hypothesis which establishes that there are no cointegration vectors $(r = 0)$, and it is rejected when the trace statistic > the 95% critical value (which in this case is 15.49). This would ultimately confirm the existence of $r = 1$ cointegration vector.

Statistical evidence for selected pairs for both Engle-Granger and Johansen

In the end, and as can be anticipated from Table 1, after the filtering process with the three tests, the candidate pairs were ultimately reduced to just two:

ticker1	ticker2	correlation	eg_p_value	jo_trace_stat	jo_crit_val	jo_coint	Strength	jo_eigenvector
CVX	OKE	0.939597	0.001715	24.707649	15.494300	True	1.594628	[0.242633, -0.256791]
ET	WMB	0.908685	0.003344	17.616239	15.494300	True	1.136950	[0.738077, -0.457965]

Table 2. Cointegrated Pairs Report (Passed All Tests)

Both pairs, CVX/OKE and ET/WMB, passed all tests and demonstrate strong statistical evidence that is ultimately only differentiated by the “Strength" metric (Trace Stat / Critical Value), which indicates how far above the threshold the relationship is. Based on the highest value, the best pair is chosen, which in this case is the combination of Chevron (CVX) and ONEOK (OKE), assets that present the strongest cointegration evidence, with the lowest Engle-Granger p-value (0.0017) and the highest Johansen "Strength" ratio (1.59).

Justification: Economic Relationship

Chevron and Oneok, although both are companies in the energy sector (oil and gas), are not actually direct competitors, as they both operate in complementary segments of the value chain. Due to the latter and given the nature of the operations and the main business of both companies, it could bed be said that both companies intrinsically share a fundamental economic relationship that justifies their long-term cointegration.

On one hand, there is Chevron (CVX), which is one of the largest integrated energy companies (upstream and downstream), and whose main business is the exploration and production (E&P) of crude oil and natural gas. This ultimately causes CVX's profitability to be intrinsically linked to global commodity prices.

And on the other hand, there is Oneok, a midstream infrastructure company that owns and operates a variety of pipelines, natural gas processing plants, and natural gas liquids fractionation facilities; and, as its business model is primarily fee-based, its profitability depends on the volume of gas it transports and processes, not directly on the price of the commodity.

Stated this way, the economic relationship and equilibrium between both assets stem from the direct dependence that the midstream segment (OKE) has on the activity generated by the upstream segment (CVX); as the success of upstream operations is what fuels the demand for midstream services. When global commodity prices (like crude oil and natural gas) are high, CVX's profitability increases, incentivizing greater investment in exploration and production. This increased drilling activity inevitably results in a larger volume of extracted product.

It is precisely this volume that must be gathered, processed, and transported from the production fields to refining or distribution centers, and it is right here where OKE's infrastructure becomes indispensable. OKE's business model, based on volume fees, benefits directly from the increase in production from CVX (and other upstream producers). Therefore, although their revenues are derived from different sources (CVX from price, OKE from volume), their financial destinies are intrinsically linked. CVX's success generates the volume on which OKE depends, creating the long-term economic anchor that justifies the statistically observed cointegration.

Charts showing price relationships and spread evolution

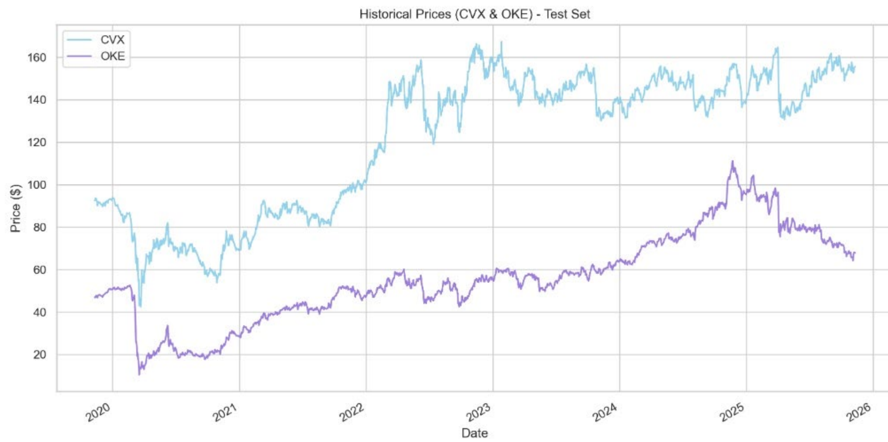


Figure 1. Evolution of historical prices of CVX and OKE.

The raw prices of CVX and OKE clearly move in the same general direction, but they operate on very different price scales. This makes it difficult to visualize their relative relationship.

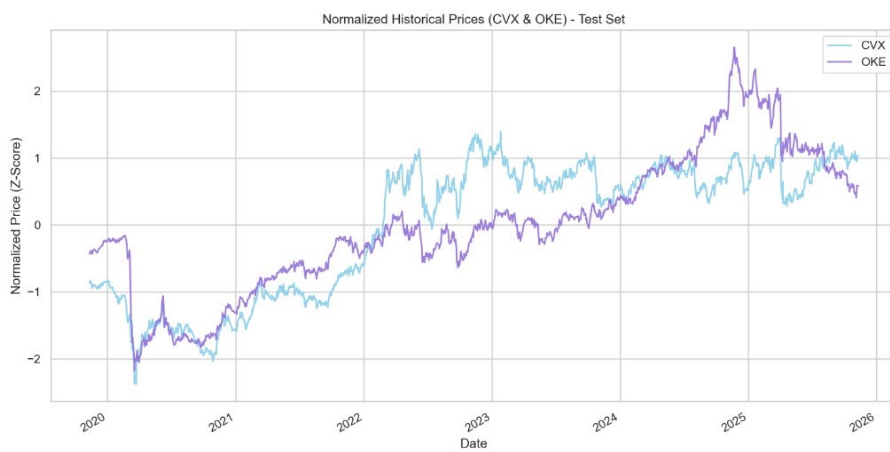
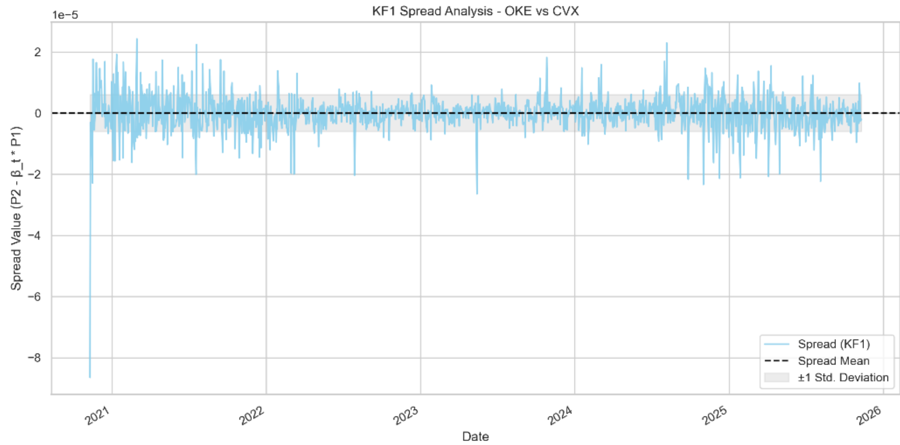


Figure 2. Normalized historical prices (Z-Score) of CVX and OKE.

By normalizing the prices (converting them to Z-Scores), their short-term co-movement becomes evident. The series repeatedly cross, diverge, and converge, visually showing the high correlation.



Evolution of the pair spread, estimated by the Kalman Filter 1.

This graph is the visual proof of cointegration. It shows the dynamic "equilibrium error." $(P_2 - (\beta_0 + \beta_1 P_1)$. It is clearly observed that the spread consistently fluctuates around a mean (the black dotted line near zero), demonstrating the property of mean reversion (stationarity). The temporary deviations from this mean (the peaks and valleys) are precisely the market inefficiencies that the strategy seeks to capitalize on.

Sequential Decision Analysis Framework

Model 1: Estimation of the Dynamic Hedge Ratio (KF1)

- Narrative: The objective is to estimate the parameters of the linear regression $p_2 = \beta_0 + \beta_1 \cdot p_1$. We assume that this relationship is not static and that the parameters $[\beta_0, \beta_1]$ follow a random walk. A filter is needed that updates its belief about these parameters each day (with each new price) to obtain a dynamic and precise hedge ratio.
- Key elements:
    - Metrics: Minimize the Mean Squared Error (MSE) of the price estimate p_2. The objective is the precision of the hedge, not profitability.
    - Decisions: The "decision" of the filter at time t is the algorithmic act of updating its state of beliefs $p_2(w_{t|t}, P_{t|t})$ after observing the new prices.
    - Uncertainty: There are two sources. The first is process uncertainty (Q), which models how the parameters β ""real" values can change randomly from day to day, and 2 the measurement uncertainty (R), which models market noise (price p_2 observed function is not a perfect function of p_1).
- Mathematical Model:
    - State S_t: As a purely belief-based state (B_t). $S_t = (w_{t|t}, P_{t|t})$, where:
        - $w_t = [\beta_0, \beta_1]^T$: Is the (2x1) state vector that has the hedge ratio
        - P_t: As the covariance matrix that measures the uncertainty of the estimate w_t
    - Exogenoues information W_t: The new information arriving at time t is the observed prices of both assets: $W_t = (p_{1,t}, p_{2,t})$
    - Transition function $S_t = S^M(S_{t-1}, x_{t-1}, W_t)$: The transition is defined by the Kalman Filter equations:
        - State: $w_t = Aw_{t-1} + n_t$ here a random walk is assumed, so the transition matrix A is the identity (2x2). Y $n_t \sim N(0, Q)$ It's the noise of the process.
        - Observation: $y_t = H_t w_t + v_t$

Where:

- $y_t = p_{2,t}$
- $H_t = [1, p_{1,t}]$
- $v_t \sim N(0, R)$ It's measurement noise.

  - Objective Function: Minimize the MSE of the state estimate

$$\min_{\pi} E\left[\sum_{t=0}^{T} (w_t - \widehat{w}_t)^2\right]$$

- Uncertainty model: Defined by the covariance matrices Q y R
- Policy design: Policy π is the filter's own update algorithm. And the Kalman gain:

$$K_t = P_{t|t} H_t^T \left(H_t P_{t|t-1} H_t^T + R\right)^{-1}$$

  And the upgrade decision is: $w_{t|t} = w_{t|t-1} + K_t(y_t - H_t w_{t|t-1})$
- Policy Evaluation: The policy (the filter) is visually evaluated by analyzing the stability of β_1 (Figure 4) and the stationarity of the resulting spread (Figure 3).

Model 2: VECM Signal Estimation (KF2)

- N arrative: The objective is to estimate the parameters of the VECM model $Spread_t = e_1 \cdot p_{1,t} + e_2 \cdot p_{2,t}$. The Johansen test in a moving window gives us a "noisy" estimate of this spread, so a KF2 is used to filter and smooth this signal, treating the Johansen output as the "noisy observation" and the parameters $[e_1, e_2]$ as the hidden state.
- Key Elements
  - Metrics: Generate a filtered VECM signal ($Signal_t = \hat{e}_1 p_1 + \hat{e}_2 p_2$), that is robustly stationary and captures equilibrium deviations (Figure 6).
  - Decisions: The filter's "decision" is to update its belief about the state vector. $w_t = [e_1, e_2]^T$
  - Uncertainty: The first is Process Noise (Q): The "true" cointegrating vector can change slowly over time. The second is Measurement Noise (R): The output of the moving-window Johansen test is only a noisy estimate of the true cointegrating vector.

- Mathematical Model:
    - State S_t: $S_t = (w_{t|t}, P_{t|t})$ where:
        - $w_t = [e_1, e_2]^T$: Is the state vector that has the components of the eigenvector.
        - P_t: The error covariance matrix
    - Exogenous Information W_t : $W_t = (p_{1,t}, p_{2,t}, v_{obs,t})$ where
        - $p_{1,t}$, $p_{2,t}$ the prices
        - $v_{obs,t} = [e_{1,obs}, e_{2,obs}]^T$ is the eigenvector obtained from the Johansen cointegration test in the moving window.
    - Transition Function $S_t = S^M(S_{t-1}, x_{t-1}, W_t)$:
        - State: $w_t = Aw_{t-1} + n_t$. I is assumed that $A = I$ (Identity) and $n_t \sim N(0, Q)$
        - Observation: $y_t = H_t w_t + v_t$, where
            - $y_t = (e_{1,obs} \cdot p_{1,t}) + (e_{2,obs} \cdot p_{2,t})$
            - $H_t = [p_{1,t} \, , p_{2,t}]$
            - $v_t \sim N(0, R)$
    - Objective Function: Minimize the MSE of the eigenvector estimation.

$$\min_{\pi} E\left[\sum_{t=0}^{T} (w_t - \widehat{w}_t)^2\right]$$

And the trading signa lis calculated as $Signal = H_t w_{t|t}$

- Uncertainty model
    - $n_t \sim N(0, Q)$
    - $v_t \sim N(0, R)$
- Policy design: The π policy is the Kalman update algorithm, identical in its mathematical formulation to that of KF1 (Prediction, Kalman Gain, Update).
- Policy Evaluation: It is evaluated by analyzing the stability of the eigenvector components e_1 and e_2 (Figure 5)

Model 3: Trading Policy Optimization (θ)

- Narrative: The goal is to find the optimal Z-score (θ) threshold for our trading policy. KF1 tells us how much to buy/sell (the hedge ratio), and KF2 gives us the signal when (the Z-score). This model determines how extreme that signal ($Z_t > \theta$) must be for us to decide to act.
- Key Elements
  - Metrics: Maximize the Calm Ratio in the backtesting set.
  - Decisions: The daily trading decision.

$$x_t \in \{open\ short\ , open\ long, close, hold\}$$

  - Uncertainty: The future sequence of prices $W_{t+1}, \dots\ , W_T$ is unknown; it is not known which π_θ policy will work best.
- Mathematical Model
  - Estate S_t : $S_t = (C_t, Pos_1, Pos_2, Z_t)$. physical and information status of the portfolio.
    - C_t: Available cash capital
    - Pos_i: Active long and short positions in Ticker 1 and Ticker 2
    - Z_t: The current Z-score (normalized KF2 output)
  - Exogenous Information W_t: $W_t = (p_{1,t}\ , p_{2,t})$ daily prices
  - Transition $S_{t+1} = S^M(S_t, x_t, W_{t+1})$
    - If $x_t = open\_long$, then C_{t+1} decreases due to the cost of the operation and are update Pos_1, Pos_2
    - If $x_t = close$, then C_{t+1} increases due to the PnL and the positions go to zero.
    - Z_{t+1} it is updated based on the KF2 with W_{t+1}
  - Objective: $\max_{\theta \in [0.5, 3.0]} Calmar\ Ratio(\pi_\theta | S_\theta)$ over the time horizon T of the backtest.
- Uncertainty model: The sequence of future prices $W_1, \dots\ , W_T$ is stochastic and is modeled using historical data.
- Policy design: Policy π is a Threshold Policy parameterized by θ. Where:
  - If $Z_t > \theta$ and there is no open position→ $x_t = open\ short$

- If $Z_t < -\theta$ and there is no open position → $x_t = open\ long$
- If $Z_t < 0.1$ and there is an open position → $x_t = close$

- Policy Evaluation: The evaluation is performed by simulation (backtesting). The code “utils.optimize_std_threshold” runs a full backtest for each θ in the range [0.5,3.0] with steps of 0.1. The Calmar Ratio of each simulation is recorded (Table 3), and the θ that maximizes it is selected, resulting in θ * = 1.6.

Kalman Filter Implementation

To address the variability and uncertainty in market relationships, this strategy is articulated upon two distinct Kalman Filters, each with a specific objective, as was discussed in the theoretical sessions. The first filter (KF1) estimates the optimal hedge ratio, while the second (KF2) stabilizes the trading signal from the Vector Error Correction Model (VECM).

Initialization procedures

- Filter Initialization 1

The first Kalman model is responsible for modeling the regression ($p_2 = \beta_0 + \beta_1 * p_1$ which is initialized with a “neutral” state. This is because the initial state vector w_0 is set as $[0.0, 1.0]$ which would represent an initial assumption of a zero intercept and a hedge ratio (β_1) of 1. This state will be updated immediately by this filter as soon as the first price observation is received.

- The second Kalman filter is responsible for modeling the spread as $VECM = e_1 * p_1 + e_2 * p_2$. In this case, instead of a random assumption, the filter is initialized using calculated statistical evidence; for example, within the code, the Johansen test is applied to the first window of the data, and as a result, the first eigenvector is generated, which is used as the initial state w_0 within the second filter.

Parameter estimation methodology

KF1, as mentioned, estimates a dynamic linear regression model where the hidden state is the parameter vector $w_t = [\beta_0, \beta_1]^T$, which yields the intercept and the hedge of the strategy at time t. For this, the system is defined by two equations:

- State or Transition Equation:

  Which assumes that the parameters behave randomly and that the state t is equal to the state $t-1$ plus noise from a process n_t. $w_t = w_{t-1} + n_t$, where $n_t \sim N(0, Q)$

- And the Observation or Measurement Equation:

  Where the observed price of asset 2 (y_t) is a linear function of the state w_t and the observation H_t (which contains the price of asset 1), plus a measurement noise v_t. $y_t = H_t w_t + v_t$, where $v_t \sim N(0, R)$.

  For this filter, the components of the observation are defined by:

  - $y_t = p_{2,t}$ which is the observed price of asset 2.
  - $H_t = [1, p_{1,t}]$ which would be the dynamic observation matrix that includes the price of asset 1.
  - $w_t = [\beta_0, \beta_1]^T$ which is the state vector to be estimated.

And iteratively for each time step t the same filter is updated:

- Prediction**:** The filter predicts the next state $\widehat{w}_{(t|t-1)}$ and its covariance $P_{(t|t-1)}$ based only on the information up to $t-1$. Such that:
  - $\widehat{w}_{(t|t-1)} = w_{(t-1|t-1)}$
  - $P_{(t|t-1)} = P_{(t|t-1)} + Q$
- And then**,** once the real Price $y_t = p_{2,t}$, is observed, the filter itself is responsible for correcting its prediction.
  - First by calculating the difference between the real price and the predicted price.

$$e_t = y_t - H_t \, \widehat{w}_{(t|t-1)}$$

- Then the covariance of this innovation

$$S_t = H_t P_{(t|t-1)} H_t^T + R$$

- The Kalman gain, which calculates the optimal weight that should be given to the residual

$$K_t = P_{(t|t-1)} H_t^T S_t^{-1}$$

- The state is updated by correcting itself

$$w_{(t|t)} = \widehat{w}_{(t|t-1)} + K_t e_t$$

- And the covariance is updated

$$P_{(t|t)} = (I - K_t H_t) P_{(t|t-1)}$$

Which give as a final result β_1 q which is extracted from the second component of the updated state vector $w_{(t|t)}$ and represents, as has been mentioned repeatedly, the hedge ratio for that day's operations.

The second Kalman filter (KF2), follows a very similar logic with the difference being how its components are defined and its objective, which is now to stabilize the VECM signal.

As a hidden state, it has the cointegration vector $w_t = [e_1, e_2]^T$ and the state and observation equations are identical to the previous one:

- State: $w_t = w_{t-1} + n_t$, where $n_t \sim N(0, Q)$
- Observation: $y_t = H_t w_t + v_t$, where $v_t \sim N(0, R)$.

However, here the components are defined differently:

- y_t : Which is the observed and noisy VECM signal. It is not a price, but the result of executing a complete Johansen test on a 252-day rolling window to obtain a vector $v_{obs} = \left[e_{1,obs}, e_{2,obs}\right]$, which is multiplied by the current prices $y_t = e_{1,obs} \cdot p_{1,t} + e_{2,obs} \cdot p_{2,t}$.
- $H_t = \left[p_{1,t}, p_{2,t}\right]$: Which is the observation matrix composed of the prices of both assets

- $w_t = [e_1, e_2]^T$: which is the vector that we are trying to estimate.

The Prediction and Update cycle of KF2 uses the same Kalman Gain and state update formulas as KF1, but applying the y_t and H_t definitions specific to this filter. The final "filtered signal," which will be used for trading, is not the state vector $w_{t|t}$ directamente, directly, but the dot product of that stabilized state $w_{t|t} = [\hat{e}_1, \hat{e}_2\ ]^T$ multiplicated by the observation matrix H_t

$$H_t \cdot w_{t|t} = \hat{e}_1 \cdot p_{1,t} + \hat{e}_2 \cdot p_{2,t}$$

Which results in a smoothed and stabilized estimate of the VECM spread, which is then normalized (converted to Z-score) to generate trading signals.

Reestimation schedule and validation approach

The re-estimation schedule is daily. Both filters, KF1 and KF2, perform a complete prediction and update cycle on each trading day of the backtest. There is no batch re-estimation; the model was designed to be purely recursive and adapt with each new observation.

On the other hand, the validation approach could be said to be a walk-forward analysis where the model is initialized with the first 252 days of data; and, from that point, it "walks forward" day by day, using only the information available up to day $t-1$ to make decisions on day t. The optimal strategy parameters (such as the θ threshold) are determined on the training set (60% of the data) and their robustness is validated on the test set (40% of the data), ensuring there is no “look-ahead bias” in the final evaluation.

Convergence analysis and filter stability

To validate that the filters are functioning correctly, a graph can be made of the evolution of the parameters they estimate, so that, if the filters are stable, the parameters should converge and adapt smoothly to regime changes, without diverging to infinity or presenting impossible situations like a negative hedge ratio.

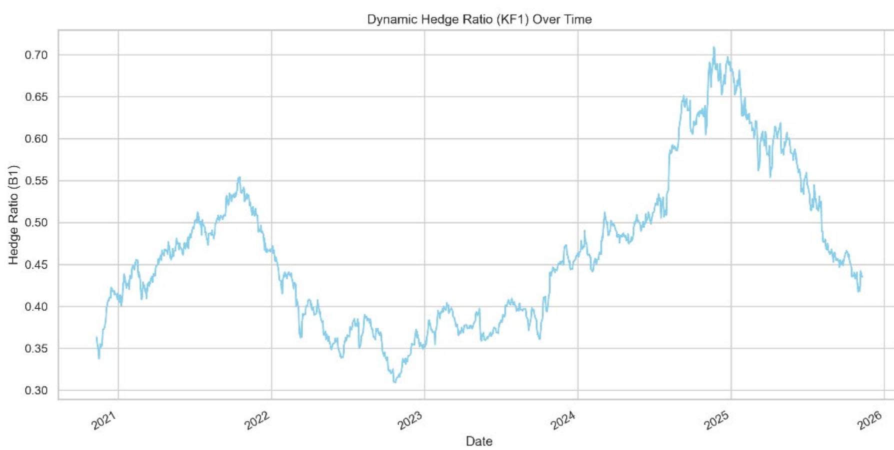
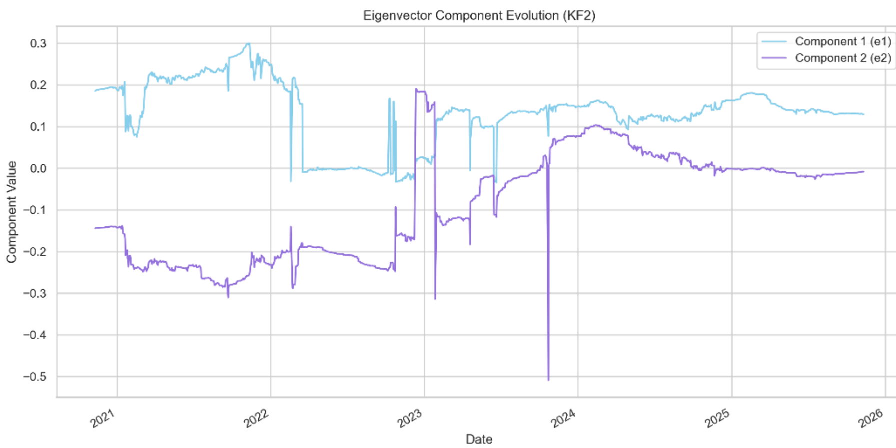


Figure 4. Evolution of the Dynamic Hedge Ratio (B1), Kalman 1

a high near 0.72 in 2025, and subsequently correcting. The interesting thing about this graph is noting that the trajectory is smooth and reactive, demonstrating that the filter converges and successfully tracks the changing relationship between the prices. This justifies the use of a dynamic filter, as a static hedge ratio (calculated only once) would have been incorrect for most of the period and would have possibly caused losses.



Evolution of the Eigenvector (e1 and e2), Kalman 2

This graph, on the other hand, shows the evolution of the two components of the KF2 state. Both parameters are highly dynamic, reflecting the changes in the observed VECM (calculated with the rolling-window Johansen), in which different market regimes can be clearly identified. For example, the strong volatility in 2022 and the abrupt changes at the end of 2023, where the parameters adjust rapidly to new information. Despite these strong shocks, the parameters remain bounded and do not diverge, demonstrating the filter's stability.

Trading Strategy Logic

Z-score definition using VECM

The execution logic of this strategy is based completely on the trading signal, which represents a stabilized and smoothed version of the VECM spread; this, as was already broken down in previous sections, is generated by the second Kalman filter.

Definition:

- Filtered signal S_t: Where for each time step t, the VECM is calculated as $S_t = \hat{e}_{1,t} \cdot p_{1,t} + \hat{e}_{2,t} \cdot p_{2,t}$, where $[\hat{e}_1, \hat{e}_2\ ]$ is the eigenvector stabilized by KF2 at time t
- Z-score (Normalization): This obtained signal S_t se almacena en "historical_filtered_signal" is stored in "historical_filtered_signal" and normalized to generate the final trading z-score using the mean and standard deviation of a rolling window containing the last 252 days. This can be seen as:

$$Z_t = \frac{S_t - \mu_{S,252}}{\sigma_{S,252}}$$

  Where μ and σ are the mean and standard deviation of the filtered signal S over the past 252 days, and Z_t is the oscillator that measures deviations from equilibrium and upon which the entry and exit decisions for the model's trades are made.

Optimal entry and exit Z-score policy found

To find the optimal entry policy, an iterative optimization process was carried out on the test set, which consisted of repeatedly running the backtest over a range of entry thresholds θ starting from 0.5 up to 3. This was done with the objective of "manually" finding the value that maximized the strategy's Calmar ratio, from which the following table was obtained.

Theta	Calmar	Sortino	Sharpe	MaxDrawdown	FinalValue	TotalReturn	Trades	WinRate	AvgPnL
0.5	-0.0137	-0.0038	-0.0033	-21.55	$ 982,411.55	-1.76	12	25	-$211,712.44
0.6	-0.03	-0.0088	-0.0538	-21.33	$ 962,375.80	-3.76	12	25	-$210,965.08
0.7	-0.036	-0.0861	-0.0766	-21.92	$ 953,714.39	-4.63	12	33.33	-$212,011.27
0.8	-0.034	-0.0784	-0.0698	-21.67	$ 956,650.19	-4.33	12	33.33	-$211,865.72
0.9	-0.0384	-0.0964	-0.0863	-22.25	$ 949,881.22	-5.01	12	25	-$211,518.36
1	-0.045	-0.1297	-0.1169	-23.09	$ 939,358.09	-6.06	12	25	-$211,613.78
1.1	0.1412	0.3408	0.3492	-14.99	$ 1,133,554.32	13.36	12	25	-$216,012.62
1.2	0.0617	0.1657	0.1748	-15.05	$ 1,056,847.22	5.68	12	25	-$216,956.56
1.3	0.0745	0.182	0.195	-14.09	$ 1,064,418.89	6.44	12	25	-$216,320.25
1.4	0.0707	0.1746	0.187	-14.09	$ 1,061,149.67	6.11	12	25	-$216,468.87
1.5	0.0871	0.2106	0.2276	-14.11	$ 1,075,834.53	7.58	10	30	-$212,262.04
1.6	0.3108	0.5286	0.6208	-10.81	$ 1,218,684.95	21.87	8	37.5	-$186,353.89
1.7	0.3021	0.5146	0.6042	-10.81	$ 1,212,018.25	21.2	8	37.5	-$187,186.79
1.8	0.288	0.4908	0.5888	-10.81	$ 1,201,335.89	20.13	8	25	-$187,570.71
1.9	0.2705	0.463	0.555	-10.81	$ 1,188,277.74	18.83	8	25	-$188,669.55
2	0.2727	0.4661	0.5607	-10.81	$ 1,189,904.04	18.99	8	25	-$188,619.33
2.1	0.1655	0.2867	0.3651	-10.81	$ 1,111,968.44	11.2	8	37.5	-$191,032.97
2.2	0.1302	0.2269	0.3002	-10.81	$ 1,087,210.80	8.72	4	25	-$177,240.58

2.3	0.1302	0.2269	0.3002	-10.81	$ 1,087,210.80	8.72	4	25	-$177,240.58
2.4	0.1302	0.2269	0.3002	-10.81	$ 1,087,210.80	8.72	4	25	-$177,240.58
2.5	0.1302	0.2269	0.3002	-10.81	$ 1,087,210.80	8.72	4	25	-$177,240.58
2.6	0.1302	0.2269	0.3002	-10.81	$ 1,087,210.80	8.72	4	25	-$177,240.58
2.7	0.1302	0.2269	0.3002	-10.81	$ 1,087,210.80	8.72	4	25	-$177,240.58
2.8	0.1302	0.2269	0.3002	-10.81	$ 1,087,210.80	8.72	4	25	-$177,240.58
2.9	0.1302	0.2269	0.3002	-10.81	$ 1,087,210.80	8.72	4	25	-$177,240.58
3	0.1302	0.2269	0.3002	-10.81	$ 1,087,210.80	8.72	4	25	-$177,240.58

Table 3. Optimization results of the entry threshold (Theta)

Within this table, it is shown that the Calmar Ratio reaches its maximum value of 0.3649 when the threshold θ is set at 1.6, and, as the threshold increases beyond this point, the Calmar Ratio decreases, which leaves this threshold as the one that provides the best risk-adjusted return for drawdown for this strategy, with this pair of assets.

- Entry Policy

And with this defined threshold ($\theta = 1.6$) he short entry on the spread is defined so that, if Z_t rises and crosses above 1.6, it would indicate that the spread is overvalued and should show signs of downward reversion; therefore, an “open_short_pair” would be executed with a long in CVX (Ticker 1) and a short in OKE (Ticker 2).

Conversely, if Z_t (vecm_norm) plummets and crosses the threshold of -1.6, it would now be assumed that the spread is undervalued, and an upward correction would be expected. And, in that case, an “open_long_pair” would be executed, opening a short in CVX (Ticker 1) and opening a long with OKE (Ticker 2).

I Regardless of the case, 80% of the available capital would always be invested (divided into 40% and 40% for each trade).

- Exit Policy

Once positions are open, the model will monitor them daily until the moment when Z_t reverts to its mean, which is when the closing signal can be activated. This signal is activated as long as the absolute value of the Z-score falls below 10%.

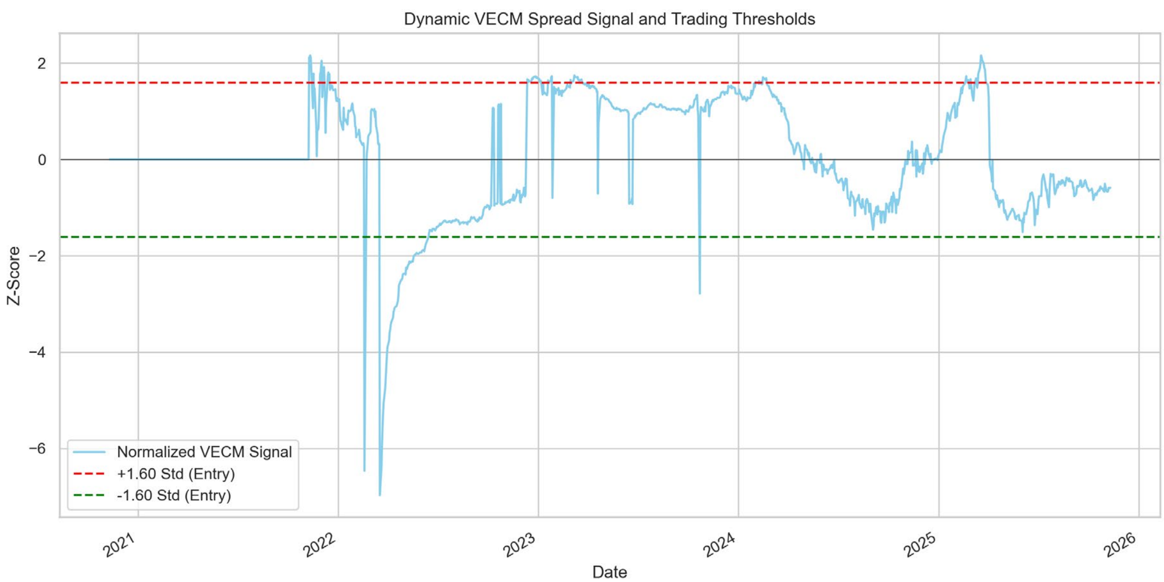


Figure 6. Normalized VECM trading signal (Z-Score) and optimal entry thresholds

Cost treatment: commissions and borrow rates

To simulate the performance of this strategy more realistically, certain conditions that affect the backtest have been included, incorporating intrinsic factors of real-world asset trading, such as commissions (cost per transaction) and the borrow rate as the cost of financing short positions.

- Commissions:

For this point, a fixed commission rate of 0.125% (COMMISSION_RATE) was implemented, which is applied to the value of each executed trade, and this is incurred both at the opening and closing of all positions, directly affecting the available capital. At the opening, for example, the capital is reduced by the total cost associated with the long position (what was paid for the shares plus the

0.125% commission) and by the commission cost of the short position (0.125% of the value of the shares sold).

- Borrow Rate

This rate, set at 0.25% annually, is applied to short positions for each day the position remains open. Within the backtesting, a function (_update_borrow_costs) is executed daily, calculating the borrowing cost based on the current market value of the short position and multiplying that value by the daily lending rate.

Results and Performance Analysis

Equity curve plots (train, test, validation periods)

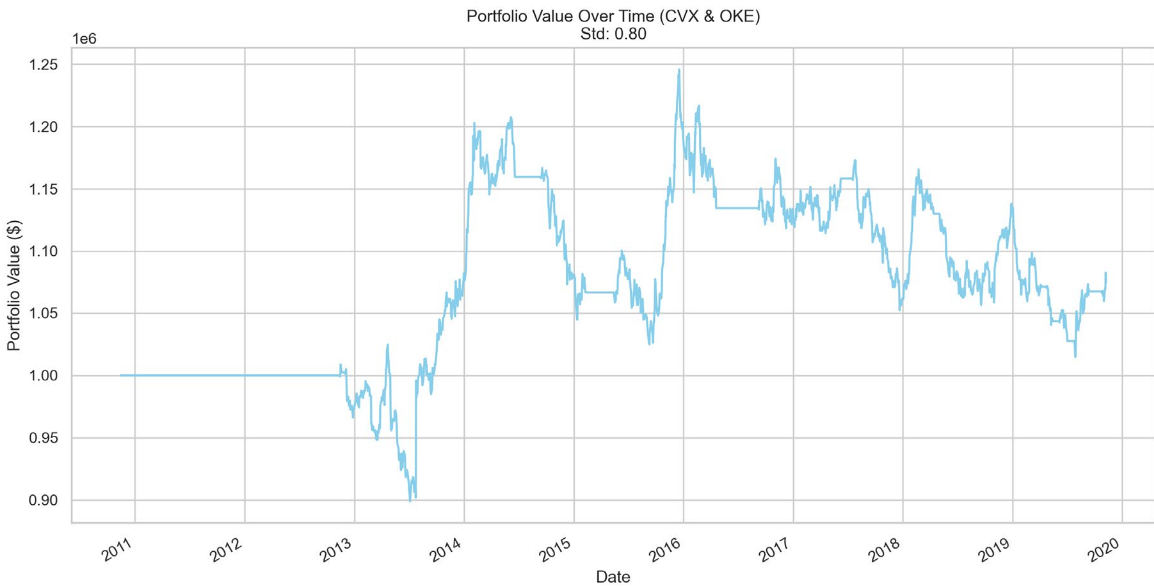


Figure 7. Portfolio Value over time (Train)

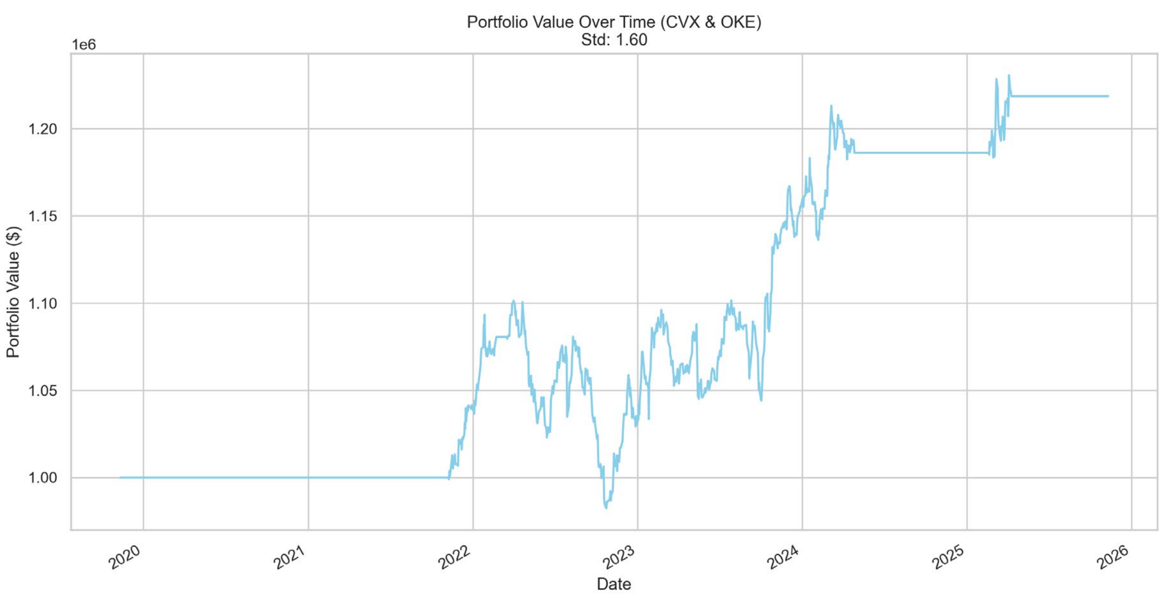


Figure 8. Portfolio Value over time (Test)

n this last graph of the “test,” it is noted how, at the beginning of the period, it remained inactive for the first two years; this is expected considering the window the model needs to begin operating (252 days). Approximately at the beginning of 2022, the strategy initiated its first trade, generating an almost immediate positive return, followed by a period of volatility and a decline at the end of 2022. All to then experience sustained growth again during 2023 and early 2024, before stabilizing at a new capital level given that the normalized VECM spread did not reach the entry threshold of plus or minus 1.60. And, finally, in 2025, the last profitable trade of the strategy would be executed. The stepped shape of the curve is to be expected in a mean-reversion strategy with wide thresholds, which executes few but long-duration trades.

Performance metrics

Despite the market downturns and the few trades executed, the model managed to generate a favorable return at the end of the period, finishing with a capital of $1,218,684.95, which represents a Total Return of 21.87% on the initial capital with the following performance metrics.

Performance Metrics			
Sharpe Ratio	**Sortino Ratio**	**Calmar Ratio**	**Maximum Drawdown**
0.6208	0.5286	0.3108	-10.81%

Table 4. Performance Metrics

The final ratios of this model ended with positive values, which would indicate that the strategy managed to generate returns that, in a certain way, “beat” the negative trends and volatilities of the market. Of all these, the important one for this exercise was the Calmar ratio, which ended at a value of 0.3108, which would indicate that the strategy generated a good return for each unit of risk taken. On the other hand, the maximum drawdown shows something interesting, which is that in the worst-case scenario, the model managed to remain robust by preserving capital as neutral to the market as possible, minimizing potential losses.

Trade statistics

In general, with an entry threshold of 1.6, and a narrow exit threshold of 0.1, the model behaved austerely: few operations, but made at the precise moment.

Total operations	Win rate
8	37.05%

Table 5. Strategy data

This result is interesting as it would be indicating that it approximately won only 3 trades that it made, but given the final portfolio results, it can be understood that the few trades the model won were large enough to compensate for the small trades that generated losses.

Figure 9. Operations performed for each asset.

This graph visually shows the operations performed by each asset simultaneously.

Figure 10. PnL Distribution

Conclusions

In closing this exercise, according to the results obtained by the developed model, the “viability” of a strategy that merges topics like cointegration and parameter estimation by Kalman filters, such as the pairs trading strategy developed in this project, could be confirmed. This three-stage selection process (correlation, Engle-Granger, and Johansen) ultimately proved successful in isolating and trading the CVX and OKE pair with a robust statistical justification.

On the other hand, the analysis of the filters ended up validating the central hypothesis of the project, which is that it is neither convenient nor efficient to operate considering a static scenario; since both the coverage ratio and the cointegration vector proved to work and adapt well by being dynamic, thus justifying the superiority of having filters that constantly adapt compared to a static parameter model. For example, in figures 4 and 5 on convergence, they demonstrated that the filters successfully tracked these regime changes, adapting to market shocks without diverging, which confirms their stability. Likewise, the strategy proved to be profitable even after applying realistic transaction costs, achieving a positive return of 21.87% with a maximum drawdown of -10.81%.

Finally, considering future improvements and expansions to this project. This project could benefit from a more adaptive and flexible parameter optimization, such as implementing a more robust approach that performs a re-optimization of theta within the rolling window so that the model can better adapt to changes in the market regime and, furthermore, not be limited in the operations it can have open simultaneously. In this regard, it could also be interesting to apply this same concept to the allocation of investment capital, which is currently fixed at a percentage and could be adjusted based on the magnitude of the Z-score or according to market volatility. Finally, the strategy could be expanded to the management of a portfolio of cointegrated pairs, which would favor diversification instead of depending on the returns of a single pair.

References

- Granger, C. W. J., Ghysels, E., Swanson, N. R., & Watson, M. W. (2001). Essays in Econometrics: Volume 2, Causality, Integration and Cointegration, and Long Memory : Collected Papers of Clive W. J. Granger (Vol. 00033). Cambridge University Press.
- The AI Quant. (2024, 21 May). Kalman Filters are a powerful tool in the world of finance for modeling and predicting time-series. Medium. https://theaiquant.medium.com/kalman-filters-are-a-powerful-tool-in-the-world-of-finance-for-modeling-and-predicting-time-series-6b4c614244d3
- QuantStart. (2017, 11 January). *Johansen Test for Cointegrating Time Series Analysis in R*. QuantStart.com. https://www.quantstart.com/articles/Johansen-Test-for-Cointegrating-Time-Series-Analysis-in-R/